Université Paris-Est Créteil

Master 2 Analyse et Modélisation des Risques

Crédit Scoring LCL
Score sur des sociétés civiles immobilières

Auteurs :
Dieudonné Malonga
Jamy Courtois

Professeur :
Alexandre Parant

20 janvier 2026

Résumé

Ce projet s'inscrit dans le cadre de la modélisation du risque de crédit pour le LCL, avec pour objectif la refonte de la grille de score destinée à la clientèle professionnelle des Sociétés Civiles Immobilières (SCI). Le portefeuille étudié, qualifié de « low default » (taux de défaut de 0,88 %), présente un fort déséquilibre de classes qui complexifie l'identification des contreparties risquées.

Nous avons déployé une méthodologie rigoureuse pour construire un modèle de régression logistique, standard de l'industrie bancaire reconnu pour son interprétabilité et sa stabilité. La démarche a inclus la sélection des variables prédictives, leur discrétisation et l'analyse des contributions au score. Bien que le modèle final affiche une performance globale satisfaisante sur l'échantillon d'apprentissage (D de Sommer de 0,761), l'analyse approfondie des résultats met en évidence certaines limites, notamment une perte de pouvoir discriminant sur les segments de clientèle présentant les scores les plus élevés.

Mots-clés : Crédit Scoring, LCL, SCI, Régression Logistique, D de Sommer, Courbe ROC, Low Default Portfolio.

Table des matières

Introduction

Dans un contexte bancaire marqué par le renforcement des exigences réglementaires, notamment avec l'entrée en vigueur de Bâle IV et la nouvelle définition du défaut (NDoD), la maîtrise du risque de crédit est un enjeu stratégique majeur pour LCL. Ce projet a pour objectif principal de développer un modèle de notation (score de risque) permettant d'évaluer la probabilité qu'un client fasse défaut à un horizon de 12 mois. Au-delà de la simple prédiction, ce score vise à constituer un outil d'aide à la décision pour l'octroi de crédit, à permettre une tarification adaptée au risque et à assurer une allocation optimale des fonds propres réglementaires.

Le périmètre de cette étude se concentre sur une catégorie spécifique de la clientèle professionnelle : les Sociétés Civiles Immobilières (SCI). L'unité statistique analysée est donc la personne morale. Pour mener à bien ces travaux, nous disposons de la base de données « sasuser.base_sci » comportant 100 013 observations. Celles-ci couvrent un historique allant de janvier à décembre 2023 pour la période d'observation, enrichi par des données collectées jusqu'à 12 mois en amont. Cette base regroupe des variables signalétiques sur l'ancienneté et le secteur d'activité, des données comportementales liées aux mouvements et incidents sur comptes, ainsi que des informations sur les engagements du client (actif, passif et hors bilan).

Le critère à modéliser, représenté par la variable cible « DDefaut_NDB », est la tombée en défaut bâlois sur une fenêtre de performance de 12 mois suivant la date d'observation. Conformément aux normes de l'EBA appliquées depuis 2021, ce défaut est matérialisé soit par des arriérés de paiement supérieurs à 90 jours consécutifs dépassant les seuils de matérialité (100 € et 1 % des engagements), soit par des signes probables d'absence de paiement (Unlikely to Pay) tels que des restructurations pour risque ou des procédures collectives.

La méthodologie adoptée repose sur la construction d'une régression logistique, choisie pour sa robustesse et son interprétabilité. Afin d'assurer la validité du modèle, la base a été scindée via un échantillonnage aléatoire stratifié sur la variable cible et la date, répartissant les données en un échantillon d'apprentissage de 70 % et un échantillon de test de 30 %. Le modèle final permettra de construire une grille de score et une échelle de notation.

Construction de la base d'analyse

2.1 Description et prise en main des données

L'étude s'appuie sur la table SAS `sasuser.base_sci`. L'analyse descriptive réalisée via la procédure `contents` permet de caractériser la structure de ce jeu de données.

La base comporte 100 013 observations et 41 variables. Outre l'identifiant client et la date d'observation, ces variables se répartissent en 10 variables continues et 29 variables catégorielles. Les informations disponibles se regroupent en quatre grandes familles :

- les informations client (ancienneté de l'activité, secteur d'activité) ;
- les données de mouvements bancaires (flux, soldes, incidents) ;
- les données d'engagement (actif, passif, hors bilan) ;
- l'historique de défaut bâlois du client.

La variable cible à modéliser est `DDefaut_NDB`, correspondant à la survenance d'un défaut bâlois sur un horizon de 12 mois.

2.2 Stratégie d'échantillonnage

Afin de construire et valider le modèle de score, la base a été scindée en deux parties distinctes : un échantillon d'apprentissage (70 % des données) et un échantillon de test (30 % des données).

Ce découpage a été réalisé à l'aide de la procédure `surveyselect`, en appliquant un tirage aléatoire simple sans remise (méthode SRS). Pour garantir la représentativité des échantillons, le tirage a été stratifié sur deux axes via l'instruction `strata` :

- la date d'observation (`DATDELHIS`) ;
- la variable cible (`DDefaut_NDB`).

Cette stratification assure que la structure temporelle et le taux de défaut initial sont rigoureusement conservés à l'identique dans la base d'apprentissage (`b_train`) et la base de test (`b_test`).

Etude exploratoire

L'étude exploratoire a pour objectif de comprendre la structure du portefeuille, de gérer et traiter les valeurs manquantes, extrêmes voir incohérentes, de vérifier la stabilité du risque dans le temps et d'identifier les variables explicatives les plus pertinentes pour la modélisation. Cette étape vise à fiabiliser la base d'analyse avant la modélisation. Les traitements sont définis sur l'échantillon d'apprentissage (`b_train`) puis répliqués strictement sur l'échantillon de test (`b_test`).

3.1 Nettoyage des données

3.1.1 Variables Quantitatives

L'analyse des fréquences sur les variables numériques révèle une structure de données manquantes très localisée.

TABLE 3.1 – Analyse des valeurs manquantes (Variables Quantitatives)

Variable	Nb Manquants	% Manquants	Décision
ANC_RELA_LCL	18	0.03%	Suppression
MVT_AFF_12M	18	0.03%	Suppression
NBJDEPDP	18	0.03%	Suppression
SOLD_CRE	18	0.03%	Suppression

Nous constatons que ces 18 valeurs manquantes concernent systématiquement les mêmes individus. L'absence d'information sur des variables structurelles comme l'ancienneté relationnelle (`ANC_RELA_LCL`) rend ces observations inexploitables. Compte tenu du très faible volume (0,03 % de l'échantillon) et de l'impossibilité de justifier ces absences par une logique métier, nous procédons à la suppression de ces observations.

3.1.2 Analyse des distributions et valeurs aberrantes

Nous avons réalisé une étude statistique descriptive pour l'intégralité des variables quantitatives présentes dans notre base de données. Toutefois, nous choisissons d'afficher uniquement les résultats de la variable NJRS_DEP_DA car celle-ci présente une anomalie particulière. En effet, l'observation d'un minimum négatif de -27 suggère la présence de valeurs aberrantes, ce qui nécessite une attention spécifique par rapport aux autres variables de l'étude.

TABLE 3.2 – Statistiques descriptives pour la variable NJRS_DEP_DA

Indicateur	Moyenne	Q. inf.	Médiane	Q. sup.	Ecart-type	Coef. var.	Skewness	Kurtosis	Min	Max
Valeur	2,65	0,00	0,00	0,00	10,59	399,82	5,42	33,07	-27,00	92,00

Le -27 s'avère être la seule observation négative contenue dans cette variable. Cette observation isolée a été supprimée de la base d'apprentissage.

L'étude des statistiques descriptives (`PROC MEANS`) permet de valider la cohérence des données et d'identifier les valeurs extrêmes. Les statistiques descriptives des autres variables quantitatives sont disponibles en Annexe 1 dans le tableau 9.1. Concernant ces variables, celles de comptage des jours débiteurs, la médiane et le 3ème quartile sont fréquemment nuls, ce qui témoigne d'une clientèle majoritairement saine en termes de trésorerie.

Concernant les montants, nous observons un écart significatif entre la moyenne et la médiane. Cela reflète l'hétérogénéité des chiffres d'affaires des clients professionnels et non des erreurs de saisie. Les distributions présentent un *skewness* et un *kurtosis* élevés, indiquant que la relation avec la variable cible ne sera probablement pas linéaire. Cette caractéristique sera traitée ultérieurement par la discrétisation des variables dans le modèle logistique.

3.1.3 Variables Qualitatives

Les variables catégorielles, notamment celles liées à l'historique des incidents sur les 6 derniers mois (Impayés, Dépassements, NDB), présentent un taux de valeurs manquantes plus élevé.

TABLE 3.3 – Valeurs manquantes sur les variables dynamiques

Variable	Nb Manquants	% Manquants
Depassement_M1	621	0,89 %
Depassement_M2	1156	1,65 %
Depassement_M3	1708	2,44 %
Impaye_M1	621	0,89 %
Impaye_M2	1156	1,65 %
Impaye_M3	1708	2,44 %
NDB_6	3420	4,88 %

Les variables catégorielles pour lesquelles nous observons des valeurs manquantes sont celles pour lesquelles nous avons introduit de la dynamique. Il s'agit de variables pour lesquelles nous observons des valeurs jusqu'à six mois avant le suivi de la tombée en défaut bâlois sur douze mois. Le fait que des valeurs manquantes puissent exister pour certaines observations se justifie par l'hypothèse suivante : ces individus ont une ancienneté inférieure à un an chez LCL. Dans le cas où l'ancienneté serait inférieure aux périodes couvertes par les variables retardées, cela expliquerait que nous ne disposions pas de ces informations pour ces individus. Ici, le manque d'information est à considérer comme une information en soi, ces variables ne sont donc pas à supprimer. Elles peuvent renfermer des données relatives à l'absence de connaissance propre au comportement en termes de tendance aux incidents. Pour gérer ces valeurs manquantes, nous allons leur imputer la valeur de 3 pour les variables Impaye et Depassement (M1, M2, M3) et la valeur de 2 pour les variables NDB.

3.1.4 Nettoyage des variables non discriminantes

Les variables qualitatives ne présentant qu'une seule modalité sur l'ensemble de la base ne peuvent pas contribuer à discriminer le risque. Elles sont donc retirées.

TABLE 3.4 – Variables exclues (Variance nulle)

Variable	**Modalité unique**	**Justification**
Top_sci	1 (Oui)	Périmètre de l'étude (Base SCI)
Top_pret_perso	0 (Non)	Non pertinent pour cette population
Top_pp	0 (Non)	Population Personne Morale uniquement

Enfin, une vérification via `PROC SQL` confirme l'absence totale de doublons dans la base nettoyée (`train_cleanf`).

3.1.5 Application à l'échantillon de test

Afin de garantir la validité externe du modèle, l'ensemble des règles de gestion définies ci-dessus a été appliqué à l'identique sur l'échantillon de test (`b_test`) :

- Suppression des observations ayant des données signalétiques manquantes (6 observations, soit 0.02%).
- Imputation des modalités spécifiques pour les historiques courts.
- Suppression des variables constantes.

Les contrôles finaux sur la base de test (`test_cleanf`) confirment l'absence de valeurs aberrantes (aucun montant d'engagement négatif, aucun jour débiteur négatif).

3.2 Statistiques descriptives

3.2.1 Statistiques descriptives univariées

Nous opérons sur un portefeuille de clients qualifié de Low Default Portfolio (LDP), au sein duquel la proportion d'individus en défaut à 12 mois s'élève à 0,88 % sur l'ensemble de notre base de données. Ce résultat souligne une maîtrise satisfaisante du risque de défaut par LCL sur le segment de la clientèle SCI.

Toutefois, nous faisons face à un déséquilibre des classes particulièrement prononcé. Cette information est déterminante dans le cadre de la modélisation, où notre objectif est de prédire avec précision les entrées en défaut. Nous pourrions être confrontés à un biais d'apprentissage lors de l'évaluation des métriques de performance ; le modèle risquerait alors de se concentrer sur la juste prédiction de la classe majoritaire (les individus sains), limitant ainsi sa capacité à identifier la classe minoritaire, pourtant critique pour la gestion du risque. Nous retrouvons la proportion suivante au sein de notre échantillon d'apprentissage :

TABLE 3.5 – Proportion des classes de la variable cible : DDefaut_NDB

DDefaut_NDB	Fréquence	Pourcentage
0	69 393	99.12 %
1	616	0.88 %
Total	**70 009**	**100.00 %**

La variable cible `DDefaut_NDB` correspond à la survenance d'un défaut bâlois à horizon 12 mois. L'analyse des fréquences sur l'échantillon d'apprentissage révèle la distribution suivante :

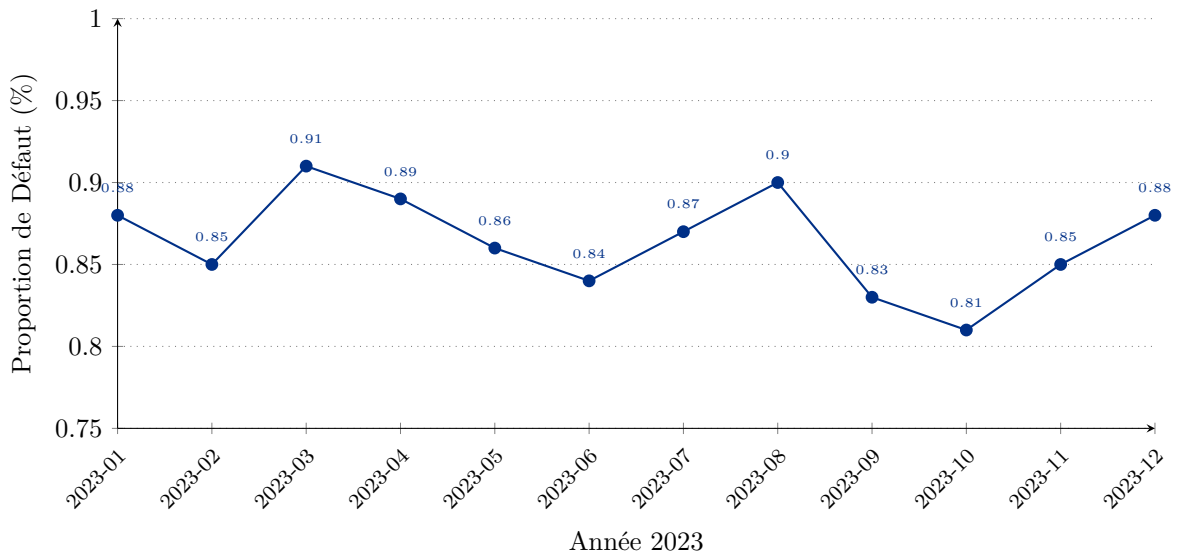


FIGURE 3.1 – Suivi de la proportion de défaut par date d'originition du prêt en 2023 (portefeuille clientèle SCI)

Le taux de défaut par date d'origination des prêts sur l'année 2023 fluctue dans une fourchette étroite comprise entre 0,81 % et 0,91 %, soit une variation marginale de 0,1 point de pourcentage. À l'échelle de l'individu, cette amplitude s'avère extrêmement réduite, confirmant ainsi la stabilité du taux de défaut bâlois à 12 mois au regard des crédits accordés durant l'année 2023.

Par ailleurs, nous pouvons affirmer qu'il n'existe aucun effet de saisonnalité notable quant à la sinistralité observée selon le mois d'origination du prêt. Cette homogénéité temporelle renforce la fiabilité de nos indicateurs de risque sur l'ensemble du cycle de production annuel.

3.2.2 Statistiques descriptives bivariées - étude des liaisons

TABLE 3.6 – Test de Kruskal-Wallis : Étude de la liaison avec DDefaut_NDB

Variable	Chi-deux	Degré de liberté	Pr > Chi-deux
NB_JR_DEB	856.42	1	<.0001
NJRS_DEP_DA	812.15	1	<.0001
SOLD_DIB	645.30	1	<.0001
NBJDEPDP	589.12	1	<.0001
SOLD_CRE	412.05	1	<.0001
MVT_AFF_12M	305.44	1	<.0001
SFMois_AG	254.18	1	<.0001
ANC_RELA_LCL	158.33	1	<.0001
NBJRDB_AT	112.09	1	<.0001
ANC_ENTR	6.22	1	0.0126
Engagement_prorat	0.01	1	0.9200

Dans un premier temps, afin d'évaluer le pouvoir discriminant de nos variables quantitatives conditionnellement aux modalités de notre variable cible, nous mettons en place le test de Kruskal-Wallis. L'hypothèse nulle (H_0) de ce test stipule que les médianes de chaque groupe sont identiques. Si cette hypothèse se confirme pour certaines variables testées, cela traduit une absence de pouvoir discriminant vis-à-vis des modalités de notre variable cible. L'hypothèse alternative (H_1) suppose quant à elle que les médianes des deux groupes sont statistiquement différentes.

L'étude du tableau Liaison_cible_quanti nous permet d'affirmer que nous ne pouvons pas rejeter l'hypothèse nulle pour la variable Engagement_prorat, quel que soit le seuil de significativité retenu (p-valeur $= 92\,\%$). En conséquence, il apparaît certain que cette variable n'intégrera pas notre modèle final.

Nous observons également une p-valeur légèrement supérieure à $1\,\%$ pour la variable ANC_ENTR ($p = 0,0126$). Selon que nous choisissions un seuil de risque à $5\,\%$ ou à $1\,\%$, l'hypothèse nulle peut être rejetée ou conservée. Étant donné le nombre important de variables candidates au modèle, nous décidons de mettre de côté ANC_ENTR en raison de l'incertitude sur la robustesse de son pouvoir discriminant vis-à-vis de la variable DDefaut_NDB.

Pour les variables restantes, nous rejetons l'hypothèse nulle au seuil de $1\,\%$. En examinant les statistiques du χ^2, nous constatons que les quatre valeurs les plus élevées concernent des variables associées au débit, notamment le solde débiteur et le nombre de jours débiteurs. Cette observation est cohérente avec l'analyse métier, car ces indicateurs constituent historiquement des signes avant-coureurs qui précèdent souvent des défauts de paiement ponctuels.

Pour les clients rencontrant les plus grandes difficultés, ces incidents évoluent ensuite vers une situation de défaut au sens de Bâle. Le pouvoir discriminant de ces variables vis-à-vis de l'entrée en défaut est donc validé tant sur le plan statistique que d'un point de vue économique et historique.

TABLE 3.7 – Matrice de corrélation de Pearson : Variables Quantitatives

Variables	ANC_R	ANC_E	ENG	MVT	NBJ	NBJR	NB_J	NJRS	SFM	CRE	DIB
ANC_RELA_LCL	1.00	0.79	-0.19	0.00	-0.01	-0.04	-0.10	-0.09	0.06	0.06	0.01
ANC_ENTR	0.79	1.00	-0.19	0.01	0.00	-0.04	-0.08	-0.07	0.07	0.06	0.01
Engagement_prorat	-0.19	-0.19	1.00	0.20	-0.01	0.08	0.05	0.02	0.08	0.08	-0.08
MVT_AFF_12M	0.00	0.01	0.20	1.00	-0.01	0.04	0.00	-0.01	0.26	0.47	-0.02
NBJDEPDP	-0.01	0.00	-0.01	-0.01	1.00	0.01	0.36	0.39	-0.02	-0.01	0.01
NBJRDB_AT	-0.04	-0.04	0.08	0.04	0.01	1.00	0.36	0.01	-0.02	0.01	-0.32
NB_JR_DEB	-0.10	-0.08	0.05	0.00	0.36	0.36	1.00	0.93	-0.05	-0.04	-0.12
NJRS_DEP_DA	-0.09	-0.07	0.02	-0.01	0.39	0.01	0.93	1.00	-0.05	-0.04	0.00
SFMois_AG	0.06	0.07	0.08	0.26	-0.02	-0.02	-0.05	-0.05	1.00	0.73	0.01
SOLD_CRE	0.06	0.06	0.08	0.47	-0.01	0.01	-0.04	-0.04	0.73	1.00	0.00
SOLD_DIB	0.01	0.01	-0.08	-0.02	0.01	-0.32	-0.12	0.00	0.01	0.00	1.00

La construction de cette matrice de corrélation à partir des variables quantitatives nous permet d'identifier d'éventuelles liaisons linéaires entre les variables candidates à l'intégration du modèle. En effet, l'une des hypothèses fondamentales de la modélisation est l'absence de multicolinéarité entre les variables explicatives. La présence de coefficients de corrélation élevés entre deux variables nous exposerait à un risque important d'instabilité des coefficients et de dégradation de la qualité prédictive du modèle.

Nous rappelons qu'à ce stade, les variables Engagement_prorat et ANC_ENTR ont été écartées à la suite des tests précédents. Par conséquent, nous ne porterons pas notre attention sur les corrélations impliquant ces variables, afin de nous concentrer exclusivement sur les prédicteurs potentiels restants. Néanmoins, plusieurs coefficients de corrélation retiennent notre attention :

- 0,26 entre MVT_AFF_12M et SFMois_AG ;
- 0,47 entre MVT_AFF_12M et SOLD_CRE ;
- 0,36 entre NBJDEPDP et NB_JR_DEB ;
- 0,39 entre NBJDEPDP et NJRS_DEP_DA ;
- 0,36 entre NBJRDB_AT et NB_JR_DEB ;
- -0,32 entre NBJRDB_AT et SOLD_DIB ;
- 0,93 entre NB_JR_DEB et NJRS_DEP_DA ;
- 0,73 entre SFMois_AG et SOLD_CRE.

Parmi ces corrélations, nous pouvons isoler un premier groupe composé des variables MVT_AFF_12M, SFMois_AG et SOLD_CRE. L'intégration simultanée de ces trois variables dans le modèle s'avère complexe, car elles génèrent deux des trois coefficients de corrélation les plus élevés de notre analyse. Un second groupe peut être constitué par l'ensemble des variables de comptage des jours de débit. Nous y observons le coefficient de corrélation le plus significatif du jeu de données entre NB_JR_DEB et NJRS_DEP_DA $(0,93)$. Il ne sera pas envisageable d'inclure ces deux variables conjointement dans le modèle, ce qui nous imposera un arbitrage. Les autres coefficients obtenus par différentes combinaisons atteignent au minimum 0,35, seuil à partir duquel nous nous devons d'être vigilants quant aux risques de redondance de l'information. Le choix définitif des variables que nous déciderons d'écarter sera affiné lors de l'étape de discrétisation de ces dernières.

Afin d'évaluer le pouvoir discriminant de nos variables qualitatives vis-à-vis de notre variable cible, nous calculons la statistique du V de Cramer. Cette mesure permet de quantifier l'intensité de la dépendance entre deux variables nominales. Ainsi, les variables explicatives maximisant le V de Cramer sont celles qui possèdent le pouvoir discriminant le plus élevé pour notre modélisation.

TABLE 3.8 – Liaison entre la variable cible et les variables qualitatives (V de Cramer)

Variable	V de Cramer
Depassement	0.3012
segment	0.2645
Depassement_M1	0.2418
NDB_6	0.1874
Impaye	0.1532
NDB_5	0.1421
Depassement_M2	0.1389
NDB_4	0.1256
Impaye_M1	0.1147
NDB_3	0.1082
Depassement_M3	0.0984
NDB_2	0.0875
Impaye_M2	0.0763
NDB_1	0.0654
Impaye_M3	0.0542
CODETAJUR	0.0412
Top_engagement	0.0387
SEC_DER	0.0298
CODNAF2	0.0254
Top_MLT	0.0187
Top_immo	0.0123
Top_pro_lib	0.0084
Top_Interfimo	0.0051
Top_hbilan_uniq	0.0023

Le tableau ci-dessus nous révèle qu'environ 40 % de nos variables présentent un V de Cramer inférieur à 0,10. Ces variables seront les premières à être écartées lors de la phase de sélection pour la construction de notre modèle. Nous observons que la variable Dépassement présente le V de Cramer le plus élevé (0,30), suivie de la variable Segment (0,26). Par ailleurs, les variables dont le coefficient est compris entre 0,1 et 0,3 concernent majoritairement, à l'exception de Segment, des indicateurs d'arriérés de paiement ou des antécédents de défaut au sens de Bâle. Nous notons que, pour une échéance donnée, la variable Dépassement affiche un pouvoir discriminant supérieur à celui de la variable Impayé. En outre, une corrélation est observée concernant la récence : plus l'arriéré est récent, plus son V de Cramer est élevé. À l'inverse, nous observons une dynamique opposée pour les variables associées aux antécédents de défaut : plus le défaut est ancien, plus son pouvoir discriminant semble marqué comparativement aux dates plus récentes. D'un point de vue historique et métier, les arriérés de paiement constituent l'un des signes précurseurs pouvant conduire au défaut bâlois. Ici, le V de Cramer fait office de confirmation statistique d'une réalité opérationnelle établie. Concernant les antécédents de défaut, ces résultats sont cohérents : un client ayant déjà été en situation de défaut par le passé présente statistiquement une probabilité de réitération plus élevée qu'un client n'ayant jamais connu d'incident majeur. Les variables restantes ont trait à la détention de crédits, aux engagements ou à la catégorie d'appartenance du client. Dans une base de données composée exclusivement de SCI, il est logique que ces variables soient les moins discriminantes. En effet, la grande majorité de ces structures ont recours à l'emprunt de manière récurrente pour financer leurs activités, ce qui tend à lisser les profils au sein de notre échantillon.

Transformation des variables

Après avoir finalisé l'étude exploratoire de la base de données, nous entamons désormais la transformation de nos variables explicatives afin de les rendre conformes aux exigences de la régression logistique. Pour rappel, à la suite de l'analyse des liaisons avec la variable cible, nous ne conservons que les prédicteurs disposant d'un pouvoir discriminant satisfaisant. Les variables retenues se répartissent comme suit :

— Variables quantitatives : MVT_AFF_12M, NBJDEPDP, NBJRDB_AT, NB_JR_DEB, NJRS_DEP_DA, SFMois_AG, SOLD_CRE, SOLD_DIB et ANC_RELA_LCL.
— Variables qualitatives : Depassement, Depassement_M1, Depassement_M2, Depassement_M3, Impaye, Impaye_M1, Impaye_M2, Impaye_M3, segment, NDB_1, NDB_2, NDB_3, NDB_4, NDB_5 et NDB_6.

À ce stade, nous disposons de 24 variables potentielles. Toutefois, afin de garantir la robustesse, la parcimonie et la stabilité du modèle final, nous visons une sélection restreinte comprenant entre 8 et 12 variables.

4.1 Discrétisation des variables quantitatives

La stratégie que nous utiliserons sera la suivante. La première étape consistera à discrétiser la distribution de chaque variable en dix déciles lorsque cela sera possible. Nous récupèrerons ensuite les valeurs des bornes de ces déciles afin de les regrouper en cinq modalités au maximum. Ce regroupement s'effectuera en fonction de la variable cible de défaut à douze mois, tout en respectant deux contraintes majeures. La première imposera un écart relatif du taux de défaut observé d'au moins 30 % entre deux modalités successives. La seconde exigera que la proportion de chaque modalité représente au minimum 1 % de l'échantillon total.

Table 4.1 – Variable d'analyse : NBJDEPDP

Variable d'analyse : NBJDEPDP			
Rang pour la variable NBJDEPDP	**Fréquence**	**Minimum**	**Maximum**
4	67766	0	0
9	2238	1	1755

Nous prendrons comme exemple ici la variable NBJDEPDP. Pour cette variable, nous discrétisons la distribution en deux modalités : une dont le minimum et le maximum sont égaux à zéro et une seconde qui prend pour valeur tout nombre différent de zéro. Nous n'avons pas besoin de procéder au regroupement des modalités car elles sont au nombre de deux. Enfin, les contraintes sont respectées. En effet, l'écart relatif en risque est de 2160 %, ce qui est supérieur à 30 %. La première modalité représente 96,80 % de notre échantillon et la seconde 3,20 %. Les deux sont supérieures à 1 %. Nous obtenons la discrétisation suivante :

TABLE 4.2 – Discrétisation d'une variable quantitative Exemple (NBJDEPDP)

NBJDEPDP en fonction de DDefaut_NDB			
Rang pour la variable NBJDEPDP	**DDefaut_NDB**		
Fréquence Pourcentage Pct de ligne Pct de col.	**0**	**1**	**Total**
4	67414 96.30 99.48 97.15	352 0.50 0.52 57.24	67766 96.80
9	1975 2.82 88.25 2.85	263 0.38 11.75 42.76	2238 3.20
Total	69389 99.12	615 0.88	70004 100.00

TABLE 4.3 – Variable d'analyse : SFMois_AG

Variable d'analyse : SFMois_AG			
Rang pour la variable SFMois_AG	**Fréquence**	**Minimum**	**Maximum**
0	6983	-41395200,00	2100,00
1	7024	2200,00	29000,00
2	6993	29100,00	82600,00
3	6999	82700,00	163300,00
4	7002	163400,00	295800,00
5	7002	295900,00	519900,00
6	6999	520000,00	937500,00
7	7002	937600,00	1833500,00
8	7000	1833800,00	4389600,00
9	7000	4389900,00	857169700

Cette variable est l'illustration parfaite d'un cas de non-linéarité détecté lors de l'étape des statistiques descriptives. Suite à la discrétisation de la variable en dix déciles, préalablement à la discrétisation finale, nous observons premièrement une relation décroissante entre le montant du solde en fin de mois et le risque de tomber en défaut bâlois. Plus le montant augmente, moins l'individu est risqué. Lorsque le montant franchit la barre des 4 389 900 centimes à la hausse, ce qui correspond à la valeur minimale prise par le dernier décile, le risque de tomber en défaut bâlois réaugmente. Lors de la discrétisation finale en cinq modalités ou moins, ce décile sera regroupé avec les déciles ayant un risque de tomber en défaut bâlois équivalent. L'ensemble des variables présentant une non-linéarité sera traité de cette façon. Nous obtenons la discrétisation suivante :

TABLE 4.4 – Exemple pour SFMois_AG_C

Table de SFMois_AG_C par DDefaut_NDB			
SFMois_AG_C	DDefaut_NDB		
Fréquence Pourcentage Pct de ligne Pct de col.	**0**	**1**	**Total**
0	13986 19.98 99.89 20.16	16 0.02 0.11 2.60	14002 20.00
1	27922 39.89 99.71 40.24	81 0.12 0.29 13.17	28003 40.00
2	6952 9.93 99.33 10.02	47 0.07 0.67 7.64	6999 10.00
3	13867 19.81 98.93 19.98	150 0.21 1.07 24.39	14017 20.02
4	6662 9.52 95.40 9.60	321 0.46 4.60 52.20	6983 9.98
Total	69389 99.12	615 0.88	70004 100.00

TABLE 4.5 – Exemple pour NBJRDB_AT_C

NBJRDB_AT_C en fonction de DDefaut_NDB			
NBJRDB_AT_C	DDefaut_NDB		
Fréquence Pourcentage Pct de ligne Pct de col.	**0**	**1**	**Total**
0	69052 98.64 99.13 99.51	607 0.87 0.87 98.70	69659 99.51
1	337 0.48 97.68 0.49	8 0.01 2.32 1.30	345 0.49
Total	69389 99.12	615 0.88	70004 100.00

Ici, le tableau ci-dessus illustre le cas d'une variable non conservée à cause d'un non respect de la proportion minimal de 1 % et d'un problème antérieur. Ici, la classe 1 ne représente que 0,49 % de notre échantillon total, ce qui va à l'encontre de notre contrainte de 1 %. Si nous ajoutons à cela que cette variable entretient une relation linéaire avec d'autres variables explicatives, ce qui est susceptible d'engendrer de la multicolinéarité, il est préférable de mettre cette variable de côté pour la suite de l'étude. Suite à la discrétisation de nos variables, nous allons mobiliser la statistique du V de Cramer afin d'arbitrer entre les variables que nous devons conserver parmi celles pour lesquelles nous avons identifié des coefficients de corrélation élevés (cf matrice de Pearson). Nous procédons au calcul du V de Cramer pour les variables de flux/stock ainsi que celles de comptage.

TABLE 4.6 – V de Cramer des variables de flux/stock

Variables	V_Cramer
DDefaut_NDB * SFMois_AG_C	0,14
DDefaut_NDB * MVT_AFF_12M_C	0,06
DDefaut_NDB * SOLD_CRE_C	0,08

La variable ayant le V de Cramer le moins important est SOLD_CRE_C avec 0,08. Pour rappel, celle-ci présen-

tait un coefficient de corrélation de 0,73 avec SFMois_AG_C ainsi qu'un autre de 0,47 avec MVT_AFF_12M_C. Nous écarterons donc SOLD_CRE_C afin de conserver SFMois_AG_C et MVT_AFF_12M_C. Ces dernières présentent une statistique de V de Cramer plus importantes ainsi qu'un coefficient de corrélation entre elles de 0,26, ce qui reste relativement faible comparé aux autres coefficients de corrélation des variables de ce premier groupe.

TABLE 4.7 – V de Cramer des variables de comptage (Jour de Débit)

Variables	V_Cramer
DDefaut_NDB * NB_JR_DEB_C	0,15
DDefaut_NDB * NBJDEPDP_C	0,21
DDefaut_NDB * NJRS_DEP_DA_C	0,15

Pour le deuxième groupe, nous constatons que la variable ayant le V de Cramer le plus élevé est NBJDEPDP_C avec 0,21. De plus, les deux autres variables ont le même V de Cramer à deux chiffres après la virgule, soit 0,15. Il s'agissait déjà des variables présentant le coefficient de corrélation le plus élevé. Nous considérons que, compte tenu du V de Cramer le plus important et des coefficients de corrélation élevés entre NBJDEPDP et les autres variables, nous ne conserverons que NBJDEPDP_C pour la suite de l'étude.

4.1.1 Création de variables

Une grande majeur partie des variables qualitatives encore en lisse ici souffrent du même problème : au moins une des modalités ne respecte pas la contrainte de proportion minimale de 1 % par rapport à l'échantillon de référence. De plus, nous constatons que les groupes ne respectant pas cette contrainte sont généralement ceux pour lesquels le risque de tomber en défaut bâlois est le plus élevé. Afin d'exploiter au mieux ces informations, la stratégie que nous adopterons consistera à croiser ces variables afin de rendre acceptables les proportions de ces modalités au sein de nouvelles variables. Pour ce faire, nous prévoyons de jouer sur la nature des variables ainsi que sur leur temporalité. Pour obtenir la variable arriérés de paiement (à date d'observation) qui est un bon prédicteur historique de la tombée en défaut bâlois, nous avons décidé de croiser la variable dépassement avec celle des impayés. Nous présentons cette variable obtenue que nous avons nommés arrieres_adate_C.

TABLE 4.8 – Exemple d'un croisement de variable (Arriérés à date) entre Dépassement et Impayé.

Arrieres_adate_C en fonction de DDefaut_NDB			
Arrieres_adate_C	DDefaut_NDB		
Fréquence Pourcentage Pct de ligne Pct de col.	**0**	**1**	**Total**
0	67263 96.08 99.55 96.94	307 0.44 0.45 49.92	67570 96.52
1	1557 2.22 95.70 2.24	70 0.10 4.30 11.38	1627 2.32
2	569 0.81 70.51 0.82	238 0.34 29.49 38.70	807 1.15
Total	69389 99.12	615 0.88	70004 100.00

De la même façon, nous avons créé la variable incidents passés qui regroupe les arriérés passés et les défauts bâlois passés. Les variables encore candidates à l'intégration de notre modèle final sont les suivantes :

- NBJDEPDP_C
- SFMois_AG_C
- MVT_AFF_12M_C
- ANC_RELA_LCL_C
- SOLD_DIB_C
- segment_c

— Arrieres_adate_C
— incident_passe_C.

4.2 Etude de stabilité

Afin de valider la pertinence et la robustesse de la construction de nos variables, plusieurs contrôles s'imposent. La première étape consiste à mener une étude de stabilité, tant en termes de risque que de volume, pour en garantir la fiabilité temporelle.

L'étude de la stabilité en volume de la variable NBJDEPDP_C révèle une forte stabilité sur la période d'observation concernée. En effet, la modalité « 0 » (absence de jours débiteurs au-delà de l'autorisation) présente une concentration constante, oscillant entre 96 % et 97 %. La modalité « 1 » (nombre de jours supérieur à 0) représente de manière pérenne environ 3,25 % de l'échantillon total sur l'ensemble de la période.

L'étude de stabilité en risque [1], quant à elle, est plus fluctuente notamment pour la modalité 1. Nous considérerons tout de même que celle-ci est stable sur la période concernée. Cela confirme donc la bonne dicrétisation de notre variable, celle ci étant apte à intégrer notre modèle.

Nous pouvons en dire de même pour toutes les autres variables qui, à ce stade, étaient candidates, à l'exception de ANC_RELA_LCL_C.

Nous précisons cependant que pour les variables ayant plus de deux modalités, les courbes associées peuvent parfois se confondre, que ce soit pour l'étude de stabilité en volume ou en risque. Ceci se justifie notamment par un problème d'échelle causé par la présence d'une modalité extrêmement conséquente en termes de risque ou de volume par rapport aux autres.

Concernant ANC_RELA_LCL_C, nous observons que même si les modalités sont stables en volume sur la période étudiée, deux d'entre elles (ancienneté comprise entre 20 et 25 ans et ancienneté supérieure à 25 ans) coïncident en termes de risque au deuxième et au quatrième trimestre avec une valeur de 0,46 %. Cela nous indique que nous devons regrouper ces deux modalités afin que cette variable soit stable en risque.

Après retraitement de la variable, nous obtenons l'étude de stabilité en risque ci-dessus. Le retraitement a consisté à regrouper le groupe dont les clients étaient compris entre 19 et 25 avec celui dont les clients étaient supérieur à 25. Suite au retraitement de ANC_RELA_LCL_C, celle ci est à présent stable en risque et en volume.

4.2.1 Etude de colinéarité

La deuxième étape nécessaire à la validation de la pertinence et de la robustesse de nos variables est l'étude de la colinéarité entre elles. Celle-ci est d'autant plus importante pour le respect de l'hypothèse d'absence de multicolinéarité de la régression logistique.

Table 4.9 – Matrice de statistiques de V de Cramer (Variables explicatives)

Variable	NBJDEPDP_C	SFMois_AG_C	MVT_AFF_12M	ANC_RELA_LCL_C	SOLD_DIB_C	segment_c	Arrieres_adate_C	incident_passe_C
ANC_RELA_LCL_C	0,049	0,073	0,038	1	0,107	0,181	0,038	0,083
Arrieres_adate_C	0,949	0,252	0,070	0,038	0,193	0,418	0,038	1
MVT_AFF_12M_C	0,073	0,284	1	0,038	0,067	0,170	0,070	0,170
NBJDEPDP_C	1	0,331	0,073	0,049	0,252	0,489	0,949	0,392
SFMois_AG_C	0,331	1	0,284	0,073	0,190	0,212	0,252	0,278
SOLD_DIB_C	0,252	0,190	0,067	0,107	1	0,238	0,193	0,285
incident_passe_C	0,392	0,278	0,170	0,083	0,285	0,383	0,386	1
segment_c	0,489	0,212	0,170	0,181	0,238	1	0,418	0,383

L'étude des statistiques du V de Cramer nous permet d'identifier trois relations entre variables pour lesquelles celui-ci est strictement supérieur à 0,4, ce qui implique une forte relation linéaire.

Les relations concernées sont :
— segment_c et NBJDEPDP_C avec un V de Cramer de 0,49
— Arrieres_adate_C et NBJDEPDP_C avec un V de Cramer de 0,94
— Arrieres_adate_C et segment_c avec un V de Cramer de 0,49

1. Les graphiques des études de stabilité de certaines de nos variables sont disponibles en annexe.

Afin de garantir l'hypothèse d'absence de multicolinéarité, nous ne conserverons qu'une seule d'entre elles en nous basant sur la force de la relation de ces variables avec la variable cible DDefaut_NDB.

Nous rappelons qu'à ce niveau de l'étude, nous avons construit nos variables avec la volonté qu'elles fassent sens économiquement et qu'elles disposent d'un réel pouvoir discriminant vis-à-vis de notre variable cible DDefaut. Le fait est que les variables identifiées ici fourniraient, si elles étaient toutes intégrées dans le modèle, la même information.

Nous notons par ailleurs que ces variables sont celles ayant le plus de relations avec d'autres variables impliquant un V de Cramer strictement supérieur à 0,3 :

- NBJDEPDP_C
- segment_c
- Arrieres_adate_C.

Éliminer deux de ces variables nous permettra d'être d'autant plus confortables par rapport à la multicolinéarité. Nous justifierons de la conservation d'une de ces variables à l'étape suivante.

Estimation du modèle

5.1 Choix de la technique de modélisation

TABLE 5.1 – V de Cramer pour déterminer l'ordre d'intégration pour la méthode pas à pas

Variable	V_Cramer
DDefaut_NDB * Arrieres_adate_C	0,34
DDefaut_NDB * incident_passe_C	0,26
DDefaut_NDB * segment_c	0,25
DDefaut_NDB * NBJDEPDP_C	0,21
DDefaut_NDB * SFMois_AG_C	0,14
DDefaut_NDB * SOLD_DIB_C	0,12
DDefaut_NDB * MVT_AFF_12M_C	0,06
DDefaut_NDB * ANC_RELA_LCL_C	0,02

Nous utilisons la méthodologie du step-by-step.

Pour ce faire, nous avons utilisé encore et toujours la statistique du V de Cramer entre nos variables explicatives restantes et notre variable cible.

La variable la plus corrélée à cette dernière se retrouve être "Arrieres_adate_C". Nous l'incorporons donc dans notre première modèle et nous décidons donc d'exclure les 2 autres variables qui lui étaient fortement corrélées (segment_c et NBJDEPDP_C).

5.2 Estimation du modèle

TABLE 5.2 – Premier modèle

Paramètre	Mod.	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept		1	-5,3895	0,0572	8876,8835	<,0001
Arrieres_adate_C	1	1	2,2879	0,1349	287,7090	<,0001
Arrieres_adate_C	2	1	4,5179	0,0961	2211,1070	<,0001
Arrieres_adate_C	0	0	0	.	.	.

Modèle 1 : D de Sommer : 0,478 ; AUC : 0,739 ; AIC : 5 486,835

L'aire sous la courbe ROC (AUC) est de 0,739 pour notre premier modèle. Il s'agit de la probabilité que le modèle classe correctement une paire d'observations (un bon client et un mauvais client). Un AUC supérieur à 0,5 signifie que le modèle fait mieux que le hasard.

L'AIC de notre premier modèle est de 5 486,835. Cet indicateur témoigne de l'équilibre du couple précision/simplicité. Il pénalise l'ajout d'une variable supplémentaire ne permettant pas d'améliorer significativement

la précision du modèle. Nous chercherons à le minimiser et à maximiser l'aire sous la courbe de ROC par l'ajout d'autres variables.

L'indice de Gini est une mesure statistique de dispersion utilisée pour évaluer le pouvoir discriminant du modèle. Dans le cadre du risque de crédit, il quantifie la capacité du modèle à différencier les clients en défaut des clients sains Il correspond à deux fois l'aire comprise entre la courbe ROC et la diagonale (qui représente un modèle aléatoire). Il est relié à l'AUC par la relation suivante :

$$\text{Gini} = 2 \times \text{AUC} - 1$$

Un indice de Gini de 0 correspond à un modèle purement aléatoire (aucune discrimination), tandis qu'un indice proche de 1 (ou 100 %) indique une discrimination parfaite. Dans notre cas, avec un AUC de 0,739, nous obtenons un indice de Gini de 0,478 après notre premier modèle. Son équivalent sur SAS et le D de Sommer. L'ajout de nos autres variables aura pour but de maximiser cette métrique.

Nous ajoutons ensuite la deuxième variable la plus corrélée à notre cible ("incident_passe_C). Nous avons répété cette méthodologie jusqu'à aboutir au modèle suivant :

TABLE 5.3 – Dernier modèle

Paramètre	Mod.	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept		1	-7,6284	0,2375	1031,7204	<,0001
Arrieres_adate_C	1	1	1,1068	0,1284	74,2975	<,0001
Arrieres_adate_C	2	1	1,8681	0,1347	192,4050	<,0001
Arrieres_adate_C	0	0	0	.	.	.
incident_passe_C	1	1	0,7485	0,1054	50,4400	<,0001
incident_passe_C	0	0	0	.	.	.
SFMois_AG_C	1	1	0,2388	0,2061	1,3430	0,2465
SFMois_AG_C	2	1	0,9117	0,2200	17,1748	<,0001
SFMois_AG_C	3	1	1,0271	0,1988	26,7055	<,0001
SFMois_AG_C	4	1	1,0261	0,2089	24,1244	<,0001
SFMois_AG_C	0	0	0	.	.	.
SOLD_DIB_C	1	1	0,8033	0,1121	51,3301	<,0001
SOLD_DIB_C	2	1	1,1870	0,1105	115,3333	<,0001
SOLD_DIB_C	0	0	0	.	.	.
MVT_AFF_12M_C	1	1	0,0165	0,0962	0,0292	0,8642
MVT_AFF_12M_C	2	1	0,3783	0,1035	13,3538	0,0003
MVT_AFF_12M_C	0	0	0	.	.	.
segment_c	1	1	1,1474	0,1696	45,7784	<,0001
segment_c	2	1	2,3558	0,1897	154,1577	<,0001
segment_c	0	0	0	.	.	.

Modèle 6 : D de Sommer : 0,761 ; AUC : 0.881 ; AIC : 5015.149

Nous avons décidé d'incorporer la variable segment_C, pour laquelle nous étions réticents au départ au vu de son V de Cramer de 0,49 avec Arrieres_adate_C. Cela nous permet d'améliorer le D de Sommer et l'AUC, tout en minimisant l'AIC. De plus, cela ne perturbe pas la stabilité de notre modèle, car les coefficients de l'ensemble de nos variables (y compris segment_c) sont significatifs.

5.3 Présentation des variables retenues

Après avoir construit un modèle avec l'objectif qu'il soit le plus robuste statistiquement, et ce à partir des données à notre disposition, nous allons présenter plus en détail les variables utilisées.

La première variable est Arrieres_adate_C. Elle fait référence à l'existence ou non d'un problème d'arriéré (impayé ou dépassement) à la date d'observation. Elle a été construite en croisant les variables Impayé et Dépassement. Elle se compose de 3 modalités. La modalité 2 correspond à au moins un arriéré (impayé ou dépassement) supérieur à 30 jours. La modalité 1 correspond à au moins un arriéré compris entre 1 et 30 jours. La modalité 0 correspond à aucun arriéré.

La seconde variable est Incident_passe_C. Elle fait référence à l'existence antérieure ou non d'un incident de type arriéré ou de tombée en défaut bâlois. Ses modalités sont les suivantes : la modalité 1 signale l'existence d'au moins un problème d'arriéré antérieur supérieur à 30 jours ou d'une tombée en défaut bâlois antérieure. La modalité 0 regroupe à la fois les clients sans aucun problème, les clients pour lesquels l'information en termes d'incident est incomplète, et enfin les clients ayant eu au moins un problème d'arriéré compris entre 1 et 30 jours.

La troisième variable renvoie au solde de fin de mois du compte courant professionnel. La modalité 3 correspond à un solde inférieur ou égal à 21 € (2 100 centimes). La modalité 2 correspond à un solde compris entre 22 € et 826 €. La modalité 1 correspond à un solde compris entre 827 € et 1 633 €. La modalité 0 regroupe les soldes compris entre 1 634 € et 9 375 €, entre 9 376 € et 43 896 €, et enfin ceux supérieurs à 43 896 €. L'effet de non-linéarité identifié lors de la première discrétisation a été absorbé par le retraitement de la variable dans la suite de l'étude. C'est globalement ce qu'il s'est passé pour toutes les variables où de la non-linéarité était présente à un moment donné.

La quatrième variable renvoie au solde créditeur moyen sur les 12 derniers mois. La modalité 2 correspond à un solde créditeur inférieur à -737,80 €. La modalité 1 correspond à un solde créditeur supérieur ou égal à -737,80 €. La modalité 0 correspond à un solde créditeur supérieur ou égal à -55,73 €.

La cinquième variable correspond au mouvement d'affaires moyen sur les 12 derniers mois. La modalité 1 correspond à un mouvement d'affaires moyen inférieur ou égal à 426,91 €. La modalité 0 correspond à un mouvement d'affaires moyen supérieur à 426,91 €.

Enfin, la dernière variable (segment_c) regroupe en son sein les individus selon trois catégories. La modalité 2 désigne les individus à haut risque (ayant eu plus de 30 jours d'impayés ou ayant été au moins une fois en défaut sur les 12 derniers mois). La modalité 1 désigne les individus détenant d'autres produits bancaires. La modalité 0 désigne les individus détenant un prêt immobilier ou MLT, garanti ou non, ainsi que ceux détenant ou non un engagement hors bilan.

5.4 Analyse des résultats

5.4.1 Odds ratios

Nous interpréterons seulement les odds ratios du modèle final. Un individu ayant au moins un arriéré supérieur à 30 jours à la date d'observation a 6,595 fois plus de risque de tomber en défaut bâlois qu'un individu n'ayant aucun problème d'arriérés. De même, un individu ayant au moins un arriéré compris entre 1 et 30 jours à la date d'observation a 2,245 fois plus de risque de tomber en défaut bâlois qu'un individu n'ayant aucun problème d'arriéré, toutes choses égales par ailleurs (TCEPA).

5.4.2 Validation sur l'échantillon test

TABLE 5.4 – Analyse des valeurs estimées du maximum de vraisemblance (Échantillon Test)

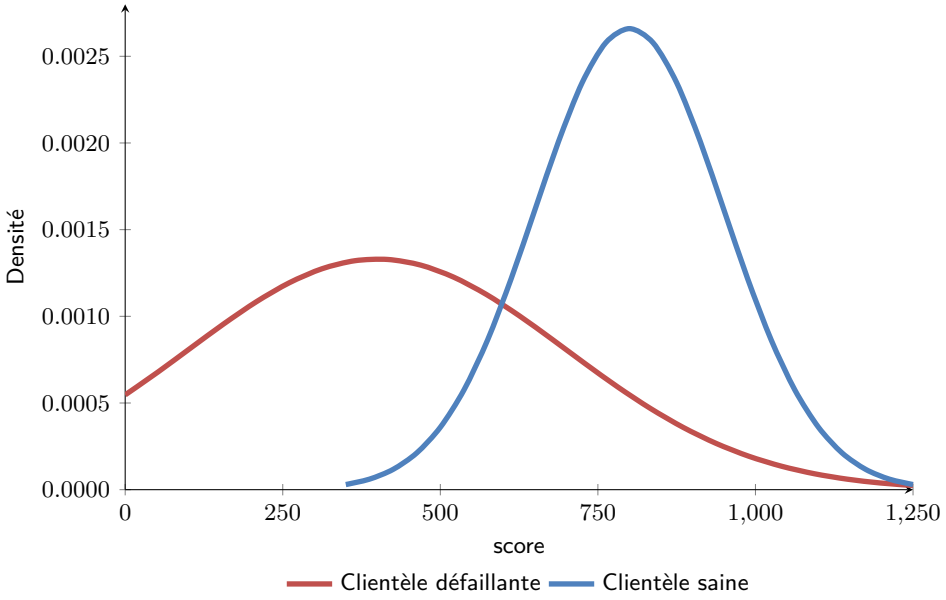
Paramètre	Mod.	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept		1	-7,3540	0,2967	614,3049	<,0001
Arrieres_adate_C	1	1	1,0896	0,2480	19,2972	<,0001
Arrieres_adate_C	2	1	1,9951	0,2678	55,5088	<,0001
Arrieres_adate_C	0	0	0	.	.	.
incident_passe_C	1	1	0,9076	0,2032	19,9555	<,0001
incident_passe_C	0	0	0	.	.	.
SFMois_AG_C	1	1	0,8017	0,2965	7,3106	0,0069
SFMois_AG_C	2	1	1,2604	0,2203	32,7233	<,0001
SFMois_AG_C	3	1	1,0592	0,2529	17,5460	<,0001
SFMois_AG_C	0	0	0	.	.	.
SOLD_DIB_C	1	1	0,7399	0,2034	13,2345	0,0003
SOLD_DIB_C	2	1	1,1895	0,1954	37,0769	<,0001
SOLD_DIB_C	0	0	0	.	.	.
MVT_AFF_12M_C	1	1	0,2221	0,1628	1,8598	0,1727
MVT_AFF_12M_C	0	0	0	.	.	.
segment_c	1	1	1,0230	0,2741	13,9332	0,0002
segment_c	2	1	1,7729	0,3296	28,9397	<,0001
segment_c	0	0	0	.	.	.

Modèle échantillon test : D de Sommer : 0,756 ; AUC : 0.878 ; AIC : 2 070,382

Nous obtenons cette fois-ci des métriques en légère baisse par rapport à l'échantillon d'apprentissage. On note de plus que la modalité 1 de la variable MVT_AFF_12M_C a une p-valeur différente (0,17). L'échantillon hors temps nous permettra de vérifier s'il ne s'agit que d'un effet dû à la taille de l'échantillon test ou non.

Analyse des performances

6.1 Densités conditionnelles



Les courbes de densités conditionnelles sont obtenues à partir des scores associés aux clients sains et des scores associés aux clients tombés en défaut bâlois. Notre modèle semblerait être dans l'incapacité de prédire correctement la tomber en défaut balois à 12 mois pour un individu ayant un score supérieur à 500. En effet, nous observons visuellement que les courbes se confondent pour un score supérieur à 500. À partir de ce seuil, nous comprenons qu'il devient difficile pour le modèle d'identifier distinctement les individus susceptibles de tomber en défaut bâlois. Nous pensons que cette faiblesse s'explique par le taux de défaut particulièrement faible du portefeuille de clientèle SCI (0,88 %).

Nous aurions pu traiter ce phénomène via des méthodes de rééquilibrage de la base de données (telles que l'oversampling ou l'undersampling) afin d'accroître la capacité prédictive du modèle sur la classe minoritaire. Cependant, compte tenu de notre objectif de création d'une grille de score opérationnelle, l'usage de ces méthodes aurait gonflé artificiellement le taux de défaut du portefeuille et aurait pu surévaluer l'importance de certaines variables. À terme, nous estimons que cela aurait risqué de biaiser les scores attribués aux nouveaux clients en les rendant indûment sévères. C'est pourquoi nous avons décidé de poursuivre avec le modèle présenté sur les données initiales.

6.2 Validation out of sample du modèle

TABLE 6.1 – Analyse des valeurs estimées du maximum de vraisemblance (Modèle Final)

Paramètre	Mod.	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept		1	-7,4497	0,1778	1755,9399	<,0001
Arrieres_adate_C	1	1	1,1070	0,1285	74,2577	<,0001
Arrieres_adate_C	2	1	1,8667	0,1347	191,9884	<,0001
Arrieres_adate_C	0	0	0	.	.	.
incident_passe_C	1	1	0,7487	0,1054	50,4910	<,0001
incident_passe_C	0	0	0	.	.	.
SFMois_AG_C	1	1	0,7385	0,1557	22,4933	<,0001
SFMois_AG_C	2	1	0,8532	0,1224	48,5902	<,0001
SFMois_AG_C	3	1	0,8494	0,1358	39,1492	<,0001
SFMois_AG_C	0	0	0	.	.	.
SOLD_DIB_C	1	1	0,8095	0,1121	52,1804	<,0001
SOLD_DIB_C	2	1	1,1930	0,1101	117,4347	<,0001
SOLD_DIB_C	0	0	0	.	.	.
MVT_AFF_12M_C	1	1	0,3727	0,0871	18,3115	<,0001
MVT_AFF_12M_C	0	0	0	.	.	.
segment_c	1	1	1,1454	0,1695	45,6889	<,0001
segment_c	2	1	2,3571	0,1896	154,5234	<,0001
segment_c	0	0	0	.	.	.

Modèle échantillon OOT : D de Sommer : 0,805 ; AUC : 0.903 ; AIC : 6753.959

Après retraitement des variables problématiques, l'ensemble des p-valeurs associées aux coefficient Les variables présentent dans le modèle, dans la base hors temps, sont toutes significatives au seuil de 1%. Cela vient nous conforter dans la construction et l'intérêt de nos variables.

Le fait que le D de Sommer soit plus élevé que sur l'échantillon de test suggère une meilleure adéquation du modèle aux données de l'année 2024.

Elaboration de la grille de score

Table 7.1 – Grille de Score

Variable	Modalités	Rép. %	Taux Déf.	Poids	Cont. Éch.	Cont. Sc.
ARRIERES_ ADATE_C	Aucun Arriéré	96,52	0,45	263,24	26,32%	12,66%
	Au moins un Arriéré (Impayé ou Dépassement) entre 1 et 30 jours	2,32	4,30	150,35		
	Au moins un Arriéré (Impayé ou Dépassement) > 30 jours	1,15	29,49	0		
INCIDENT_ PASSE_C	Regroupe les clients sans incident, info incomplète, ou incident 1-30j.	97,97	0,54	93,87	9,39%	5,15%
	Existence d'au moins un incident antérieur > 30 jours ou défaut balois.	2,03	16,99	0		
MVT_AFF_ 12M_C	Mouvement d'affaire moyen > 426,91€	80,00	0,64	26,97	2,70%	4,19%
	Mouvement d'affaire moyen ≤ 426,91€	20,00	1,85	0		
SEGMENT_C	Prêt immobilier ou MLT, garantie ou non, engagement hors bilan	26,90	0,16	314,25	31,43%	32,63%
	Individus détenant d'autres produits bancaires	69,04	0,49	143,69		
	Haut Risque (> 30j impayés ou défaut 12 derniers mois)	4,06	12,17	0		
SFMOIS_ AG_C	Solde fin de mois ≥ 1634 €	60,00	0,23	145,98	14,60%	26,27%
	Solde fin de mois 827 € - 1633 €	10,00	0,67	35,92		
	Solde fin de mois 22 € - 826 €	9,98	4,60	1,44		
	Solde fin de mois ≤ 21 €	20,02	1,07	0		
SOLD_ DIB_C	Solde débiteur ≥ -55,73 €	80,00	0,35	155,68	15,57%	19,10%
	Solde débiteur ≥ -737,80 €	10,00	1,94	84,35		
	Solde débiteur < -737,80 €	10,00	4,04	0		

La grille de score ci-dessus nous permet d'attribuer un score de risque à un client en sommant les points récoltés selon ses caractéristiques personnelles, compte tenu des variables de comparaison entre les individus. La grille présente, par ailleurs, des informations sur ces variables. L'une de ces données est la contribution d'échelle, soit le poids théorique maximum de la variable considérée sur le score final. Une autre information renvoie à la contribution au score de la variable, soit son influence statistique réelle au vu de la distribution des scores des individus observés.

Nous constatons que, pour la variable SEGMENT_C, il existe une correspondance (en points de pourcentage) entre la contribution d'échelle et la contribution au score. La contribution théorique maximale de cette variable au score total est semblable à sa contribution réelle. Nous observons également deux cas intéressants. Pour la variable ARRIERES_ADATE_C, la contribution théorique maximale est environ deux fois plus importante que la contribution réelle. À l'inverse, pour la variable SFMOIS_AG_C, la contribution théorique maximale est environ deux fois moins importante que la contribution réelle. Le solde de fin de mois et le segment auquel appartient un individu expliquent plus de 50 % de la variation du score final entre les individus.

Construction de l'échelle de rating

Pour la construction des classes homogènes de risque, nous avons appliqué la même méthodologie de discrétisation que celle utilisée précédemment pour les variables quantitatives. L'objectif principal était de maximiser l'homogénéité intra-classe tout en assurant une forte hétérogénéité inter-classe. Cette segmentation a été réalisée en respectant deux contraintes majeures : un effectif minimal de 1 % de contreparties par classe et une différenciation des taux de défaut d'au moins 30 % entre chaque classe. Cette démarche nous a permis d'obtenir la grille de risque présentée ci-après :

TABLE 8.1 – Classes homogènes de risque

Classes	Bornes	Nombre de clients	Taux de défaut	Ecart relatif
1	888 - 1000	11950	0.03%	
2	829 - 887	27141	0.16%	433.33%
3	658 - 828	19753	0.54%	237.5%
4	< 658	11160	4.13%	664.81%
Ensemble		**70 004**	**0.88%**	

Le tableau ci-dessus présente la segmentation finale du portefeuille en quatre classes de risque distinctes. L'analyse des résultats confirme la robustesse du modèle sur plusieurs dimensions statistiques. Nous validons tout d'abord la stricte monotonie des taux de défaut qui augmentent progressivement à mesure que la qualité de crédit se dégrade. Le risque passe ainsi d'un niveau marginal de 0,03 % pour la classe 1, regroupant les meilleures contreparties (scores entre 888 et 1000), à 4,13 % pour la classe 4, qui concentre les dossiers les plus fragiles (scores inférieurs à 658).

Nous constatons également que la contrainte de différenciation du risque est largement respectée. Les écarts relatifs entre les classes adjacentes sont très nettement supérieurs au seuil minimal de 30 % fixé ex-ante. Le passage de la classe 1 à la classe 2 matérialise une augmentation du risque de plus de 433 %, tandis que la rupture la plus significative s'opère vers la classe 4, avec un saut de risque de près de 665 % par rapport à la classe 3. Cette forte discrimination permet d'isoler efficacement la population la plus risquée.

Enfin, la répartition des effectifs assure une bonne représentativité statistique de chaque segment. Aucune classe ne représente moins de 1 % du volume total des 70 004 clients. La concentration la plus forte s'opère sur la classe 2, qui regroupe 27 141 clients, reflétant la structure globalement saine de ce portefeuille dont le taux de défaut moyen s'établit à 0,88 %.

Conclusion / Préconisations

Le projet avait pour objectif de construire une grille de score à partir d'un portefeuille dit « low default » (0,88 %) sur une clientèle professionnelle (SCI).

Pour cela, nous avons mobilisé un modèle de régression logistique qui, parmi l'ensemble des modèles de classification, constitue la référence dans ce genre d'exercice en raison de sa capacité à fournir des coefficients interprétables.

Nous soulignons que, malgré un D de Sommer conséquent sur notre base d'apprentissage (0,761), l'étude des courbes de densité conditionnelles montre que notre modèle peine à identifier un client en défaut bâlois face à un client sain lorsque les scores sont supérieurs à 500. Afin de l'améliorer dans cette tâche, nous pensons qu'une des solutions pourrait être d'utiliser une régression logistique pondérée. Cela permettrait de pénaliser plus fortement les erreurs de prédiction du modèle sur la classe sous-représentée, afin que celui-ci ne se concentre pas exclusivement sur la bonne prédiction de la classe majoritaire.

Les coefficients utilisés pour le calcul des poids attribués aux modalités des variables dans la grille de score n'auraient été que faiblement affectés. Nous aurions cependant potentiellement amélioré l'homogénéité intra-classe et l'hétérogénéité inter-classe, nous permettant ainsi d'obtenir une meilleure segmentation des classes homogènes de risque.

Annexes

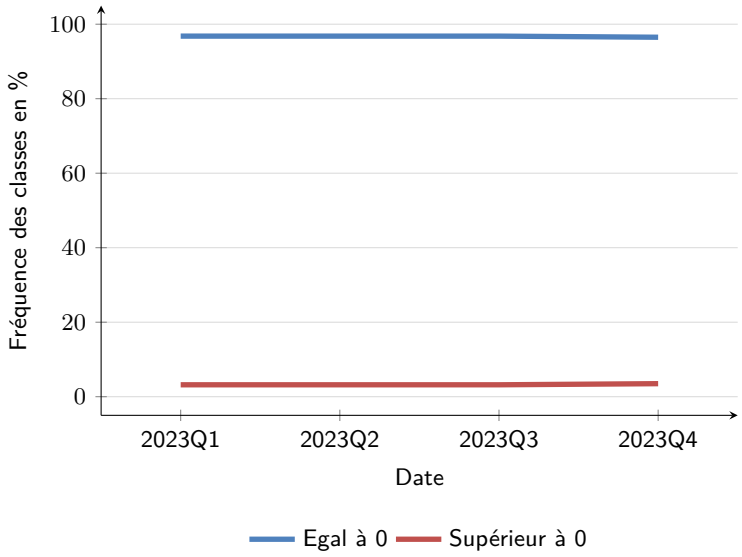
Table 9.1 – Statistiques descriptives complètes des variables quantitatives (hors NJRS_DEP_DA)

Variable	Moyenne	Médiane	Écart-type	Coef. var.	Skewness	Min	Max
ANC_ENTR	162,21	137,00	134,04	82,63	1,56	-1,00	1 487,00
ANC_RELA_LCL	11,45	9,00	9,86	86,05	1,19	0,00	74,00
Engagement_prorat	166 220,60	607,25	582 292,95	350,31	32,84	0,00	43 264 110,91
MVT_AFF_12M	587 288,95	188 333,00	3 563 432,77	606,76	57,45	0,00	368 081 847,00
NBJDEPDP	1,48	0,00	20,60	1 390,37	40,12	0,00	1 755,00
NBJRDB_AT	0,23	0,00	3,97	1 731,88	19,51	0,00	92,00
NB_JR_DEB	2,88	0,00	11,33	393,67	5,26	0,00	92,00
SFMois_AG	2 079 818,83	295 900,00	9 699 522,17	466,36	35,76	-41 395 200,00	857 169 700,00
SOLD_CRE	68 314 669,59	11 459 218,00	328 931 116,11	481,49	38,60	0,00	30 217 927 537
SOLD_DIB	-923 652,32	0,00	49 367 166,17	-5 344,78	-91,37	-6 017 151 801	0,00

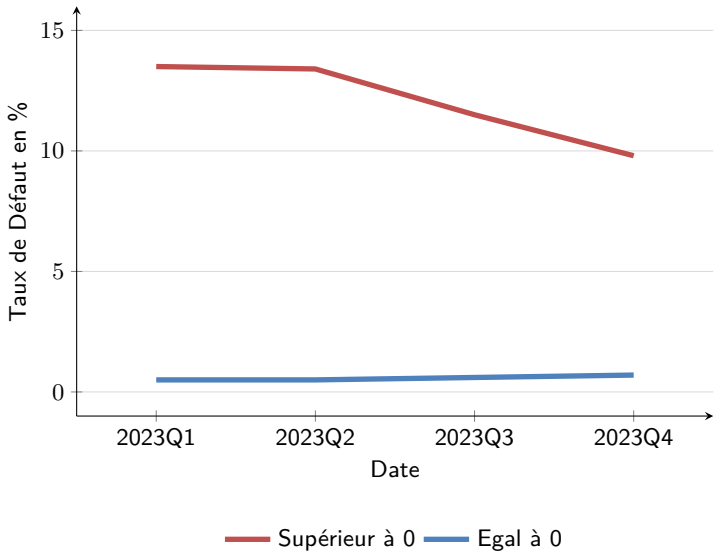
Table 9.2 – Table de d_SFMois_AG par DDefaut_NDB

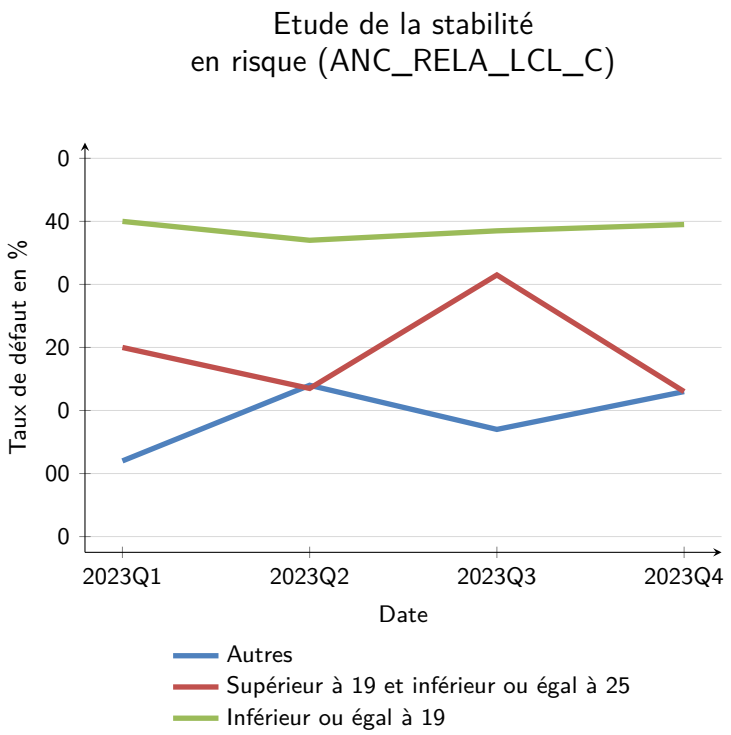
Table de d_SFMois_AG par DDefaut_NDB			
d_SFMois_AG (Rang pour la variable SFMois_AG)	**DDefaut_NDB**		
Fréquence Pourcentage Pct de ligne Pct de col.	**0**	**1**	**Total**
0	6662 9.52 95.40 9.60	321 0.46 4.60 52.20	6983 9.98
1	6941 9.92 98.82 10.00	83 0.12 1.18 13.50	7024 10.03
2	6926 9.89 99.04 9.98	67 0.10 0.96 10.89	6993 9.99
3	6952 9.93 99.33 10.02	47 0.07 0.67 7.64	6999 10.00
4	6980 9.97 99.69 10.06	22 0.03 0.31 3.58	7002 10.00
5	6984 9.98 99.74 10.06	18 0.03 0.26 2.93	7002 10.00
6	6981 9.97 99.74 10.06	18 0.03 0.26 2.93	6999 10.00
7	6995 9.99 99.90 10.08	7 0.01 0.10 1.14	7002 10.00
8	6991 9.99 99.87 10.08	9 0.01 0.13 1.46	7000 10.00
9	6977 9.97 99.67 10.05	23 0.03 0.33 3.74	7000 10.00
Total	69389 99.12	615 0.88	70004 100.00

Etude de la stabilité
en volume (NBJDEPDP_C)

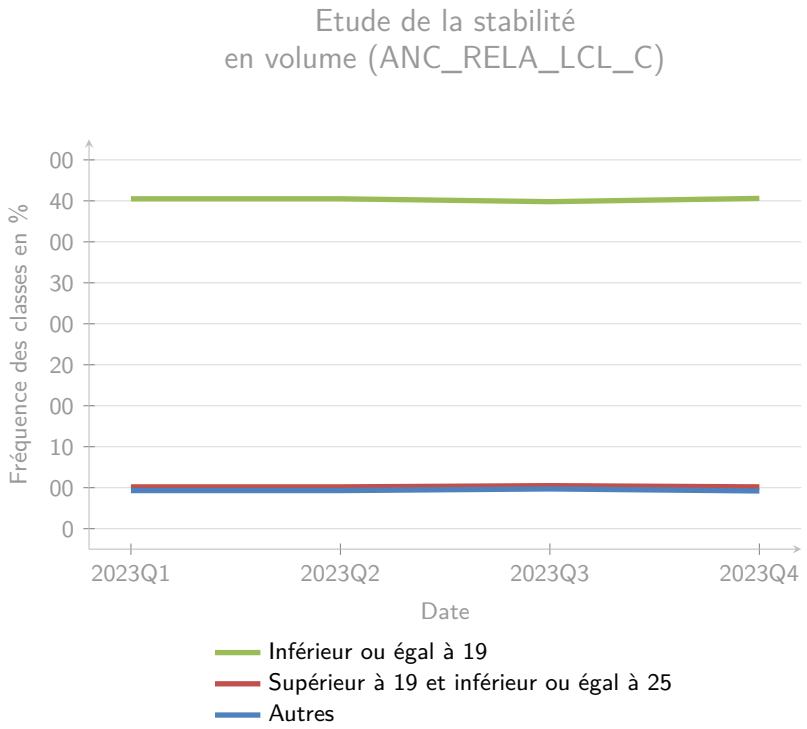


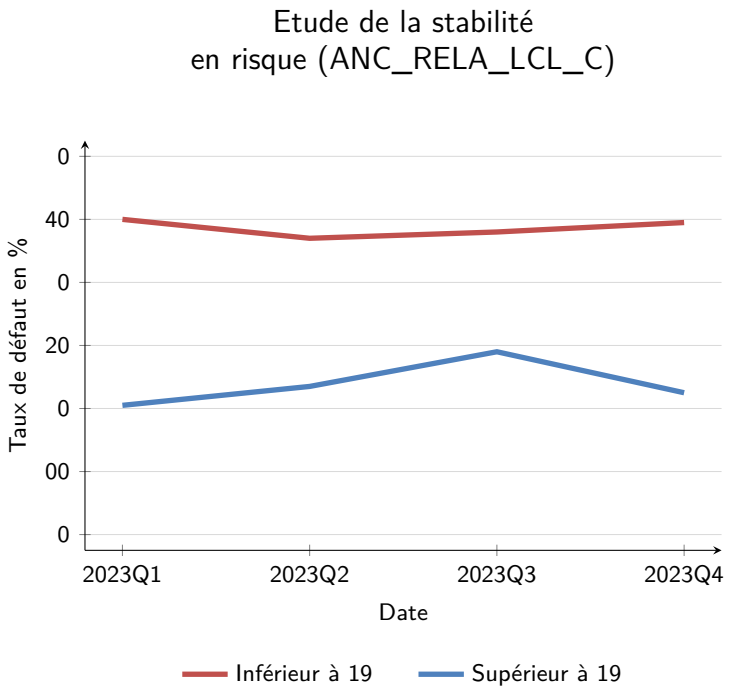
Etude de la stabilité
en risque (NBJDEPDP_C)





Problème concernant la stabilité en risque pour la variable ANC_RELA_LCL_C




Etude de la stabilité
en risque (ANC_RELA_LCL_C)
Taux de défaut en %
0
40
0
20
0
00
0
2023Q1
2023Q2
2023Q3
2023Q4
Date
Inférieur à 19
Supérieur à 19