



# Modélisation Risque de Crédit

Probabilité de Défaut Baloise

**Intervenant**

Aryan Razaghi

**MoSEF Data Science**

# Plan



1. Probabilité de Défaut – Notions
2. Données en risque de crédit
3. Différenciation du risque
4. Quantification du risque



# 1. Probabilité de Défaut – Notions

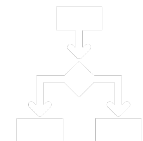
# 1.1. Définition de la Probabilité de Défaut

- La probabilité de défaut est une mesure utilisée dans le domaine de la réglementation bancaire pour évaluer le risque de non-remboursement d'un emprunteur ou d'une contrepartie. C'est une estimation de la probabilité qu'un emprunteur ou une contrepartie ne puisse pas honorer ses obligations de paiement.
- La réglementation considère la probabilité de défaut comme une mesure de la probabilité d'occurrence d'un défaut sur une contrepartie à un horizon donné (un an).
- La probabilité de défaut est utilisée comme base pour estimer les paramètres de risque et les exigences en fonds propres.
- Les taux de défaut sont souvent utilisés comme mesure directe de la probabilité de défaut.

## 1.2. Motif de défaut en Banque

### Définition du défaut :

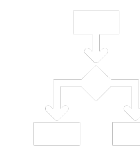
Une contrepartie entre en défaut selon la réglementation baloise dès que l'une des trois conditions suivantes est vérifiée :

- 1.L'arriéré du débiteur sur un crédit important dépasse 90 jours.
- 2.Une détérioration significative de la situation financière de la contrepartie faisant douter de sa capacité à rembourser en totalité son crédit, sans que des mesures appropriées telles que la réalisation d'éventuelles garanties soient vérifiées. Cette première condition de nature subjective constitue un jugement d'expert.
- 3.La contrepartie fait l'objet d'une procédure judiciaire en cours.

## 1.3. Taux de défaut

- **Le Taux de Défaut**, pour une cohorte  $T$ , est défini comme le ratio entre les contrats sains en  $T$  et qui font défaut entre  $T$  et  $T+12$  et les contrats sains en  $T$  :

$$TD_T = \frac{\text{Contrats sains en } T \text{ qui font défaut entre } T \text{ et } T + 12}{\text{Contrats sains en } T}$$





## 2. Données en risque de crédit

## 2.1. Nature des données / Stock

Une source dites stock archive pour chaque arrêté de fin de mois l'ensemble des informations de types signalétique. Ces données permettent aux institutions financières d'identifier et de comprendre leurs clients, de suivre leurs comportements financiers et de prévoir les risques potentiels.

On peut retrouver :

### **L'historique des emprunteurs :**

Les données signalétiques archivées permettent de suivre l'historique des emprunteurs, y compris leurs précédents prêts, leur comportement de paiement, et toute tendance concernant l'utilisation responsable du crédit.

### **Données personnelles et financières :**

Ces données stockées peuvent inclure des informations détaillées sur les emprunteurs, telles que les revenus, l'emploi, l'état civil, etc.



## 2.2. Nature des données / Flux

Les établissements financiers utilisent généralement une cinquantaine de paramètres quantifiés et pondérés par leur système de notation interne. Ces données sont alimentée en vision flux.

Les différentes données utilisées peuvent être regroupées en 3 catégories :

### **Données financières :**

La notation d'une entreprise passe d'abord par une saisie et une retranscription des informations comptables, disponibles dans les comptes annuels de la société. Le système de notation interne utilise ces données pour calculer des ratios financiers s'articulant notamment autour des fondamentaux financiers : autonomie financière, niveau d'endettement, gestion de l'exploitation et gestion de la liquidité ; ainsi que sur les performances commerciales, économiques et financières de l'entreprise : croissance de l'activité, niveau de marge et rentabilité.

## 2.2. Nature des données / Flux

### **Données comportementales :**

Ces données concernent notamment le fonctionnement courant du compte, le niveau d'utilisation des lignes de crédit, les incidents de paiement et les données sur les risques fournies par la Banque de France. La collecte de ces informations permet d'analyser le fonctionnement courant de la société afin d'en détecter les vulnérabilités potentielles.

### **Données qualitatives :**

La pertinence des informations qualitatives repose principalement sur la complétude de la connaissance du client (le KYC, Know Your Customer) et de sa mise à jour. Les critères qualitatifs s'articulent généralement autour des thématiques suivantes : la qualité et la fiabilité de l'information financière, le marché et le positionnement stratégique de l'entreprise, l'actionnariat et le management, l'exposition aux risques.



### **3. Différenciation du risque**



# 3.1 Processus de Différenciation du risque

Le processus de modélisation est composé des 5 étapes suivantes :

1. Analyse descriptive des variables
2. Découpages des variables quantitatives et qualitatives
3. Étude de la relation entre les variables
4. Estimation du modèle
5. Construction d'une grille de score

## 3.2 Études des variables



La première étape de la modélisation est d'effectuer une analyse univariée des variables explicatives. Cette étape permet d'identifier, pour chacune des variables à disposition le pourcentage de valeurs :

- 1. Non applicables** : certaines valeurs non renseignées sont dues au fait que la variable ne s'applique pas à l'individu (par exemple, la variable peut concerner un produit non détenu par le client).
- 2. Manquantes** : Les valeurs non renseignées qui n'ont pas d'explication métier.
- 3. Aberrantes** : Les valeurs qui présentent un écart anormal par rapport aux autres valeurs observées dans un échantillon tiré aléatoirement. Les seuils à partir desquels une variable est considérée comme aberrante sont définis en relation avec les équipes métiers.
- 4. À exclure** : Pour certaines variables, il existe des valeurs non aberrantes mais qu'il est tout de même nécessaire d'exclure car elles nécessitent un traitement spécial.

## 3.3. Discrétisation



### Objectifs :

- Des effets non linéaires peuvent alors être pris en compte dans la modélisation.
- La stabilité du modèle construit est accrue du fait du regroupement en quelques modalités.
- Les valeurs extrêmes se retrouvent groupées dans des classes, leur influence sur la modélisation devient alors limitée.
- Les valeurs manquantes peuvent être regroupées dans la modalité souhaitée.
- La compréhension et la lecture du modèle final sous forme de grille de score sont facilités.

## 3.3. Discrétisation

### Conditions :

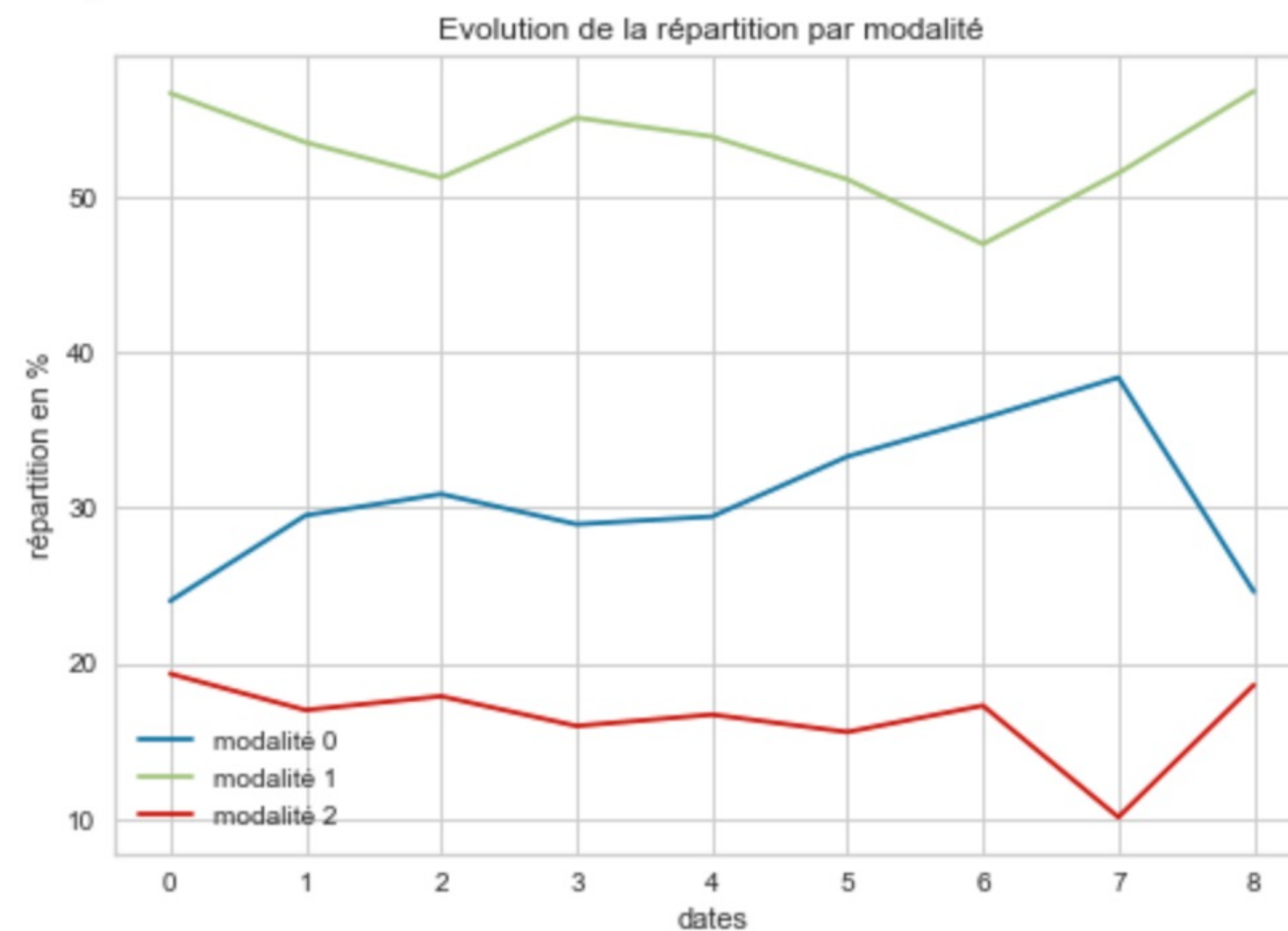
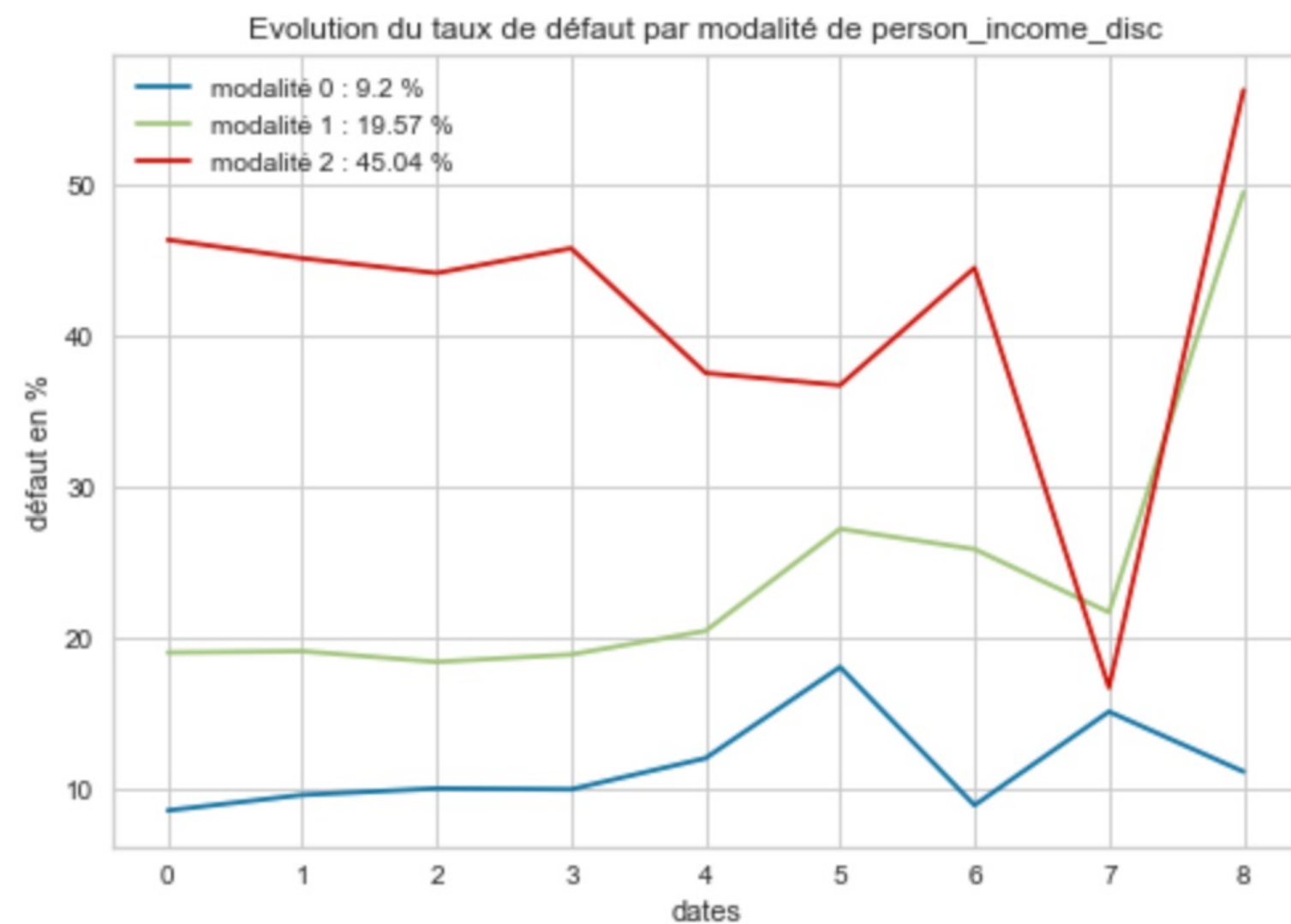
- Un faible nombre de modalités afin de garder une certaine aisance de lecture (max 4, 5).
- Chaque modalité doit contenir au moins 5% des observations.
- Chaque modalité doit rester stable en risque et en volume.
- La mesure de stabilité est calculée à l'aide d'un *Indicateur de Stabilité* par variable :

$$IS = \sum_k (p_k - b_k) * \ln \left( \frac{p_k}{b_k} \right) \geq 0$$

où  $p_k$  et  $b_k$  sont les deux différentes proportions de la modalité  $k$  que l'on cherche à comparer.



## 3.3. Discrétisation





## 3.3. Discrétisation



### Algorithmes de discrétisation :

1. Clustering hiérarchique
2. Chi-merge
3. Optimum Chi2
4. Minimum Description Length Principle

## 3.4. Études des corrélations



### Méthodes :

- Une fois la discrétisation réalisée, on effectue un test de corrélation entre nos variables explicatives, afin de palier au problème de multi-colinéarité.
- Pour vérifier la corrélation entre les variables discrétisées, et entre les variables explicatives et la variable à expliquer, on effectue le test d'indépendance du  $\chi^2$ .
- Le test d'indépendance du  $\chi^2$  consiste à déterminer si la valeur observée d'une variable dépend de la valeur observée d'une autre variable.

## 3.5. Régression logistique / Problématique

- Le but d'un modèle de Probabilité de Défaut est de pouvoir prédire au mieux la probabilité de réalisation d'un critère binaire (le défaut) sur un horizon  $H$  à partir de données disponibles à l'instant  $T$ .
- Pour cela on utilise la régression Logistique. Cette méthode permet de répondre aux 3 objectifs inhérents à tout modèle de classification :
  - 1. Précision:** Les prédictions du modèle se doivent de correspondre le plus possible au risque de défaut associé aux contreparties étudiées.
  - 2. Robustesse:** Les modèles servant à estimer le risque associé à chaque contrepartie ne doivent pas être trop sensibles à la volumétrie.
  - 3. Interprétabilité:** Les modèles doivent être facilement compréhensibles et interprétables

## 3.5. Régression logistique / Problématique

- Soit  $Y = (Y_1, Y_2, \dots, Y_m)^t$ , le vecteur des variables réponses des  $m$  individus observés.  $Y$  est une variable aléatoire binaire, indiquant si un individu est défaillant ou non (la valeur 1 représente le défaut).
- Soit  $j$  tel que  $1 \leq j \leq m$  et soit  $X_j = (X_{j,1}, X_{j,2}, \dots, X_{j,k})^t$ , l'ensemble des caractéristiques des  $d$  variables explicatives sélectionnées de chaque individu  $j$ .
- L'objectif est de réussir à prédire l'appétence au défaut d'un individu  $z$  dont les caractéristiques sont connues.
- Il faut estimer la probabilité de défaut  $Y = 1$  sachant les valeurs prises par les variables  $X_1, X_2, \dots, k$ .
- Ainsi, nous voulons modéliser, pour un individu  $z$  dont les caractéristiques  $x_z = (x_1, \dots, x_k)$  sont connus, la probabilité a posteriori suivante :

$$\pi(z) = P(Y = 1 | X = x_z)$$

- La variable  $Y | X = x_z$  suit une loi de Bernoulli de paramètre  $p = P(Y | X = x)$  qui est le paramètre à estimer. Le modèle Logit est adapté pour modéliser la probabilité  $\pi$

## 3.5. Régression logistique / Formule

- La régression logistique est une technique qui permet d'exprimer le logit de  $\pi(z)$  sous la forme d'une combinaison linéaire des variables  $X_1, X_2, \dots, X_d$
- c'est-à-dire nous cherchons des constantes,  $\beta_0, \beta_1, \dots, \beta_k$  telles que :

$$\forall z, \quad \text{logit}(\pi(z)) = \ln \left( \frac{\pi(z)}{1 - \pi(z)} \right) = \beta_0 + \beta_1 X_{z,1} + \beta_2 X_{z,2} + \dots + \beta_k X_{z,k} = X_z \beta$$

- Avec  $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)$  et  $X_z = (1, X_{z,1}, \dots, X_{z,d})^t$ . Il est facile de retrouver à partir de l'équation précédente :

$$\pi(z) = P(Y = 1 | X = x_z) = \frac{\exp(X_z \beta)}{1 + \exp(X_z \beta)}$$

## 3.5. Régression logistique / Estimation

$$\ln \left( \frac{\pi(z)}{1 - \pi(z)} \right) \beta_0 + \beta_1 X_{z,1} + \beta_2 X_{z,2} + \dots + \beta_k X_{z,k} = X_z \beta$$

- Les constantes  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$  sont estimées par maximum de vraisemblance. La log-vraisemblance associée à l'échantillon de  $m$  individus est la suivante.

$$\mathcal{L}(Y, \beta) = \sum_{z=1}^m (Y_z \ln(\pi(z)) + (1 - Y_z) \ln(1 - \pi(z))) = \sum_{z=1}^m (Y_z (X_z \beta) + \ln(1 - \exp(X_z \beta)))$$

- L'objectif est de maximiser la log-vraisemblance, pour cela il faut optimiser les coefficients  $\beta$  de manière à avoir :

$$\frac{\partial \mathcal{L}}{\partial \beta} = 0$$

L'algorithme le plus souvent utilisé est celui de Newton Raphson

## 3.5. Régression logistique / Test de Wald

- Pour vérifier la significativité des coefficients d'une régression logistique, on applique le test de Wald.
- Pour tester la nullité simultanée de  $q$  coefficients, nous utilisons la généralisation de la statistique de WALD  $W_{(q)}$ . Elle suit une loi du  $\chi^2$  à  $q$  degrés de liberté.

$$W_{(q)} = \hat{a}'_{(q)} * (\hat{V}_{(q)})^{-1} * \hat{a}_{(q)}$$

Avec :

$\hat{a}_{(q)}$  le vecteur des valeurs observées des coefficients que l'on souhaite tester

$\hat{V}_{(q)}$  la sous matrice de variance covariance associé à ces coefficients.

## 3.5. Régression logistique / Sélection de variables



- Dans le cadre du risque de crédit on se basera sur une approche de sélection de variables basée sur les critères statistiques.
- En effet, la sélection de variables va nous permettre d'éliminer les variables n'ayant aucune ou très peu d'influence sur le défaut ou encore de choisir la variable la plus discriminante parmi plusieurs variables transmettant plus ou moins la même information.
- Trois méthodes sont souvent utilisées en risque de crédit :
  1. Forward
  2. Backward
  3. Stepwise



## 3.5. Régression logistique / Sélection de variables



### Forward :

- Cette méthode consiste à sélectionner les variables explicatives étape par étape en partant d'un modèle à 1 variable jusqu'à obtenir  $i$  variables.
- À chaque étape, la variable explicative ayant la statistique du  $\chi^2$  la plus élevée est sélectionnée parmi les variables restantes.
- Si aucune variable n'a une statistique du  $\chi^2$  assez élevée, la sélection s'arrête.

## 3.5. Régression logistique / Sélection de variables



### Backward :

- Cette méthode part d'un modèle composé de toutes les variables à disposition et élimine à chaque étape la variable la moins significative par rapport au test de Wald jusqu'à obtenir  $k$  variables.
- Si toutes les variables obtiennent des valeurs au-dessus d'un certain seuil de significativité, aucune variable n'est éliminée et la sélection s'arrête.

## 3.5. Régression logistique / Sélection de variables



### Stepwise :

- La méthode Stepwise consiste à sélectionner à chaque étape une variable à l'aide de la méthode Forward tout en regardant si une des variables précédemment sélectionnées n'a pas perdu en significativité avec l'ajout d'autres variables (méthode Backward) si c'est le cas la variable est retirée du modèle.
- Cette méthode nous indique une combinaison de variables possibles qui va nous permettre de construire un modèle à  $p$  variables en combinant les résultats statistiques fournis par la méthode et nos connaissances métiers.

## 3.5. Grille de score

- Chacune des grilles de score créées regroupe les informations suivantes et donne une note sur 1000 au client :
  - Variables explicatives.
  - Classes des variables explicatives.
  - P-Value associée au test de Wald pour chaque classe.
  - Note normée attribuée à chaque classe.
  - La contribution de la variable.
  - Taux de défaut de chaque classe.
  - Effectif de chaque classe.

## 3.5. Grille de score



- **Règles d'acceptations**
  - P-Value associée à chaque classe : Toutes les classes doivent être significatives au seuil de 5%. Le non respect de cette règle induit une existence de corrélation résiduelle entre classes et/ou variables
  - Le signe des coefficients doit être cohérent.

## 3.5. Grille de score

- **Calcul des pondérations pour chaque modalité (Note)**

$$N_i^j = \frac{|\max(\beta_i^1, \dots, \beta_i^p) - \beta_i^j|}{\sum_{i=1}^k \max(\beta_i^1, \dots, \beta_i^p) - \min(\beta_i^1, \dots, \beta_i^p)} * 1000$$

- $\beta_i^j$  le coefficient estimé par la régression logistique de la modalité  $j$  de la variable  $i$
- $\max(\beta_i^1, \dots, \beta_i^p)$  le coefficient maximum de la variable  $i$
- $\min(\beta_i^1, \dots, \beta_i^p)$  le coefficient minimum de la variable  $i$
- $p$  le nombre de modalités de la variable  $i$
- $k$  le nombre de variables dans le modèle
- $N_i^j \in [0; 1000]$  : La note 0 étant le profil de risque le moins risqué et 1000 le plus risqué.

## 3.5. Grille de score

- **Calcul des contributions des variables**

$$c_i = \frac{\sqrt{\sum_{j=1}^p r_j \left(N_i^j - \overline{N_i^j}\right)^2}}{\sum_{i=1}^k \sqrt{\sum_{j=1}^p r_j \left(N_i^j - \overline{N_i^j}\right)^2}}$$

- $\overline{N_i^j}$  la note moyenne de la variable  $i$
- $r_j$  la part de la population avec la modalité de la variable  $i$
- $p$  le nombre de modalités de la variable  $i$
- $k$  le nombre de variables dans le modèle

# 3.5. Grille de score



Variable	Classe	P-value / Significativité	Note	Contribution	Taux de défaut en %	Effectif de chaque classe en %
X1	Modalité 1 . . Modalité 3	*** . . **		%	% . . %	% . . %
X2	Modalité 1 . . . Modalité 4	***   *		%	% . . . %	% . . . %
X3	Modalité 1 . . Modalité 2	**   ***		%	% . . %	% . . . %
X4	Modalité 1 . . . Modalité 4	*   **		%	% . . %	% . . . %



## 3.6. Indicateurs de performances



### Métriques

- Courbe ROC : Courbe représentant l'arbitrage entre vrai et faux positifs.
- Indice de Gini : Indicateur compris entre 0 et 1 qui rend compte de la répartition d'une variable dans une population. Dans un score, on s'attend qu'il soit proche de 1, signe que les défauts sont concentrés dans les classes les plus risquées.

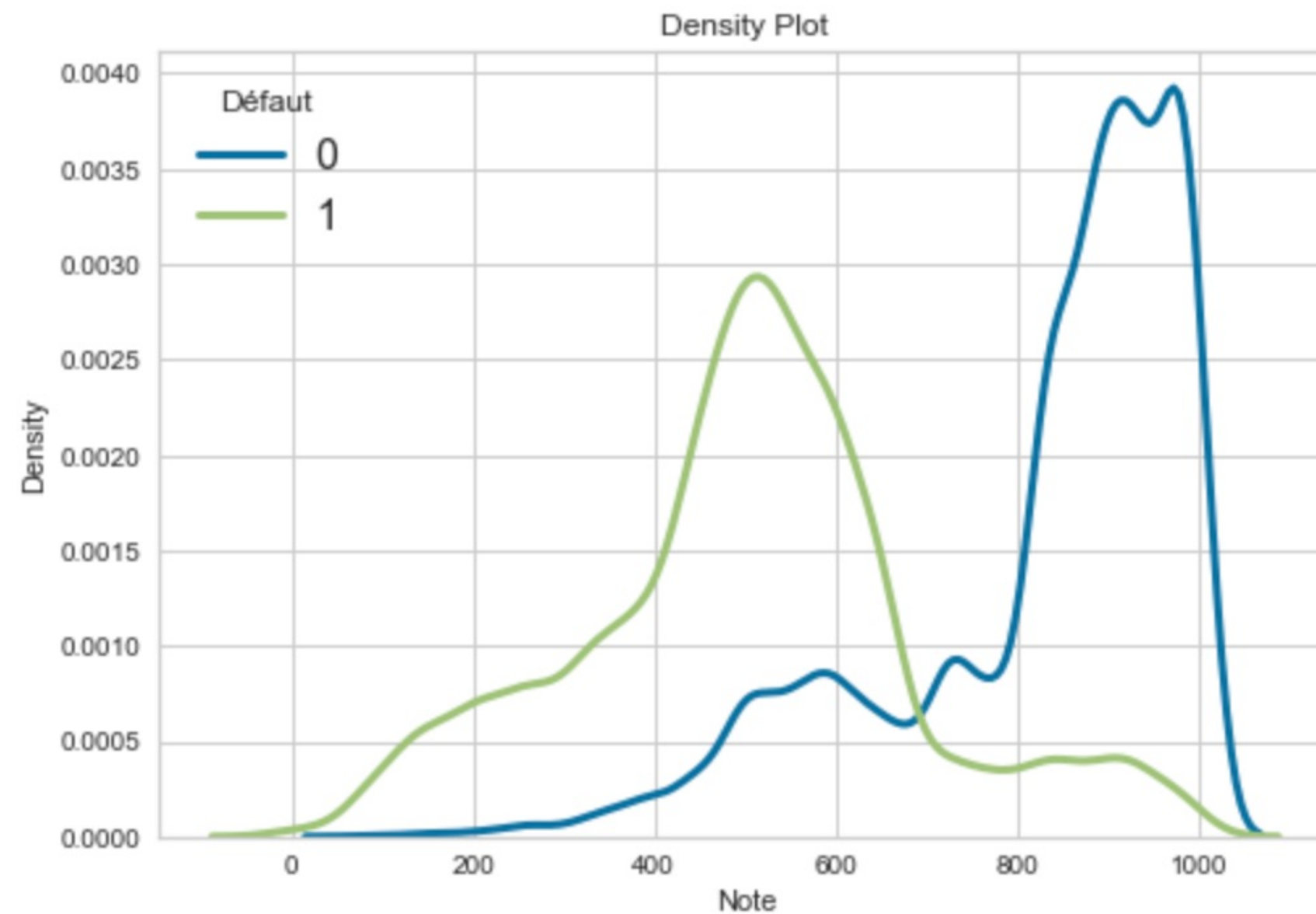
$$Gini = 2 * AUC - 1.$$

## 3.6. Indicateurs de performances



### Densités conditionnelles

- Distribution des scores des individus conditionnellement au défaut.
- Plus les distributions sont éloignées, plus le score est discriminant.

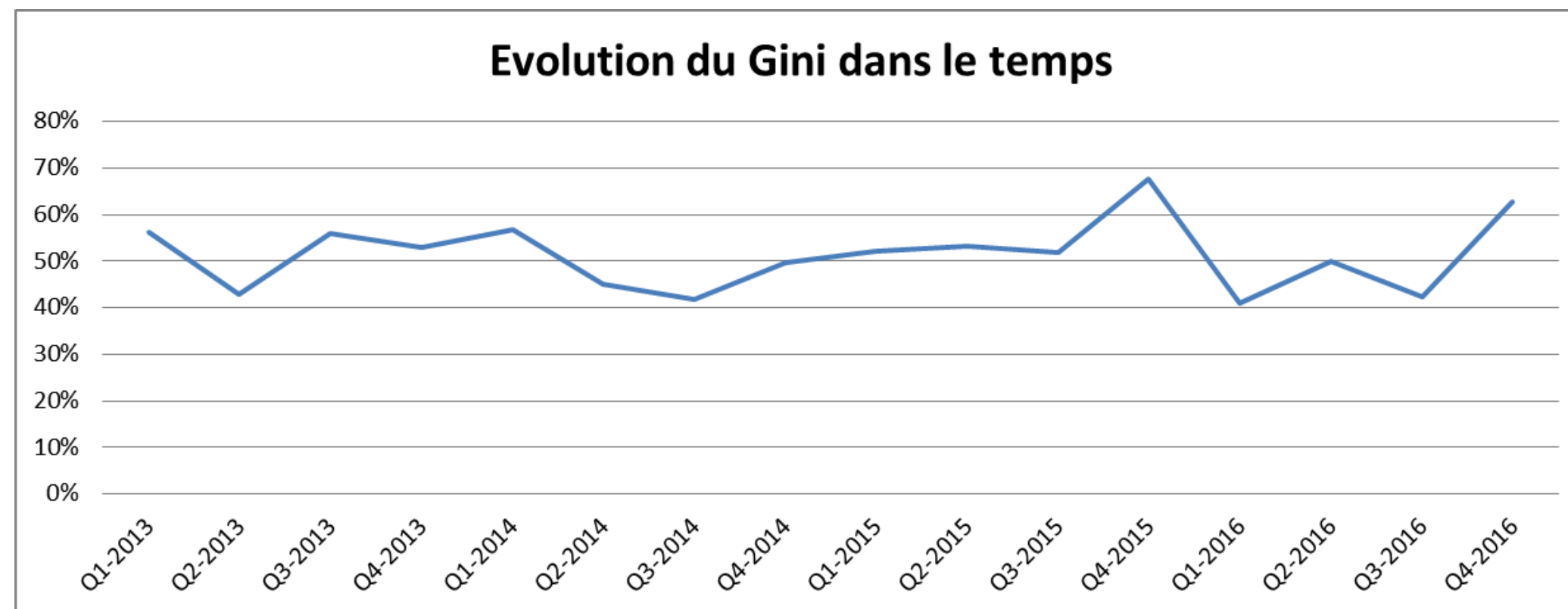


## 3.6. Indicateurs de performances



### Stabilité de la performance dans le temps

- Le score construit doit faire l'objet d'une étude afin de vérifier qu'il est robuste dans le temps : un score est robuste (stabilité temporelle) s'il est indépendant du temps.
- Il est ainsi nécessaire de s'assurer de la stabilité des performances dans le temps.





## 4. Quantification du risque

# 4.1. Segmentation



## Ce que dit l'EBA :

Article 69 EBA guidelines : “institutions should check the **homogeneity** of obligors or exposures assigned to the same grades or pools. In particular, grades should be defined in such a manner that each obligor within **each grade or pool has a reasonably similar risk of default**”.

Article 97 EBA guidelines : “Institutions may split exposures covered by the same PD model into **as many different calibration segments as needed** where one or more subsets of these exposures carry a significantly different level of risk”.

Les institutions doivent isoler au maximum tous les profils présentant des niveaux de risque significativement différents.

Pour construire des classes homogènes de risque, les institutions doivent s'assurer que les contrats rassemblés au sein d'une même classe ne constituent pas un trop large éventail de niveau de risque.

# 4.1. Segmentation



- Une fois la grille de scores défini, celle-ci doit être associée à une échelle de notation afin de créer les CHR (Classes Homogènes de Risques).
- Il existe plusieurs algorithmes permettant de construire les CHR : les arbres de décisions, les algorithmes génétiques, Jenks natural breaks optimization...
- Pour faire une bonne segmentation il faut respecter plusieurs contraintes réglementaires :
  1. Homogénéité au sein de chaque classe
  2. Hétérogénéité entre les classes
  3. Eviter une concentration excessive au sein de chaque classe (max 30%)
  4. Assurer une augmentation régulière des taux de défaut à mesure que l'on progresse d'une classe à l'autre
  5. ...

## 4.2. Calibrage



- Les PD sont estimées au sein de chaque CHR. Chaque note de la fonction de score est ensuite associée à un niveau de PD.

$$PD_{LRA_{CHR_i}} = LRA_{CHR_i} + MOC_C + MOC_A + MOC_B$$

où la  $LRA$  est donné par la relation suivante :

$$LRA_{CHR_i} = \frac{\sum DR_{YEAR\ N,CHR_i}}{T}$$

Avec :

$DR_{YEAR\ N,CHR_i}$  représente le taux de défaut TTC au sein de chaque CHR.

$MOC_C$  ,  $MOC_A$  et  $MOC_B$  représente les marges de conservatisme.

## 4.2. Calibrage



- Pour ce faire, les établissements doivent classer les incertitudes identifiées selon les catégories suivantes :
  - A. Incertitudes liées aux données et à la méthodologie : données manquantes ou aberrantes, les évolutions des seuils de matérialité du défaut, etc...
  - B. Incertitudes liées aux changements dans les procédures, à l'appétence au risque, à la politique de recouvrement et à d'autres sources d'incertitudes additionnelles.
  - C. Incertitudes liées aux erreurs d'estimation.
- La marge de conservatisme finale doit correspondre à la somme des marges associées à chaque catégorie d'incertitude.



## 4.3. Calibrage / MOC C



- la MOC pour l'erreur d'estimation générale (EEG) devrait refléter la dispersion de la distribution de l'estimateur statistique et doit être appliqué à tout modèle.
- Le modèle mathématique est certainement la meilleure estimation pour le paramètre cible, mais il reste une approximation avec une dispersion et une erreur sous-jacente.
- Le GEE est là pour couvrir cette erreur d'estimation et plus précisément l'erreur de prédiction après l'étape de calibrage.
- Si les CHRs sont définis, la MOC C doit être calculée au niveau de la classe. Si la sortie du modèle est une valeur continue et est utilisée comme telle, le calcul du MOC devrait être effectué au niveau du segment de calibrage à condition qu'il ne soit pas significativement différent entre les grades de PD.

## 4.3. Calibrage / MOC C



- La méthodologie proposée est une estimation asymptotique de la variance des paramètres. Le calcul du MOC est basé sur l'utilisation du 90e percentile du paramètre prédit sur un ensemble de données bootstrapp.
- Les étapes sont les suivantes :
  1. Construction de 1000 échantillons aléatoires avec remise de l'échantillon de calibrage. Le nombre d'observations dans les échantillons devrait être le même que dans l'ensemble de données de calibrage.
  2. Calcul de la LRA au niveau du segment de calibrage pour chaque sous-échantillon.
  3. Calcul des 90e centiles des 1000 valeurs de LRA.
  4. Le MOC est calculé comme la différence entre le 90e percentile de la moyenne bootstrappée et la valeur de la moyenne bootstrappée.