Munich Personal RePEc Archive

Optimization of Post-Scoring Classification and Impact on Regulatory Capital for Low Default Portfolios

Genest, benoit and Fares, Ziad

27 April 2014

Online at https://mpra.ub.uni-muenchen.de/62907/
MPRA Paper No. 62907, posted 16 Mar 2015 15:45 UTC

© CHAPPUIS HALDER & CIE

Optimization of Post-Scoring Classification and Impact on Regulatory Capital for Low Default Portfolios

17/04/2014

By Benoit GENEST & Ziad FARES

Global Research & Analytics[1]

[1] This work was supported by the Global Research & Analytics Dpt. of Chappuis Halder & Cie.
E-mail : zfares@chappuishalder.com (Paris), bgenest@chappuishalder.com (London)

Optimization of Post-Scoring Classification and Impact on Regulatory Capital for Low Default Portfolios

Abstract

After the crisis of 2008, new regulatory requirements have emerged with supervisors strengthening their position in terms of requirements to meet IRBA standards. Low Default Portfolios (LDP) present specific characteristics that raise challenges for banks when building and implementing credit risk models. In this context, where banks are looking to improve their Return On Equity and supervisors strengthening their positions, this paper aims to provide clues for optimizing Post-Scoring classification as well as analyzing the relationship between the number of classes in a rating scale and the impact on regulatory capital for LDPs.

Key words: Basel II, Return On Equity, RWA, Classification trees, Rating scale, Gini, LDP

1. Introduction

The 2008 crisis was the main cause of tough market regulation and banks' consolidation basement. These new constraints have caused a major increase in banks' capital. As a result, banks need to optimize their return on equity, which has suffered a big drop (due both to a drop in profitability as well as an increase in capital).

Moreover, as this crisis showed loopholes in risk measurement and management, regulators' tolerance becomes more and more stringent. Consequently, banks are facing challenges while dealing with Low Default Portfolios (LDP) and the way to meet regulatory requirements in terms of credit risk management under the Advanced Approach (IRBA) of Basel rules.

The purpose of this paper is to focus on post-scoring classification for LDPs with the aim to study the possibility of building a rating scale for these portfolios that meets regulatory requirements as well as to identify the opportunities to optimize RWA, by studying the relationship between the number of classes within a rating scale and the impact on RWA. The analysis will follow different steps.

First, the different classification techniques will be detailed and studied from a theoretical point of view. The purpose is to understand the underlying statistical indicators and algorithms behind each technique. This will allow understanding the differences between these techniques.

Second, once the classification techniques are analyzed, we will be placed from a credit scoring perspective and define the constraints that must be respected by the rating scale in order to be compliant with the regulatory requirements.

Third, a study will be conducted on a real Retail portfolio with a low frequency of loss events. This portfolio is constituted of mortgage loans, with a significant number of loans making the granularity assumption reliable. The aim of the study is to build rating scales using the different techniques and to analyze the difference between the results. This will allow identifying which classification technique provides the best results (in terms of stability, discrimination, robustness…) for LDPs. Finally, the relationship between the number of risk classes and the impact on RWA will be established in order to identify potential opportunities for RWA optimization for LDPs.

2. The main credit risk measurement methods

In order to provide the reader with a global view of the framework of this study, we will briefly overview the main credit measurement methods. These methods are a prerequisite for understanding the context in which this paper has been written.

2.1. Heuristic models

The "Heuristic" branch corresponds to all kinds of rather "qualitative" methods. These heuristic models are based on acquired experience. This experience is made of observations, conjectures or business theories. In the peculiar field of credit assessment, heuristic models represent the use of experience acquired in the lending business to deduce the future creditworthiness of borrowers. They determine subjectively the factors relevant to this creditworthiness as long as their respective weights in the analysis. But the adopted factors and weights are not validated statistically.

Four main different types of heuristic models are presented here.

2.1.1. Classic rating questionnaires

Based on credit experts' experience, they are designed in the form of clear questions regarding factors that are determined to be of influence for creditworthiness assessment. To each type of answer are attributed an amount of points: for instance, as an answer to a "Yes/No" question, "Yes" means 1 point and "No" means 0. The amounts of points reflect the experts' opinion on the influence of the factors, but there is no place for subjectivity in the answer. The answers are filled in by the relevant customer service representative or clerk at the bank, and the points for all answers are added to obtain an overall score.
The questions are usually about the customer's sex, age, marital status, income, profession, etc. These factors are considered to be among the most relevant ones for retail business.

2.1.2. Fuzzy logic systems

These systems are a new area of research and they are sometimes considered as a part of the expert systems. They bring to those systems the ability to use fuzzy logic, which allows to replace simple dichotomous variables ("high/low") by continuous ones. For instance, when it comes to evaluating a return, a classical rule would be that the return is low under 15% and high above 15%. But that would mean that a 14.99% return would be "low" whereas 15% is "high", which is absurd. The "fuzzification" process introduces continuity where it did not exist, as it is shown on figure 2.

In the example shown there, a return of 10% would be rated 50% "low", 50% "medium" and 0% high, whereas in the first model it would have been rated simply "low". Thus it determines the degree to which the low/medium/high terms apply to a given level of return.

After this, the determined values undergo transforming processes according to rules defined by the expert system, and they are aggregated with other "fuzzi-fied" variables or simple ones. The result is a clear output value, representing the creditworthiness of the counterpart. It must be noted that inference engine, being fed with fuzzy inputs, will produce fuzzy outputs, which must be "defuzzified" in order to obtain the final, clear output (e.g. a credit rating).

2.1.3.Expert systems

These are software solutions aiming to recreate human reasoning methods, in a specific area. Just as a human brain would do, they try to solve complex and poorly structured problems and to make reliable conclusions. They belong to the "artificial intelligence" field. They are made of 2 parts: a knowledge base, with all the knowledge acquired for the specific area, and an inference engine, which often uses "fuzzy" inference techniques, "if/then" production rules to recreate the human way of thinking.

2.1.4.Qualitative systems

Qualitative systems require credit experts to assign a grade for each factor which, overall, determines a global rating regarding one person's credit. Technically speaking, this solution is used for corporate businesses.

2.2. Statistical models

These statistical models derive from two main theories that give birth to two sets of calculation methods.

2.2.1.Market theory

The models that derive from the so-called "Market Theory" are models that are based on the classical market finance models. The most famous of them is the Merton model, based on the theory on the evolution of financial assets formulated by Black, Scholes and Merton in the early 70's. The "Market theory" models are also called "Causal" models as they focus on the explanation of default: what is the process that leads a firm to default? The Merton model, for instance, defines default as the moment that the firm's asset value falls below its debt value.

2.2.2.Credit theory

The models that derive from the "Credit theory" are a different kind of credit assessment models, as they do not focus on the causes of default. Instead, they try to emphasize the external conditions and factors that lead to default, by using statistical tools in order to identify what factors are really significant for default prediction. The statistical models based on Credit Theory are of 3 different kinds.

Regression and Multivariate Discriminant Analysis (MDA) is the first one, based on the search of the variables (for instance financial statements) that are the most discriminant between solvent and insolvent counterparts.

The second one is a rather new method, called Neural Networks. It tries to reproduce the processing mechanisms of a human brain. A brain is made of neurons which are connected to one another by synapses. Each neuron receives information through synapses, processes it, then passes information on through other neurons, allowing information to be distributed to all neurons and processed in parallel across the whole network. Artificial neural networks also attempt to reproduce the human brain ability to learn.

A third kind of statistical model based on Credit Theory is the Bayesian model. Bayesian models are based on real time, which means they are updated when the information is processed. As a result, Bayesian models help refresh in the portfolio instantly. Furthermore, Bayesian systems use a technology of time-varying parameters that add a dynamic notion to the analysis. For instance, these models can use methods such as Markov Chain Monte Carlo that are very effective for conjunction. Actually, these models have provided accurate estimations and predictions.

3. The Theory and the Methodology behind classification trees

3.1. The use of classification techniques in credit scoring models

When building a credit scoring model and more specially when modeling the Probability of Default (PD) of counterparties (for instance for a retail portfolio), the dependent variable (Y) is binary and can take two possible values:

$$Y = \begin{Bmatrix} \mathbf{0}\ if\ the\ counterparty\ does\ not\ default\ within\ the\ following\ year \\ \\ \mathbf{1}\ if\ the\ counterparty\ defaults\ within\ the\ following\ year \end{Bmatrix}$$

Generally, the PD is modeled by using a logistic regression where a score is attributed to each counterparty based on explanatory variables that were accurately chosen (based on a statistical study showing their ability to predict the default of a counterparty) when building the model:

$$score = \ln\left\{\frac{P(Y=1\,|X)}{1-\ P(Y=1\,|X)}\right\} = \propto_0 + \sum_{i=1}^{n} \propto_i (j).x_i(j)$$

With:

- α_i the parameters of the regression;
- x_i the explanatory covariates;
- $X = \{x_i\}_{i=1..n}$
- P(Y=1|X) is the conditional probability that Y = 1 (default of a counterparty) knowing a given combination of the explanatory variables;
- j reflects the modality of the explanatory variables (for qualitative variables as well as discretized variables).

The conditional probability can then be written as:

$$P(Y=1|X) = \frac{1}{1+e^{-score}} = \frac{1}{1+e^{-(\propto_0 + \sum_{i=1}^{n} \propto_i (j).x_i(j))}}$$

Consequently, for each counterparty a score is attributed. Yet, giving that the portfolio is constituted of many counterparties (for retail portfolios for instance), the score values might be too numerous making it hard to manage risks accurately. In other words, it is essential to regroup the individuals of the portfolio in clusters or classes to better reflect risk categories and make it more convenient for operational use.

There are many techniques used to classify the individuals and build clusters such as:

- Neural Networks;
- Bayesian Networks;
- Kernel estimation (k-nearest neighbors);
- Decision trees;

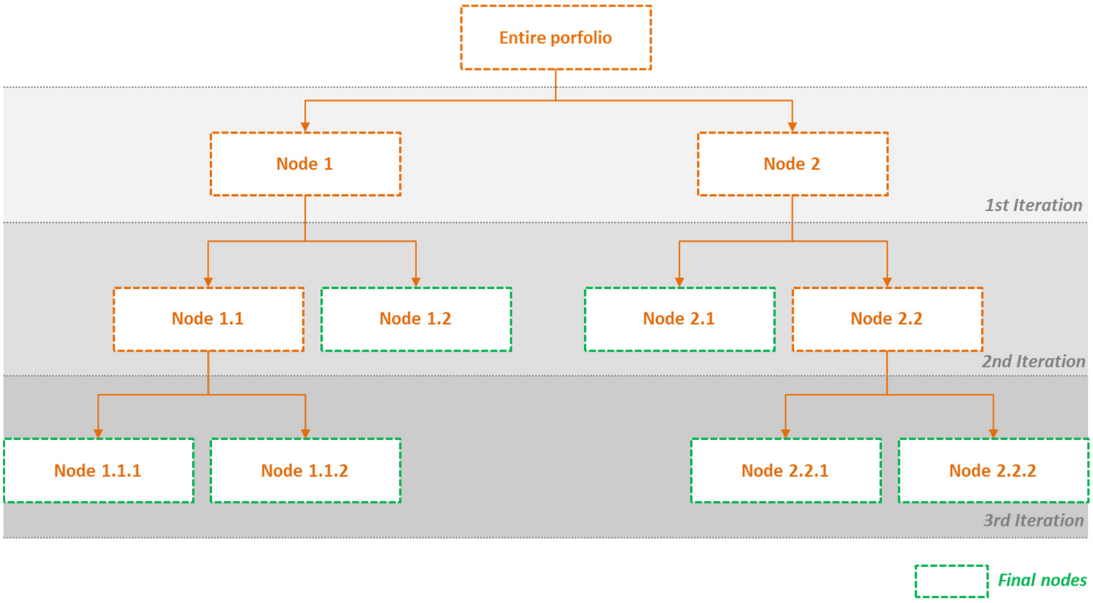
- Classic classification.

This article is dedicated to classification techniques that are frequently used in the banking sector and more precisely in the credit scoring process. Consequently, a focus will be made on Decision trees and classic classification techniques. More precisely, it is the impact of the portfolio classification on RWA (and by extension on capital) that will be analyzed following two main axes:

- The nature of the classification technique;
- The number of classes created.

3.2. Principle of the decision trees and classic classifications

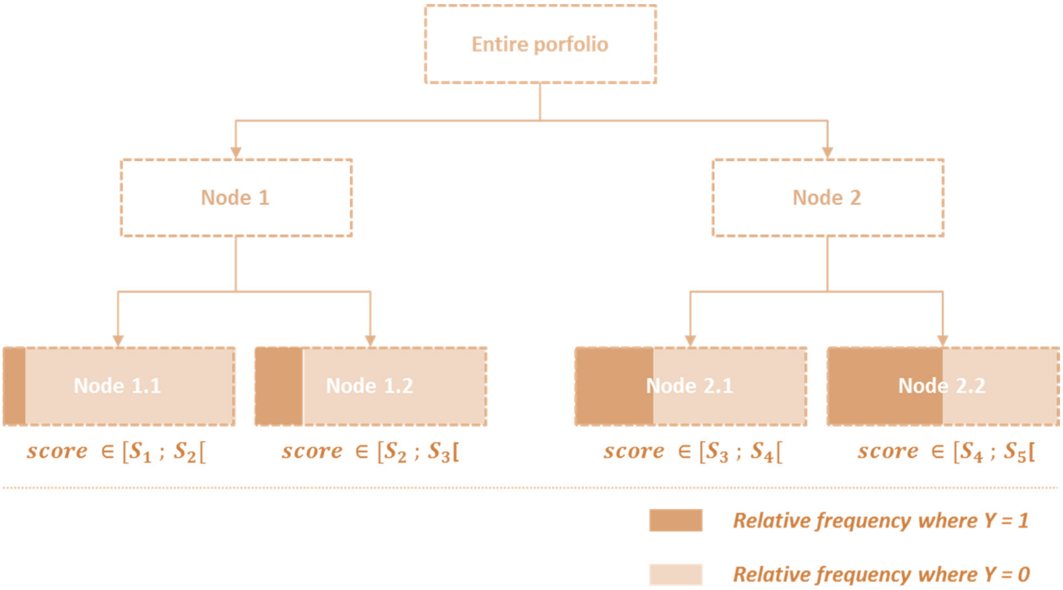
The decision tree technique is one of the most popular methods in credit scoring models. In fact it provides explicit rules for classification as well as it presents features that are easy to understand. Actually, there are two types of decision trees: classification trees and regression trees. Consequently, decision trees are on the boundary between predictive and descriptive methods. For the purpose of this article only classification trees are considered since the prediction of the dependent variable has already been modeled using a logistic regression procedure. The principle of decision trees is to classify the individuals of a population based on a dividing (or discriminating) criterion. Based on this criterion, the portfolio is divided into sub-populations, called nodes. The same operation is repeated on the new nodes until no further separation is possible. Consequently, decision trees are an iterative process since the classification is built by repeating the same operation several times. Generally, a parent node is separated into two child nodes:



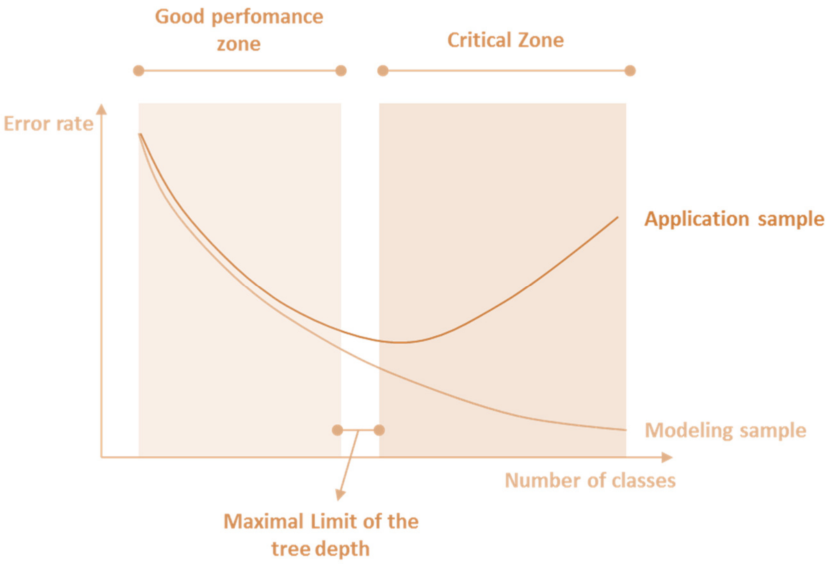
Furthermore, there are mainly three steps in creating decision trees:

- **First Step | Initialization**: the first step aims to define which variable will serve as a dividing criterion and which variable(s) will be used to build and define the clusters or the classes. For instance, in the credit scoring process, the dividing variable is the default variable and the variable used to build the classes is the score. In fact, the aim is to build classes that reflect different levels of credit risk, or in other words, different values of default rate. Consequently, the separation will be built in order to isolate into classes, individuals with the same credit risk level. Thus, the score values will be grouped into classes where the default rate in each class is significantly different. Moreover, the purpose of the credit scoring model is to predict the default probability of a performing loan. Consequently, a score value will be attributed to this loan which will allow assigning it to a risk category (or class). Finally, during this step, the nature of the tree is defined (CART, CHAID, etc…) and its depth (or the potential maximum number of classes) is fixed;

- **Second Step | Distribution of the population in each node:** After initializing the tree, the separation process is launched. The purpose is to obtain classes with significantly different distribution. This process is repeated until no further separation is possible. The stop criterion of the process depends on the tree used, but in general it is related to the purity of the tree, the maximum numbers of nodes, the minimum number of individuals in a node, etc…:



- **Third Step | Pruning:** Pruning is essential when the tree is too deep considering the size of the portfolio. This might lead to classes with low number of individuals which will introduce instability in the classification as well as a random component for these classes. Moreover, if the tree is too deep, the risks of overfitting are high. In other words, the classification on the training (or modeling) sample might provide good results but the error rate might be too high when this classification is used on the application sample :



Concerning classic classification techniques, the paper focuses mainly on two types:

- Create classes with the same number of individuals: this first technique allows the creation of classes with the same number of individuals by class. For instance, when applied on a credit portfolio, the score values are classified in the descending order. Then, knowing the number of classes, a classification is built by regrouping the same number of (ordered) individuals;
- Create classes with the same length: this second technique allows the creation of classes with the same length or in other words with the same intervals. For instance, when applied on a credit portfolio, the score values are also classified in the descending order. Then, individuals are grouped into classes with the same length.

These techniques are simple in terms of application as they are not based on statistical indicators. Consequently, there is no limitation in terms of number of classes.

3.3. Methodological basis of classification trees

Classification trees are mainly based on a statistical approach. The methodology used differs from one tree to another. More precisely, it is the splitting criterion (or in other words the statistical indicator used to assess the level of differentiation between two classes) that differs. The purpose of this paragraph is to explain for each type of tree the nature of the splitting criterion used and its methodological basis.

3.3.1.Classification And Regression Tree (CART)

The CART tree was invented in 1984 by L.Breiman, R.A Olshen, C.J Stone and J.H Friedman. It is one of the most used decision trees and is renowned for its effectiveness and performance.

The main characteristics of this tree are various, which makes it applicable in every circumstance:

- It can be used on all kinds of variables (qualitative, quantitative, …);

- It can also be used for continuous variables (infinite number of categories) or discrete variables (finite number of categories);
- It has a pruning algorithm that is more elaborated that the CHAID technique. In fact, the CART builds a tree with the maximum number of nodes. Then, the algorithm compares the purity (by comparing the error rate) of sub-trees using a sequence of pruning operations and selects the one which maximizes the purity of a tree;
- Finally, one key element in the performance of this technique is its ability to test all possible splits and then to select the optimal one (using the pruning algorithm).

Even though this technique is reliable it has some drawbacks:

- It is binary (one parent node gives two or less child nodes) which might provide trees that are too deep and somehow complex in terms of interpretation;
- The processing time might be long. In fact, this can happen when applied for credit risk purposes, where there are too many values of score (knowing that a retail portfolio with high granularity is considered).

The splitting criterion used in the CART is the Gini index. This index measures the population diversity or in another words the purity of a node. The Gini index of a node is measured as follows:

$$Gini_{node} = 1 - \sum_{i=1}^{m} P_i^2 = \sum_{i=1}^{m} \sum_{\substack{j=1, \\ j \neq i}}^{m} P_i . P_j$$

Because,

$$\left(\sum_{i=1}^{m} P_i \right)^2 = 1 = \sum_{i=1}^{m} P_i^2 + \sum_{i=1}^{m} \sum_{\substack{j=1, \\ j \neq i}}^{m} P_i . P_j$$

Where,

- $Gini_{node}$ is the Gini index of a node;
- m is the number of categories of the dependent variable;
- P_i is the relative frequency of the i_{th} category of the dependent variable.

If the gini index increases, the purity of the node decreases. Consequently, during the separation phase, the best split is the one that provides the best decrease in the Gini index. In other words, the criterion to be minimized is:

$$Gini_{separation} = \sum_{k=1}^{r} \frac{n_k}{n} . Gini_k = \sum_{k=1}^{r} \left(\frac{n_k}{n} . \sum_{i=1}^{m} \sum_{\substack{j=1, \\ j \neq i}}^{m} P_i(k) . P_j(k) \right)$$

Where,

- r is the optimal number of child nodes that minimizes the $Gini_{seperation}$ index;
- n_k is the number of individuals in the child node k;

- n is the total number of individuals in the parent node :

$$n = \sum_{k=1}^{r} n_k$$

Finally, it is also possible to use another splitting criterion. Using the same mechanism, the entropy reduction process could also lead to good results. The splitting criterion is:

$$Entropy_{node} = \sum_{i=1}^{m} P_i . \log(P_i)$$

3.3.2.CHi-squared Automation Interaction Detection (CHAID)

This technique was developed in South Africa in 1980 by Gordon V. Kass.

The main characteristics of this tree are:

- It is mainly used for discrete and qualitative dependent and independent variables;
- It can be used for both classification and prediction;
- It can be interpreted easily and it is non-parametric. In other words, it does not rely on the normality distribution of data used;
- CHAID is not a binary tree which provides wider trees rather than deeper.

Yet, this technique presents also some drawbacks:

- This technique uses the X^2 (Chi-squared index) as a splitting criterion. Consequently, this test requires a significance threshold in order to work accurately. Somehow, this threshold is a parameter to be defined and its level is often arbitrary and depends on the size of the sample;
- It requires large samples and might provide unreliable results when applied on small samples.

As mentioned above, the splitting criterion of this technique is the X^2. The steps that lead to the tree are described below in the context of a credit scoring model where the dependent variable indicates whether a counterparty has defaulted or not during the year. Consequently, it has 2 categories (0 and 1, see paragraph 2.1). The independent variable is the score values:

- Each adjacent pair of the score (the independent variable) is cross-tabulated with the categories of the dependent variable. This leads to sub-tables with 2 x 2 dimensions:

		Y = Dependent Variable	
Pair 1		*0*	*1*
X = Independent Variable = Score Values	*Score 1 = S1*	$n_{1,0}$	$n_{1,1}$
	Score 2 = S2	$n_{2,0}$	$n_{2,1}$

		Y = Dependent Variable	
Pair 2		*0*	*1*
X = Independent Variable = Score Values	*Score 2 = S2*	$n_{2,0}$	$n_{2,1}$
	Score 3 = S3	$n_{3,0}$	$n_{3,1}$

		Y = Dependent Variable	
Pair 3		*0*	*1*
X = Independent Variable = Score Values	*Score 3 = S3*	$n_{3,0}$	$n_{3,1}$
	Score 4 = S4	$n_{4,0}$	$n_{4,1}$

- For each sub-table, the X^2 is computed. The general formula where the independent variable has K categories and the dependent variable has L categories is:

$$\chi^2 = \sum_{k=1}^{K}\sum_{l=1}^{L}\frac{\left(n_{kl} - \frac{n_k.n_l}{n}\right)^2}{\frac{n_k.n_l}{n}}$$

Where,

$$n_k = \sum_{l=1}^{L} n_{k,l}$$

$$n_l = \sum_{k=1}^{K} n_{k,l}$$

$$n = n_k + n_l$$

- For each sub-table, the p-value associated to the X^2 is computed. If the p-value is superior to the merging threshold (in general this threshold is set to 0.05), then the two categories are merged since this means that these categories differ the least on the response. This step is repeated until all pairs have a significant X^2. If the X^2 is significant, the categories could be split into child nodes. Moreover, if the minimum number of individuals is not respected in a node, it is merged with an adjacent node;
- Finally, the Bonferroni correction could be used to prevent the over-evaluation of the significance of multiple-category variables.

3.3.3.Quick Unbiased Efficient Statistical Tree (QUEST)

This technique was invented in 1997 by Loh and Shih.

The main characteristics of this tree are:

- It can be used for all kinds of variables (continuous, discrete, ordinal and nominal variables);
- The algorithm yields to a binary tree with binary splits;
- It has the advantage of being unbiased along with having high speed computational algorithm.

The steps of the technique are presented below in the context of a credit scoring model:

- First, as mentioned above, the independent variable is well-known. Consequently, the description of the entire Quest algorithm is unnecessary since the Anova F-test as well as Levene's F-test (and Pearson's X^2) have been used to select which independent variables are best suited in explaining the dependent variable. Knowing that in the credit scoring model only one independent variable is used, this part of the algorithm might be skipped;
- Second, the Split criterion is based upon the Quadratic Discriminant Analysis (QDA) instead of the Linear Discriminant Analysis for the FACT algorithm. The process used is as follows:
  - In credit scoring models, the final tree is generally constituted of multiple classes (in general from 5 to 10 classes). Consequently, knowing that more than 2 classes are required, the first split is based on separating the portfolio into two super-classes. The two-mean clustering method is used to do so. The purpose is to create two classes where the means are the most distant. Consider that the sample is constituted of n observations. These observations are grouped into 2 classes by optimizing the Euclidean distance to the mean of each class :

$$\min \sum_{i=1}^{2} \sum_{x_j \in C_i} |x_j - m_i|^2$$

    Where x_j are the observations in the class C_i and m_i the mean of the class C_i

  - Then a QDA is performed on these two super-classes in order to the find optimal splitting point. The QDA makes the assumption that the population in the two classes is normally distributed. Consequently, the normal densities are considered and the split point is determined as the intersection between the two Gaussian curves. Consequently, the equation to be resolved is :

$$f(x, m_1, \sigma_1) = \frac{1}{\sqrt{2\pi}.\sigma_1}.e^{-\frac{(x-m_1)^2}{2.\sigma_1}} = f(x, m_2, \sigma_2) = \frac{1}{\sqrt{2\pi}.\sigma_2}.e^{-\frac{(x-m_2)^2}{2.\sigma_2}}$$

    Where m is the mean and σ the standard deviation of the class (1 or 2).

    By applying the logarithm function on the equation we obtain:

$$\log\left(\frac{1}{\sqrt{2\pi}.\sigma_1}.e^{-\frac{(x-m_1)^2}{2.\sigma_1}}\right) = log\left(\frac{1}{\sqrt{2\pi}.\sigma_1}\right) - \frac{(x-m_1)^2}{2.\sigma_1} = log\left(\frac{1}{\sqrt{2\pi}.\sigma_2}\right) - \frac{(x-m_2)^2}{2.\sigma_2}$$

$$-log(\sigma_1) - \frac{1}{2.\sigma_1}(x^2 + m_1^2 - 2xm_1) = -log(\sigma_2) - \frac{1}{2.\sigma_2}(x^2 + m_2^2 - 2xm_2)$$

Finally, the quadratic equation becomes:

$$\left(\frac{1}{2.\sigma_2} - \frac{1}{2.\sigma_1}\right)x^2 + \left(\frac{m_1}{\sigma_1} - \frac{m_2}{\sigma_2}\right)x + \left(\frac{m_2^2}{2.\sigma_2} - \frac{m_1^2}{2.\sigma_1} + log\left(\frac{\sigma_2}{\sigma_1}\right)\right) = 0$$

- This step is repeated until the maximal tree length is obtained and the pruning algorithm is activated to optimize the tree.

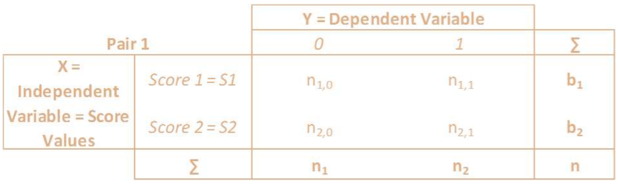
3.3.4.Belson criteria

The characteristics of the Belson criteria technique are:

- It is used to optimize the discretization by building classes that optimizes the correlation between the dependent and the independent variable;
- It does not yield to a binary tree;
- It might provide trees that are wide;
- No thresholds (for splitting or merging) are necessary for this technique.

The steps that lead to the segmentation are:

- This algorithm is very similar to the CHAID algorithm in terms of process and steps of segmentation. What really changes is the splitting criterion;
- In fact, the Belson criterion is a simplified version of the X^2 test and it is very useful when the variables are dichotomous. Somehow in the credit scoring model the dependent variable is dichotomous. The steps leading to the segmentation contain a contingency table with 2 x 2 dimension making the independent variable dichotomous at each iterative calculus;
- More precisely, the algorithm is :



		Y = Dependent Variable		
Pair 1		*0*	*1*	Σ
X = Independent Variable = Score Values	*Score 1 = S1*	$n_{1,0}$	$n_{1,1}$	b_1
	Score 2 = S2	$n_{2,0}$	$n_{2,1}$	b_2
	Σ	n_1	n_2	n

The Belson criterion is computed for each pair:

$$Crit_B = \left|n_{1,0} - \frac{b_1}{n}n_1\right| = \left|n_{2,0} - \frac{b_2}{n}n_1\right| = \left|n_{1,1} - \frac{b_1}{n}n_2\right| = \left|n_{2,1} - \frac{b_2}{n}n_2\right|$$

Where,

$$b_1 = n_{1,0} + n_{1,1}$$

$$b_2 = n_{2,0} + n_{2,1}$$

$$n_1 = n_{1,0} + n_{2,0}$$

$$n_2 = n_{1,1} + n_{2,1}$$

$$n = n_1 + n_2 = b_1 + b_2$$

The splitting point is the one that maximizes the Belson Criterion.

3.4. Conditions and constraints to optimize the segmentation

In a credit scoring model, while building the risk classes, there are some conditions and constraints that must be fulfilled in order to have a robust model that will allow measuring accurately the risks within the portfolio.

In fact, regulators and central banks pay a lot of attention on the way risk classes have been constructed. Moreover, for each risk manager having a rating scale that provides accurate results is a must have since it will allow him to fully understand the structure of the portfolio as well as drive him to take actions if the risk limits are not respected.

The conditions and constraints that must be taken into account are:

- **Number of individuals in each class:**
  - First, it is important that the number of individuals in each class does not exceed a certain limit since it will show that the discrimination power of the rating scale is not sufficient enough. In general, the size of a class must not exceed 40% of the total portfolio;
  - Second, the minimum number of individuals within a class must not be too low. In fact, this might introduce instability in the rating scale since a little change in the population might have a significant impact on the default rate. Moreover, a class with few individuals might not be sufficiently consolidated or not sufficiently representative of a homogenous risk class. In general, each class must at least have more than 100 counterparts.

- **Homogeneity within a class / heterogeneity between classes :**
  - First, the aim of the rating scale is to build classes that are homogenous in terms of risk profile. Consequently, the intra-class dispersion must be low in order to ensure that the individuals that are grouped into a risk class share the same characteristics as well as the same risk profile. If this dispersion is high, a recalibration is needed;
  - Second, the level of risk must be different from a class to another. In fact, the rating scale purpose is to build clusters where each cluster represents a different

level of risk. Consequently, the dispersion between classes must be high showing that these classes are differentiated in terms of risk profile.

- **Stability of the classes :**
  - First, the classes must guarantee a stability of the population across time (or a giving horizon). This is very important since it will affect the predictive power of the model. Consequently, indicators such as the Population Stability Index (PSI) must be performed and monitored across time ;
  - Second, the rating scale must also ensure that the default rate within a class is stable across time which means that the risk profile within a class is stable allowing the model to have a good predictive power;
  - Third, the stability (as well as the discrimination power) could also be appreciated by making sure that the default rates series between classes do not cross each other. This means that there are no inversions between classes making the model stable across time.

- **Discrimination and progressiveness of the default rate :**
  - The levels of default rate from one class to another must be progressive. This means that the function that represents the default rate by risk class is a monotonously increasing function;
  - This guarantees that there are no inversions between classes.

In the following part, a rating scale is built for a Retail Low Default Portfolio (LDP) using the different techniques mentioned above. When building the rating scales, the conditions and constraints mentioned above have a significant meaning since we are dealing with a LDP.

4. Building a rating scale on a Low Default Portfolio and the impacts on the regulatory capital

4.1. Framework of the study

4.1.1.Purpose of the analysis

The analysis presented in this paragraph is about building an efficient rating scale for a Low Default Portfolio and analyzing the relationship between the number of classes and the impact on regulatory capital.

In fact, a LDP has specific characteristics that present challenges for risk quantification. In fact, given the low frequency of loss events, applying a statistical treatment to build a score and a rating scale might not be an easy task. Consequently, the industry participants expressed concerns regarding LDPs since their structure could lead to unstable models which could imply that these LDPs do not meet the regulatory requirements and standards and therefore excluding them from the IRB perimeter.

The purpose of this part is to show that it is possible to build a stable rating scale that respects the conditions and constraints mentioned in paragraph 2.4 by applying a statistical treatment. Then, the optimization of RWA is also analyzed. More specifically, for a LDP, does the number of classes have a significant impact on the regulatory capital and by adjusting the number of classes is it possible to optimize efficiently the RWA?

4.1.2. What are the dimensions involved in the study?

There are three dimensions that are taken into account for the study:

The first dimension concerns the nature of the classification technique used to build the rating scale. In fact, all types of techniques mentioned above in this article are performed. The purpose is to understand how the results differ from one technique to another and where the difference comes from. Another purpose is to test which technique is best suited for LDP by providing accurate and stable results in line with the regulatory standards and risk management.

The second dimension concerns the number of classes. In fact, for each segmentation technique, different simulations are realized by taking the number of classes as an input. More precisely, the number of classes is varied until the maximum size of the tree is reached. Then the relationship between the number of classes and the impact on RWA (and thus, on regulatory capital) is studied. Another purpose is to understand how the different techniques perform while changing the number of classes and thus the depth of the tree.

The third dimension is time. The purpose here is to make sure that the rating scale built is stable across time and thus that its predictive power is good. Moreover, comparing the different techniques and their behavior across time allows a better understanding of the compatibility between the underlying algorithm of each technique and the characteristics of a LDP.

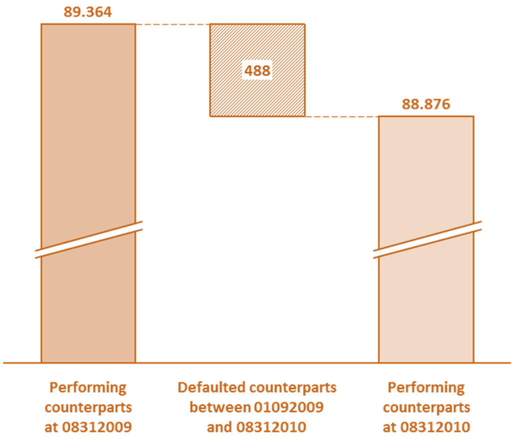
By mixing these three dimensions, the study will provide answers and solutions to the concerns raised by market participants towards the treatment of LDP and its compliance with IRB requirements.

4.1.3. Description of the portfolio on which the study was conducted

As mentioned above, the portfolio on which the study was conducted is a Mortgage portfolio. This portfolio is characterized by a low frequency of loss events. The modeling window is the date 08312009 which means that all performing loans at 08312009 are considered. These loans are analyzed from 09012009 to 08312010 by trying to identify which counterparts defaulted. Consequently it is the one year default rate that is modeled.

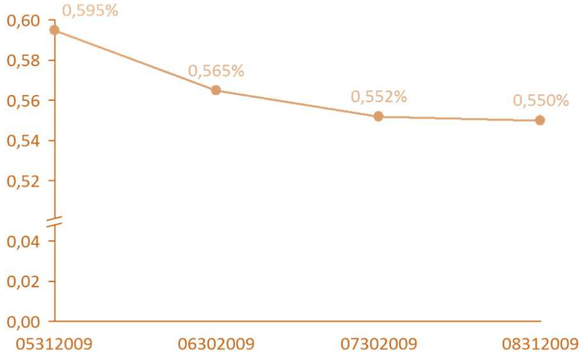
For the purpose of the study, the score has already been constructed, based on the best practices in the industry. The score shows good performances even though the portfolio considered is a LDP. The methodology used to build the score is not the core subject of the study. In fact, the focus is on the rating scale and therefore the methodology of the score won’t be detailed here.

The distribution of the number of counterparts in the portfolio is:



The one-year default rate at 08312009 is 0.55%. The number of counterparties is acceptable making the portfolio sufficiently granular.

The evolution of the default rate on different time windows is:



The default rate series presents a slight decreasing tendency.

4.2. Presentation of the results

4.2.1. Building the rating scale on the modeling sample

The first step of the study is to build the rating scale on the modeling sample. As mentioned above, the modeling window is 08312009. The different techniques have been performed on this sample by changing the depth of the tree and the number of classes. The different results are summarized in the following matrix:

Distribution of default rate by number of classes and segmentation technique

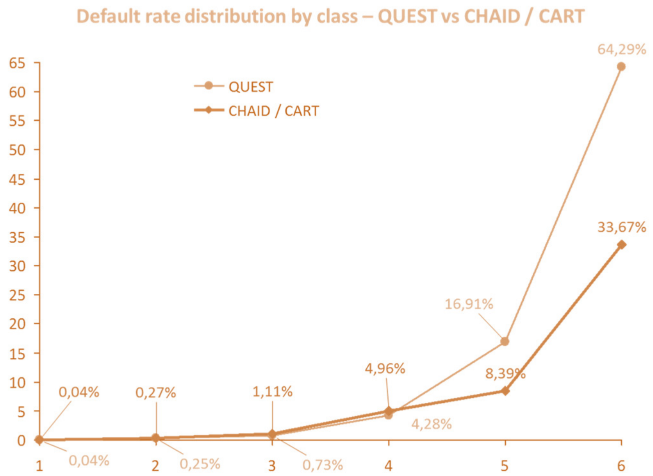
Nb of classes	QUEST	CHAID	CART	BELSON	ISO interval	ISO distrib.
1	0,4%	0,4%	0,0%	0,1%	0,3%	0,0%
2	20,8%	33,7%	1,0%	3,0%	12,2%	1,0%
1		0,3%	0,0%	0,1%	0,2%	0,0%
2		6,0%	0,3%	1,1%	3,1%	0,1%
3		33,7%	2,9%	9,2%	33,3%	1,5%
1	0,1%		0,0%	0,1%	0,1%	0,0%
2	1,6%		0,3%	1,1%	1,2%	0,1%
3	16,9%		1,1%	5,6%	9,2%	0,2%
4	65,5%		8,7%	22,0%	47,6%	1,9%
1	0,0%	0,1%		0,1%	0,1%	0,0%
2	0,2%	1,1%		0,4%	0,5%	0,1%
3	0,7%	5,0%		1,1%	4,8%	0,1%
4	4,3%	8,4%		5,6%	19,7%	0,3%
5	20,8%	33,7%		22,0%	64,3%	2,2%
1	0,0%	0,0%	0,0%		0,0%	0,0%
2	0,2%	0,3%	0,3%		0,3%	0,0%
3	0,7%	1,1%	1,1%		2,2%	0,1%
4	4,3%	5,0%	5,0%		7,6%	0,1%
5	16,9%	8,4%	8,4%		26,0%	0,4%
6	64,3%	33,7%	33,7%		66,0%	2,6%
1		0,0%	0,0%		0,0%	0,0%
2		0,3%	0,3%		0,2%	0,0%
3		1,1%	1,1%		1,1%	0,1%
4		5,0%	5,0%		5,6%	0,1%
5		8,4%	8,4%		12,2%	0,3%
6		23,7%	23,7%		32,1%	0,4%
7		58,1%	58,1%		63,0%	3,0%
1				0,1%	0,0%	0,0%
2				0,4%	0,2%	0,0%
3				0,8%	0,6%	0,0%
4				2,1%	3,4%	0,1%
5				5,5%	6,6%	0,1%
6				12,1%	20,7%	0,3%
7				23,3%	40,2%	0,4%
8				54,4%	61,9%	3,3%

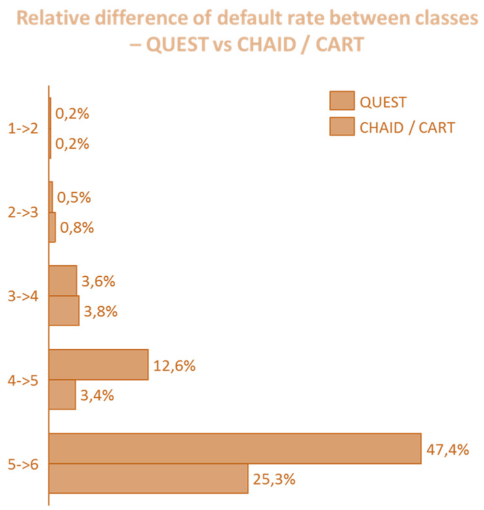
The first finding concerns the variety of results obtained. In fact, the rating scale and the clustering might differ significantly from one technique to another. More precisely, it is easy to observe that the distribution of the default rate differs on the technique used. This is also true when the number of classes vary. There is only one exception and it concerns the CART and the CHAID procedure. In fact, from a certain tree depth, the results are identical for both trees. For instance, for 6 and 7 classes, the distribution of the default rate by class is identical.

The second finding shows that the statistical algorithms (QUEST, CART, CHAID and Belson) could not always build a tree with a specified number of classes. For instance, the CHAID algorithm was not able to build a rating scale with 4 classes. This could be explained by the iterative process of the algorithm as well as the splitting criterion. In fact, the input for a tree is its depth. This depth represents the maximal number of iterations. In fact, a CART algorithm with a specified depth of n iterations yields to a tree with a maximum number of classes of 2^n (since CART is binary). For instance, a CART tree with 3 iterations will yield to a tree with 7 classes or 8 classes (maximum number of classes). This means that one of the solutions is unlikely. Moreover, the study shows that the possible number of classes vary from one tree to another. In fact, the QUEST could not yield to a tree with 3 classes whereas the CHAID could not yield to a tree with 4 classes. Moreover, CART could not yield to a tree with 5 classes whereas Belson Criteria could not yield to a tree with 6 classes. This shows that the behavior of statistical algorithms could yield to various results when applied on a LDP. This could be explained by the sensitivity of statistical indicators (splitting criterion) when the frequency of loss events is low.

The third finding shows that the depth of the tree is limited. In fact, for the statistical methods, it was not possible to build trees with more than 8 classes while respecting the constraints presented above (and more specifically the one concerning the minimum number of individuals in a class). In fact, considering the low number of defaults in a LDP, the algorithms are unable to build trees that both optimize the splitting criterion as well as respecting the minimum number of individuals. Consequently, when building a rating scale for a LDP, it is sometimes challenging to respect the minimum number of classes as mentioned by the regulators (a minimum of 7 classes). This point is even more important given the relationship between the number of classes and the impact on regulatory capital.

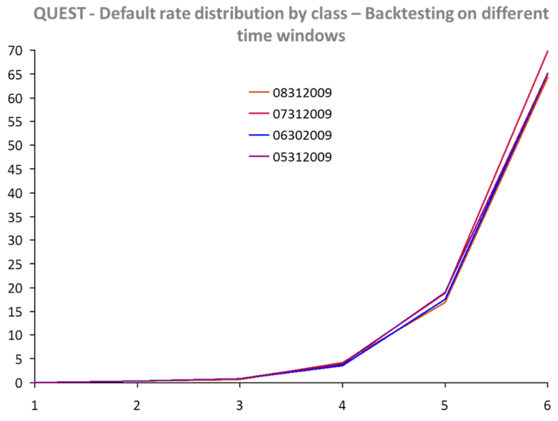
The fourth finding shows that it is possible to build a rating scale on LDPs using statistical algorithms as well as respecting the constraints described above. In fact, the rating scales built allow respecting the number of individuals in a node or class. More precisely, for statistical algorithms this constraint was not always respected. This is why the human judgment is important. In fact, to solve this problem, it is possible to manually merge adjacent classes in order to build a sufficiently consolidated class. This is a common practice in the industry. Moreover, as the study shows, the discrimination of individuals and the progressiveness of default rates are also respected. In fact, there a no inversions observed. Nonetheless, it is important to pinpoint the fact that the discrimination power, measured here by the levels of default rates in each class, is different from one technique to another. Clearly, the discrimination power is low for the iso-distribution technique. It is obviously much more important for the QUEST algorithm. In fact, by comparing the default rates obtained for 6 classes between QUEST and CHAID, the study shows that the progressiveness of default rates is slower on the first 4 classes for QUEST, and is it is more important on the last 2 classes:

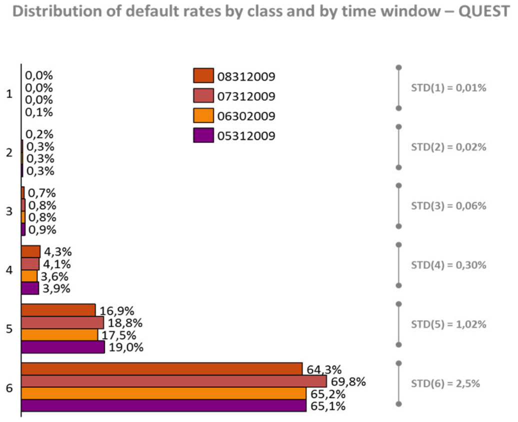

Default rate distribution by class – QUEST vs CHAID / CART

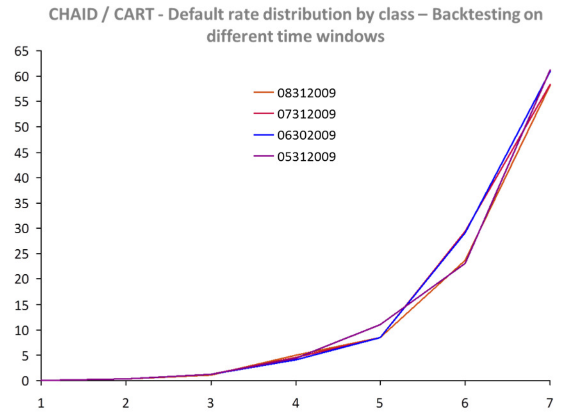

Relative difference of default rate between classes – QUEST vs CHAID / CART

4.2.2. Verifying the stability of the rating scale

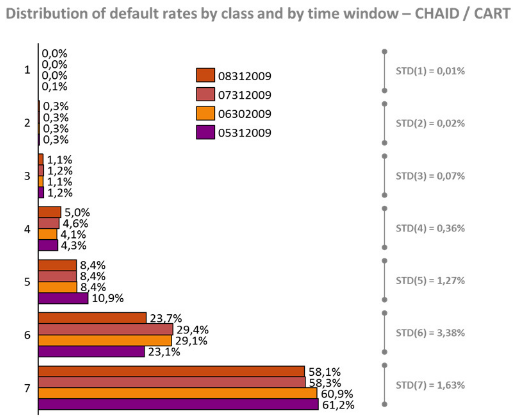
Once the rating scale is built on the modeling sample, its stability across time is analyzed. In fact, one of the major challenges when building a rating scale, and especially on a LDP, is to have stable results and avoid volatility as well as disturbance. Consequently, out-of sample tests are performed and the stability of default rates in each class is analyzed as well as the number of inversions between classes. The study has been performed on the rating scales with the maximum number of classes (6 classes for QUEST, 7 for CHAID and CART, 8 for BELSON and 8 for iso-distribution).


QUEST - Default rate distribution by class – Backtesting on different time windows

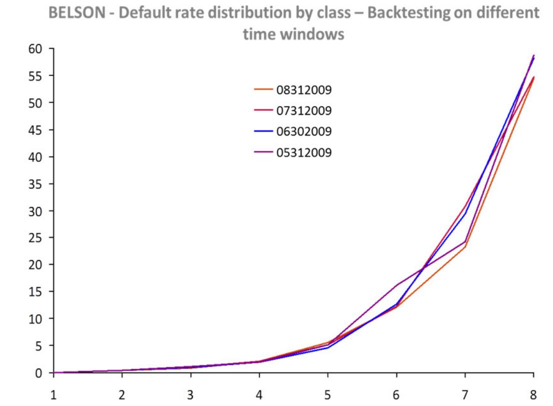

Distribution of default rates by class and by time window – QUEST



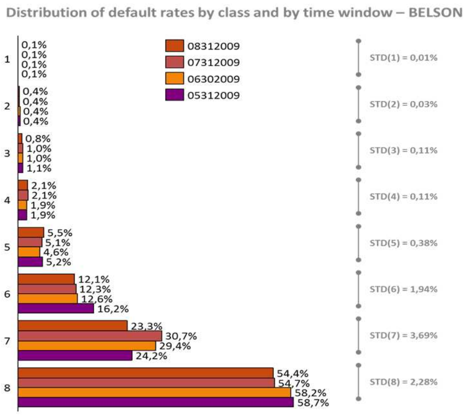
CHAID / CART - Default rate distribution by class – Backtesting on different time windows



Distribution of default rates by class and by time window – CHAID / CART



BELSON - Default rate distribution by class – Backtesting on different time windows

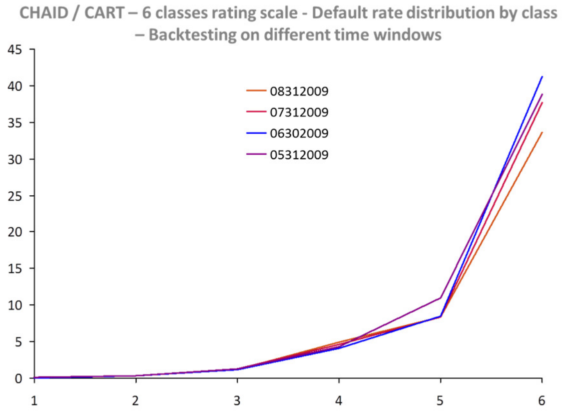


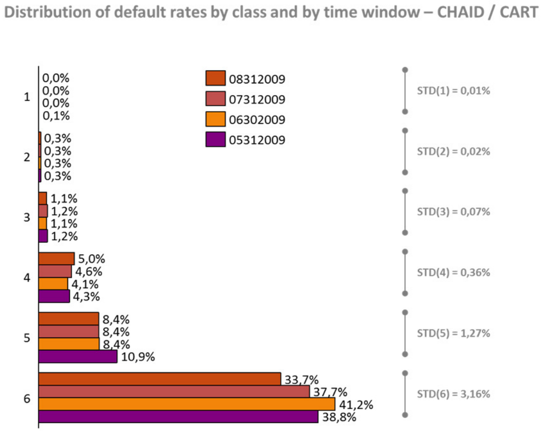
Distribution of default rates by class and by time window – BELSON

The study shows that the rating scales provide stable results in time. In fact, as the graphs above show, the distribution of default rates by class is relatively stable across time. The degree of stability depends on the technique used. Moreover, the discrimination power on the application (or validation) sample is good since there are no inversions between risk classes.

Yet, the comparison between the different classification techniques provides interesting results. In fact, the study shows that the QUEST algorithm provides the best results in terms of stability. In fact, the distribution of default rates on the different time windows is narrower even though a gap is observed on the last class (the 6th class) for the time window 07312009. For the other out-of-sample time windows (06302009 and 05312009), default rates are stable (65.2% and 65.1% respectively, in comparison to 64.3% on the modeling sample). Moreover, the standard deviations of default rates for each class are the lowest for the QUEST algorithm in comparison to the other techniques showing a higher degree of stability and precision in measuring risk levels across time.

Finally, one can question whether the performances of the QUEST algorithm in comparison to CHAID, CART and BELSON are related to the number of classes (6 classes for QUEST vs 7 for CHAID / CART and 8 for BELSON) and to the fact that the depth of the QUEST rating scale is lower than the other techniques making it more compact which might introduce a bias in the conclusions. Consequently, a study has been conducted by comparing the performances of the algorithms on the same number of classes:





The graphs above show the distribution of default rates using the CHAID / CART rating scale, with 6 classes. This study shows that the stability and homogeneity are not better in comparison with the 7-class rating scale. In fact, the dispersion of the class 6 is relatively high in comparison with the results obtained for the QUEST algorithm.

4.2.3. Establishing a relationship between the number of classes and the impact on regulatory capital

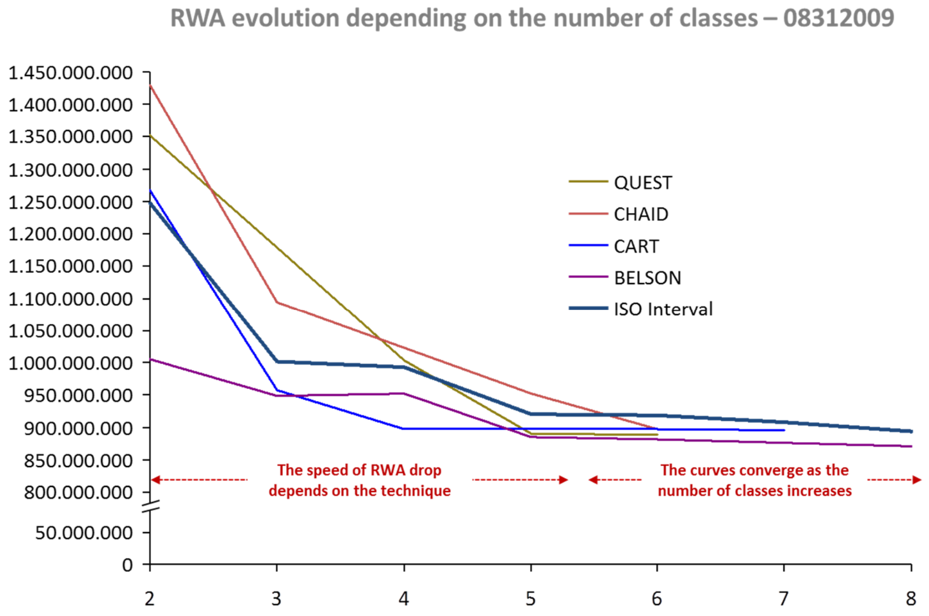
The purpose of this paragraph is to establish a relationship between the number of classes within a rating scale and the impact on regulatory capital. This relationship is quite important in times where the number of regulatory constraints is getting higher with important pressure on banks' capital. Consequently, optimizing the Risk Weighted Assets (RWA) is a key challenge in times where banks try to reach acceptable levels of pre-crisis ROE.

To establish this relationship, a RWA simulation has been conducted. The Exposure-At-Default (EAD) of the portfolio has been measured by considering that each loan has the same EAD. This hypothesis was preferred to the real EAD of the portfolio since it gives a similar weight to each loan and consequently assumes that the portfolio is perfectly granular. Knowing this, if the real portfolio was taken into account, the conclusions will be unchanged meaning that this hypothesis does not affect the conclusions of the study.

Consequently, each loan is supposed to have an EAD of 50 K€, which, by experience, is the average EAD in a mortgage portfolio. The EAD of the portfolio is then 4,468.2 M€. The simulation's results are:

RWA Evolution in €

Nb of classes	QUEST	CHAID	CART	BELSON	ISO Inter
2	1 353 002 207	1 429 770 868	1 267 430 967	1 006 010 969	1 248 486 555
3		1 093 752 831	958 332 588	949 965 394	1 002 168 236
4	1 004 689 499		897 623 004	952 481 099	993 520 579
5	891 161 468	953 158 707		885 518 321	920 448 561
6	888 269 923	898 215 593	898 215 593		918 323 611
7		896 594 095	896 594 095		907 526 622
8				871 872 815	893 343 772

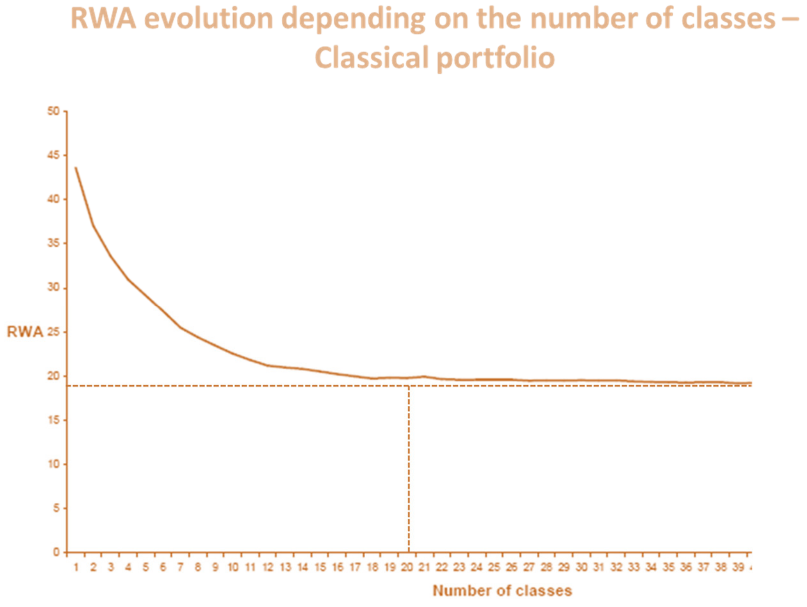


The first finding of this study is that the RWA decrease with the number of classes. In fact, the more the number of classes is high the less the RWA are and consequently the regulatory capital.

The second finding is that the slope of the curve (which is negative, reflecting the negative correlation between RWA and the number of classes) depends on the segmentation technique. In other words, the speed of the RWA drop is different from one technique to another. Nonetheless, the common point is that the curves converge as the number of classes increases. In fact, as the study shows, the results of the different technique tend to be closer with the number of classes increasing.

The third finding is that there is a point where the slope of the curve becomes close to 0. In other words, more than just getting closer with the increasing number of classes, the curves converge to a certain limit. This shows that RWA do not decrease indefinitely with the number of classes.

The last finding concerns the point where the decrease in RWA becomes limited. In fact, this point is reached rapidly for this portfolio, which is a LDP. As the study shows, the decrease in RWA becomes too small starting from 6 classes. This is a characteristic of a LDP which does not necessarily apply on classical portfolios. In fact, the same study was conducted on a classical portfolio showing that the curve linking the RWA to the number of classes also presents a negative slope and converges to a certain limit. Nonetheless, the point from which this limit is reached (or at least considered as unchanged) is higher than a LDP:



RWA evolution depending on the number of classes – Classical portfolio

As the graph above shows, the limit is reached for 20 classes. This means that the opportunities of RWA optimization are higher for a classical portfolio than a LDP.

Conclusion

As LDPs are characterized by low frequency of loss events, building a rating scale based upon statistical algorithms might be challenging. Yet, this study showed that it is possible to meet regulatory requirements for these portfolios. In fact, the QUEST algorithm might be best suited for LDPs since it provides better results in terms of discriminatory power, stability, and robustness. Finally, the negative correlation between the number of risk classes and RWA showed a path for RWA optimization. Yet, these opportunities are less significant for LDPs but might have significant impact for classical portfolios with a decent number of defaults. A point to bear in mind with an increasing pressure on ROE and regulatory capital.

Bibliographie

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.*

Kass, G. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics, Vol. 29, No. 2.*, 119-127.

Loh, W.-Y., & Shih, Y.-S. (1997). Split Selection Methods for Classification Trees. *Statistica Sinica, Vol. 7*, 815-840.

Rakotomalala, R. (2005). Arbres de décision. *Modulad,33*, 163-187.

Rakotomalala, R. *Les méthodes d'induction d'arbres, CHAID, CART, C4, C5.*

Ritschard, G. (2010). *CHAID and Earlier Supervised Tree.*

Rokach, L., & Maimon, O. (2010). *Data Mining and Knowledge Discovery Handbook.* Springer.

Fovea. *La segmentation (2010).*

Stéphane Tufféery, *Data Mining and Statistics for Decision Making, University of Rennes, France*

Techniques of Decision Trees Induction, Chapter 2.

Lior Rokah, Oded Maimon, *Data Mining And Knowledge Discovery Handbook, Decision Trees, Chapter 9.*

Jerome Frugier, Benoit Genest, *Optimization of Post-Scoring Classication Methods and their Impact on the Assessment of Probabilities of Default and on RegulatoryCapital.* 2005