Université Paris 1 Panthéon-Sorbonne

École d'Économie de La Sorbonne

Master 2 Modélisations Statistiques Économiques et Financières (MoSEF)

Année universitaire 2023-2024

Modélisation de la probabilité de défaut bâlois

Étudiants :
Cécile Huang
Jynaldo Jeannot
Yoan Jsem
Alice Liu

Superviseurs :
Armand L'Huillier
Aryan Razaghi

Table des matières

1 Introduction

Dans le cadre du Challenge Nexialog, nous modélisons la probabilité de défaut dans le respect de la réglementation Bâloise. Pour cela, nous avons à disposition de nombreuses bases de données comprenant les caractéristiques du client (état civil, situation familiale, région...), des données sur les paiements, des données externes... Sur une période allant de 2013 à 2020. Le but de ce projet est de faire le traitement nécessaire pour sélectionner les meilleures variables possibles afin de modéliser la probabilité de défaut.

Pour mieux structurer nos variables, nous les avons catégorisées en variables numériques, catégorielles et binaires. Cette approche nous permet de mieux comprendre la distribution de nos données et de préparer le terrain pour une analyse plus approfondie.

Dans cette optique, notre travail se divise en plusieurs phases. Tout d'abord, nous nous penchons sur l'analyse des variables binaires, en effectuant des tests de stabilité et des tests de χ^2. Ensuite, nous nous concentrons sur les variables catégorielles, avec une distinction entre celles qui ont moins de 4 modalités et celles qui en ont davantage. L'analyse des variables numériques occupe également une place centrale, impliquant des étapes telles que le feature engineering, les tests de Mann-Whitney, la discrétisation, et le calcul de l'information value. Après avoir soigneusement préparé et structuré nos variables, nous nous lançons dans la construction du modèle logistique. Une fois le modèle établi, nous réalisons une grille de score pour nous permettre d'attribuer une note a chaque demande de prêt. Enfin, nous abordons la phase cruciale de la quantification du risque, en effectuant une segmentation en classes de risques et en calculant la Marge de Conservatisme (MoC) selon les deux approches, MoC C et MoC A. L'analyse se poursuit avec la Long-Run Average (LRA) et la détermination de la Probabilité de Défaut (PD).

2 Analyses des variables binaires

Pour les variables binaires (2 modalités), notre objectif est de garantir la stabilité de ces variables en termes de risque et de volume, tout en assurant leur significativité dans le contexte de la régression logistique.

2.1 Test de stabilité

Pour évaluer la stabilité en risque, nous avons calculé le taux de défaut pour chaque modalité de ces variables au fil des années, permettant ainsi d'observer leur évolution. La stabilité en risque est atteinte lorsque les courbes des différentes modalités ne se croisent

pas, indiquant que l'impact de la variable sur la variable cible reste constant.

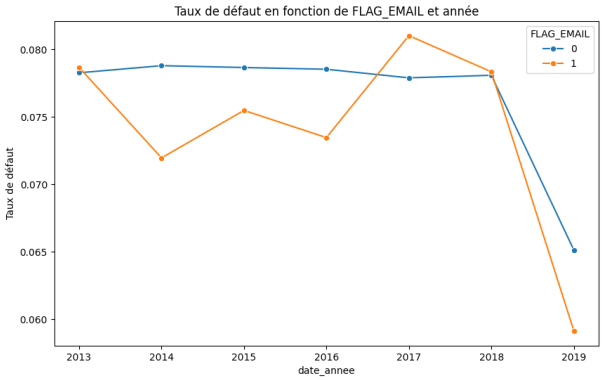

FIGURE 1 – Stabilité en risque d'une variable binaire

Un exemple illustratif montre que la variable `FLAG_EMAIL` n'est pas stable en risque, ses courbes se croisent. Ce test de stabilité en risque nous permet de mettre à l'écart certaines variables qui ne sont pas stables en risque comme `FLAG_MOBIL, FLAG_CONT_MOBILE, FLAG_EMAIL, REG_REGION_NOT_LIVE_REGION, REG_REGION_NOT_WORK_REGION et LIVE_REGION_NOT_WORK_REGION`. En parallèle, nous réalisons un test de stabilité en volume, vérifiant si les modalités conservent une proportion relativement constante dans le temps. Une variable est stable en volume si toutes ses modalités ont un pourcentage au-dessus de 5%. Nous conservons donc les variables avec une stabilité en volume des modalités toutes supérieures à 5%.

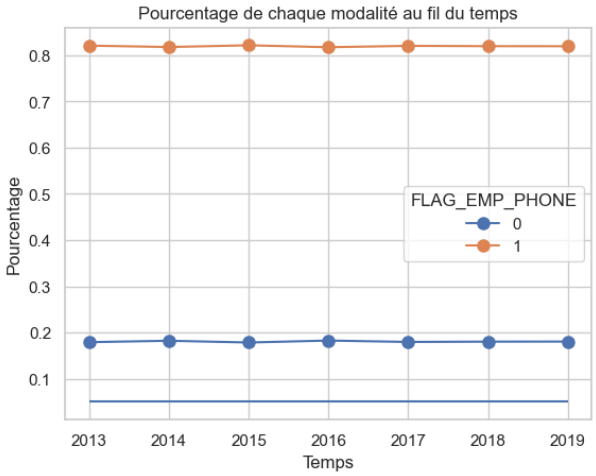

FIGURE 2 – Stabilité en volume d'une variable binaire

La stabilité en volume et en risque des variables est cruciale pour assurer la robustesse, l'interprétabilité et la fiabilité des modèles de régression logistique.

2.2 Test de χ^2

Par la suite, les variables binaires restantes ont été soumises à un test du χ^2 afin d'évaluer l'indépendance entre chaque variable binaire et la variable cible (`TARGET`). Les résultats ont démontré que toutes les variables binaires que nous avons conservées sont significatives. Cela signifie qu'elles ont un impact statistiquement significatif sur la variable cible, renforçant ainsi leur pertinence potentielle pour expliquer les variations de la variable cible.

3 Analyses des variables catégorielles (nombre de modalités inférieur à 4)

Nous distinguons les variables à faible nombre de modalités des autres variables catégorielles. Ces variables à faible nombre de modalités se caractérisent par une gamme restreinte de catégories, inférieur à 4 modalités. En procédant ainsi, nous simplifions notre approche analytique des variables catégorielles.

3.1 Test de stabilité

Comme nous avons procédé pour les variables binaires, nous effectuons le test de stabilité en volume puis un test de stabilité en risque.

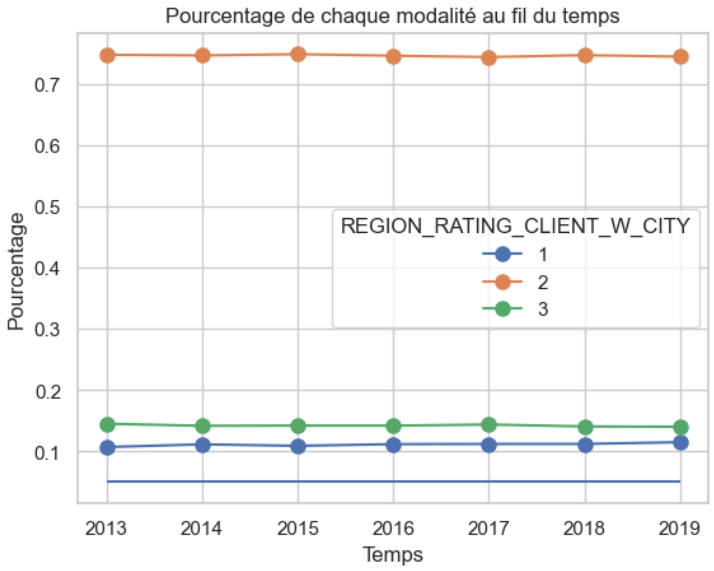

FIGURE 3 – Stabilité en volume d'une variable à faible modalité

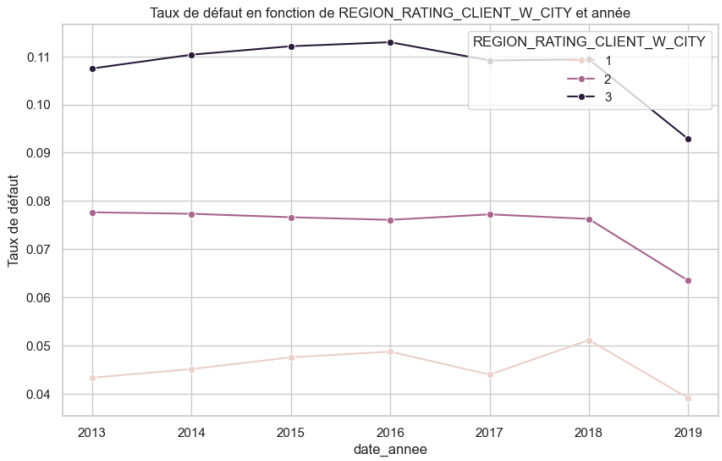

FIGURE 4 – Stabilité en risque d'une variable à faible modalité

Certaines variables sont identifiées comme non stables en risque, telles que `FLAG_OWN_REALTY`, `FONDKAPREMONT_MODE`, `HOUSETYPE_MODE` et `EMERGENCYSTATE_MODE` sont exclues de nos variables potentielles pour la régression logistique.

4 Analyses des variables catégorielles

Dans cette partie, nous allons nous intéresser aux variables catégorielles dans sa généralité, afin d'avoir une vision plus générale. Parmi celles-ci nous pouvons retrouver par exemple les variables suivantes : `'NAME_TYPE_SUITE'`, `'NAME_INCOME_TYPE'`, `'NAME_EDUCATION_TYPE'`, `'NAME_FAMILY_STATUS'`, `'NAME_HOUSING_TYPE'`, `'OCCUPATION_TYPE'` etc...

4.1 Tests statistiques

Des premiers tests statistiques comme celui du χ^2 (test d'indépendance entre la variable explicative et la target), et du celui du V de Cramer (qui mesure le degré de dépendance de la relation) sont réalisés afin de nous fournir une compréhension des variables catégorielles quant à la sélection de variables.

Parmi nos variables catégorielles pré-définies, toutes ont une p-value inférieure à 0.05, donc que les variables sont statistiquement significatives, et qu'il y a une relation entre la variable considérée et la target.

Cela n'est pas une condition suffisante quant à la sélection de variables, donc on réalise des tests plus approfondis dans les paragraphes suivants.

4.2 Test de stabilité

Comme évoqué dans les précédentes sections, des tests de stabilité en volume et des tests de stabilité en risque dans le temps sont effectués. Prenons ici un exemple pour illustrer nos propos.

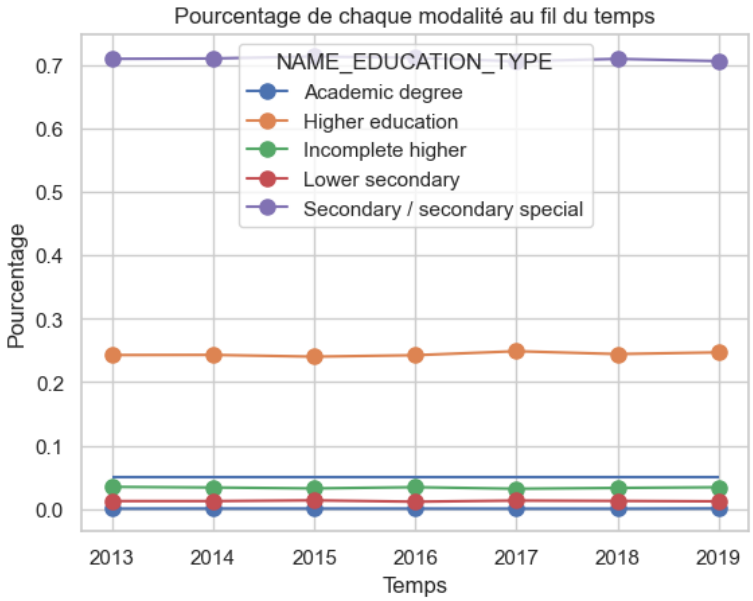

FIGURE 5 – Test de stabilité en volume dans le temps pour la variable `NAME_EDUCATION_TYPE`

Avec ce graphe du test de stabilité en volume de `NAME_EDUCATION_TYPE`, nous pouvons observer que des courbes de pourcentage de modalités se situent en deçà du seuil critique de 5%. Pour rendre pertinente cette variable, nous pouvons recourir à des regroupements de modalités.

On vérifie également que les variables catégorielles sont stables en risque dans le temps. On voit par exemple pour la variable `NAME_EDUCATION_TYPE` que les courbes se croisent donc que la variable n'est pas stable en risque dans le temps. On peut penser qu'on ne doit pas la garder mais nous verrons dans la prochaine section ce que nous en ferons.

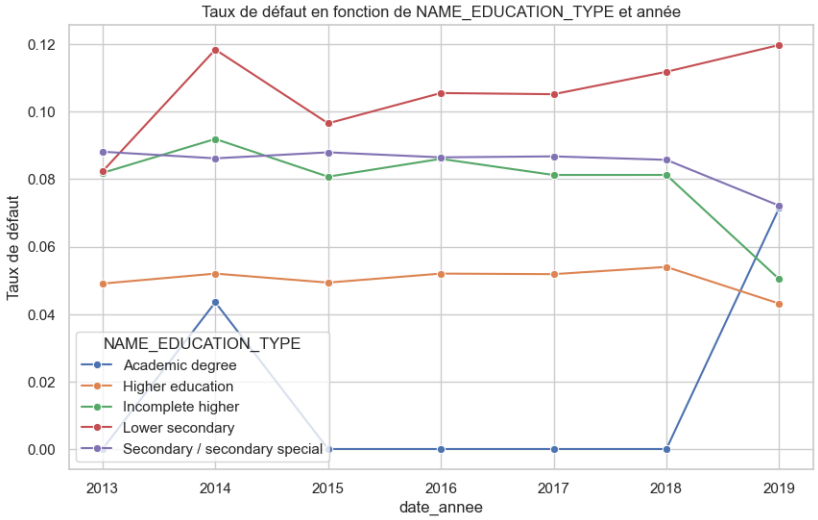

FIGURE 6 – Test de stabilité en risque dans le temps pour la variable `NAME_EDUCATION_TYPE`

4.3 Regroupement des modalités

Dans le but d'apporter de l'information et de l'interprétabilité au modèle, nous faisons donc nos propres regroupements de modalités, selon nos connaissances générales de la variable considérée. Cela nous permet d'accroître les chances que la variable respecte les critères de stabilité en risque et en volume dans le temps.

Par exemple, considérons la variable `NAME_EDUCATION_TYPE`. Les modalités initiales étaient au nombre de 5, que nous avons regroupé en deux modalités :

— Graduated : "Academic degree", "Higher education"

— Non graduated : "Lower secondary", "Secondary / secondary special", "Incomplete higher"

Voici les tests de stabilité en risque et en volume après avoir regroupé les modalités :

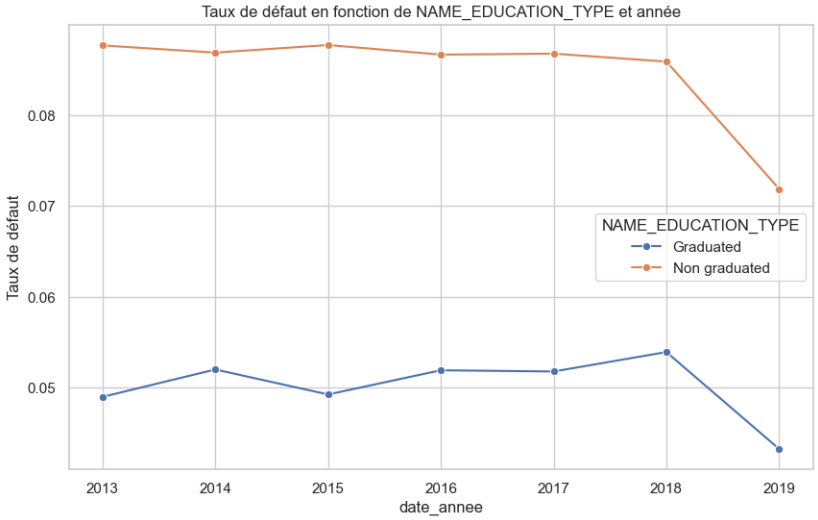

FIGURE 7 – Test de stabilité en risque dans le temps pour la variable `NAME_EDUCATION_TYPE` après regroupement de modalités

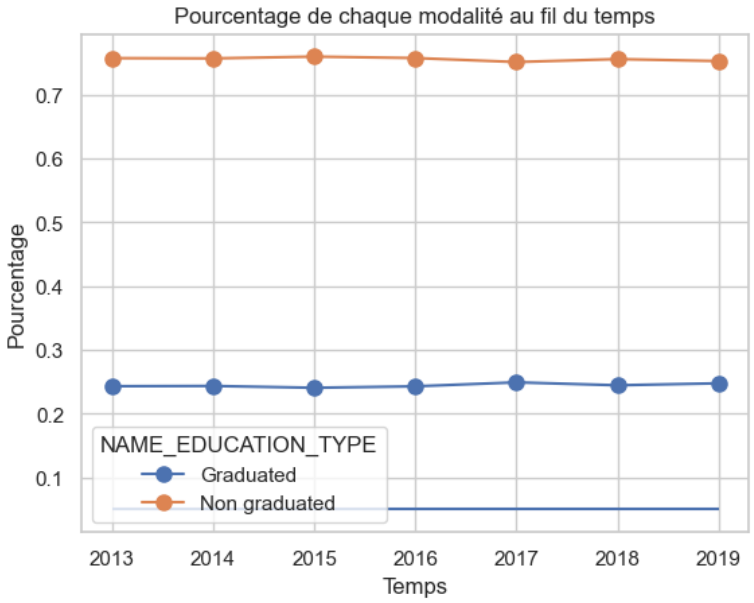

FIGURE 8 – Test de stabilité en volume dans le temps pour la variable `NAME_EDUCATION_TYPE` après regroupement de modalités

Ce qui montre que le regroupement de modalités n'était pas sans intérêt, le pourcentage de chaque modalité dans le temps étant supérieur à 5%.

D'autres variables ont également eu recours au regroupement des modalités comme `NAME_FAMILY_STATUS`, ou aussi des regroupements de variables comme `FAM_STATS_CHILD` qui est composé d'informations sur le statut familial et le fait d'avoir des enfants ou non.

Les variables ne respectant pas les tests de stabilité en volume et en risque dans le

temps, même après regroupement de modalités, sont exclues de notre choix de variables potentielles pour notre modèle.

4.4 Information value

Nous évaluons la valeur informative de chacune des variables en utilisant une table de contingence qui récapitule les fréquences d'occurrence des différentes modalités par rapport à la variable cible. Pour chaque modalité de la variable, nous calculons les taux d'événement (`TARGET`= 1) et de non-événement (`TARGET`= 0).

L'*Information Value* (IV) de la variable est mesurée à l'aide de la formule suivante :

$$IV = \sum_i (event\_rate_i - non\_event\_rate_i) \times \log\left(\frac{event\_rate_i}{non\_event\_rate_i}\right)$$

De ce fait, après avoir créé de nouvelles variables par un regroupement de modalités, nous calculons l'information value pour celles-ci afin d'évaluer le pouvoir prédictif de la variable considérée par rapport à notre cible. Plus la valeur est élevée, plus la variable explicative a un pouvoir prédictif. Dans notre cas, nous avons les IV suivantes pour :

- `NAME_FAMILY_STATUS_2` : 0.009
- `FAM_STATS_CHILD` : 0.016
- `HAS_CHILDREN` : 0.005
- `NAME_EDUCATION_TYPE` : 0.05

Au vue des valeurs et du nombre de variables que l'on peut retenir dans le modèle final, on peut garder `NAME_EDUCATION_TYPE` qui a le pouvoir prédictif le plus élevé.

5 Analyses des variables numériques

Après une analyse approfondie des variables catégorielles visant à identifier leur impact sur la variable cible tout en assurant leur stabilité en volume et en risque, nous dirigeons notre attention vers l'examen des variables numériques. L'objectif principal est de comprendre le rôle et l'influence de ces variables dans l'explication de notre variable cible.

5.1 Feature engineering

Dans cette section, nous mettons en lumière nos travaux quant à la création de nouvelles variables dans le but de mieux exploiter les informations de nos données. Nous introduisons tout d'abord la variable :

$$AMT\_CREDIT\_TO\_INCOME = \frac{AMT\_CREDIT + CB\_AMT\_CREDIT\_SUM}{AMT\_INCOME\_TOTAL}$$

Cette nouvelle variable exprime la part de charge financière liée au crédit dans le revenu de l'individu. Nous définissons également la variable :

$$AMT\_CREDIT\_NORM = \frac{AMT\_CREDIT}{AMT\_GOODS\_PRICE}$$

Cette variable offre une mesure du rapport entre le montant du crédit et le prix des biens. Nous avons également créé les variables :

- `BORROWER_AGE`
- `BORROWER_SENIORITY`
- `BORROWER_FIDELITY`

Représentant respectivement l'âge du demandeur, l'ancienneté de l'emploi, et le nombre d'année depuis son enregistrement dans la banque en tant que client.

5.2 Test de Mann-Whitney

L'analyse des nouvelles variables numériques créées a été soumise à des tests de Mann-Whitney, et tous ces tests ont révélé une signification statistique. La p-value significativement faible obtenue lors de ces tests indique qu'il existe une différence statistiquement significative entre les distributions de la variable pour les deux groupes. Ce constat suggère que ces variables sont potentiellement informatives pour discerner les deux niveaux de la variable cible. Nous pouvons donc procéder à la discrétisation de ces variables.

5.3 Discrétisation

Nous avons mis en œuvre une approche de discrétisation des variables continues à l'aide d'un modèle basé sur un arbre de décision. Notre objectif était de créer des intervalles significatifs pour chaque variable continue, permettant ainsi une meilleure compréhension de leur impact sur la variable cible. Les feuilles de l'arbre nous permettent d'identifier les seuils qui sont utilisés pour diviser la variable continue en intervalles. Afin de conserver la monotonie du taux de défaut sur les intervalles, nous avons utiliser un arbre à très peu de

profondeur(2). Ensuite nous vérifions la stabilité en volume et en risque de ces variables discrétisées.

5.4 Information value

Suite à la discrétisation des variables, nous évaluons leur valeur informative.

Cette mesure, l'*Information Value* (IV), sert de critère de sélection des variables les plus informatives pour notre modèle de régression logistique. Les 7 variables sélectionnées pour notre modèle sont à la fois informatives, stables en volume et en risque. Cette double stabilité garantit que la relation entre ces variables et la variable cible reste constante au fil du temps et que les proportions des modalités demeurent relativement constantes.

Un autre aspect crucial est l'ordonnancement monotone des modalités par rapport aux valeurs originales de la variable continue. Cette disposition monotone assure une correspondance cohérente entre les catégories discrètes et les niveaux croissants ou décroissants de la variable continue d'origine. Ainsi, les variables retenues sont informatives et stables, mais conservent également une relation monotone avec la variable continue sous-jacente, renforçant ainsi leur pertinence pour notre modèle.

6 Construction du modèle

Après avoir réalisé une discrétisation des variables ainsi qu'une sélection des variables poussées nous passons à la modélisation. On choisit de modéliser sans constante (non explicable) et décidons des modalités de références.

Pour les modalités de références nous décidons de choisir la modalité qui représente le pire taux de défaut. Ce détail nous servira à l'interprétation de la grille de score par la suite, il est donc bon de le retenir.

La régression logistique est composée des variables suivantes : `OCCUPATION_TYPE`, `NAME_EDUCATION_TYPE`, `AMT_CREDIT_NORM`, `BORROWER_AGE`, `BORROWER_SENIORITY`, `CB_NB_CREDIT_CLOSED` et `CB_DAYS_CREDIT`.

Les variables sont toutes discrétisées avec `NAME_EDUCATION_TYPE` une variable binaire. Toutes les modalités représentées et donc les variables en général sont significatives à 1%. On s'assure aussi que tous les coefficients de la régression sont du même signe, ici tous sont négatifs. On passe donc à l'étape suivante.

Variable	Modalités	Coefficient	std err	z	P>|z|	[0.025, 0.975]
OCCUPATION_TYPE	Treatment(reference=0)[0]	-0.9269	0.059	-15.723	0.000	-1.042, -0.811
OCCUPATION_TYPE	Treatment(reference=0)[1]	-0.7332	0.046	-15.868	0.000	-0.824, -0.643
OCCUPATION_TYPE	Treatment(reference=0)[2]	-0.5365	0.045	-11.963	0.000	-0.624, -0.449
OCCUPATION_TYPE	Treatment(reference=0)[3]	-0.3485	0.051	-6.806	0.000	-0.449, -0.248
NAME_EDUCATION_TYPE	Treatment(reference='Non graduated')[T.Graduated]	-0.4997	0.025	-20.222	0.000	-0.548, -0.451
AMT_CREDIT_NORM	Treatment(reference=3)[Interval(-inf, 1.158, closed='right')]	-0.6339	0.026	-24.132	0.000	-0.685, -0.582
AMT_CREDIT_NORM	Treatment(reference=3)[Interval(1.158, 1.211, closed='right')]	-0.3376	0.032	-10.608	0.000	-0.400, -0.275
AMT_CREDIT_NORM	Treatment(reference=3)[Interval(1.211, 1.317, closed='right')]	-0.1996	0.033	-6.083	0.000	-0.264, -0.135
BORROWER_AGE	Treatment(reference=0)[Interval(30.5, 38.5, closed='right')]	-0.0779	0.026	-3.020	0.003	-0.129, -0.027
BORROWER_AGE	Treatment(reference=0)[Interval(38.5, 52.5, closed='right')]	-0.2929	0.025	-11.748	0.000	-0.342, -0.244
BORROWER_AGE	Treatment(reference=0)[Interval(52.5, inf, closed='right')]	-0.4591	0.033	-14.102	0.000	-0.523, -0.395
BORROWER_SENIORITY	Treatment(reference=0)[Interval(2.5, 4.5, closed='right')]	-0.1048	0.025	-4.163	0.000	-0.154, -0.055
BORROWER_SENIORITY	Treatment(reference=0)[Interval(4.5, 10.5, closed='right')]	-0.3358	0.024	-13.854	0.000	-0.383, -0.288
BORROWER_SENIORITY	Treatment(reference=0)[Interval(10.5, inf, closed='right')]	-0.4539	0.029	-15.664	0.000	-0.511, -0.397
CB_NB_CREDIT_CLOSED	Treatment(reference=0)[Interval(0.5, 1.5, closed='right')]	-0.2592	0.026	-9.857	0.000	-0.311, -0.208
CB_NB_CREDIT_CLOSED	Treatment(reference=0)[Interval(1.5, inf, closed='right')]	-0.4585	0.022	-20.916	0.000	-0.502, -0.416
CB_DAYS_CREDIT	Treatment(reference=3)[Interval(-inf, -2921.5, closed='right')]	-0.7503	0.036	-20.567	0.000	-0.822, -0.679
CB_DAYS_CREDIT	Treatment(reference=3)[Interval(-2921.5, -254.5, closed='right')]	-0.7457	0.036	-21.003	0.000	-0.815, -0.676
CB_DAYS_CREDIT	Treatment(reference=3)[Interval(-254.5, -47.5, closed='right')]	-0.3522	0.035	-10.117	0.000	-0.420, -0.284

TABLE 1 – Régression Logistique

Concernant les résultats de la régression logistique, nous obtenons sur le train un Gini égal à 0.330 et sur le test un Gini égal à 0.322.

7 Réalisation d'une grille de score

Maintenant que nous avons notre régression logistique avec des coefficients tous de même signe et chaque modalité significative à 5 %, on s'attache à réaliser une grille de score. On suit la méthodologie fournie afin de construire une grille de score de 0 à 1000. Puisque nous avons décidé de prendre précédemment comme modalité de référence la pire modalité, plus notre note se rapproche de 0 plus le taux de défaut attendu sera élevé. Inversement plus la note est élevée, plus le taux de défaut des individus est faible.

Variable	Modalités	Coefficient	Effectif	p-value	Note	Contribution	Taux de défaut
AMT_CREDIT_NORM	(-inf, 1.158]	-0.6339	64.4	0.000	165	19.15	6.31
AMT_CREDIT_NORM	(1.158, 1.211]	-0.3376	15.2	0.000	88	19.15	8.42
AMT_CREDIT_NORM	(1.211, 1.317]	-0.1996	11.3	0.000	52	19.15	10.14
AMT_CREDIT_NORM (Référence)	(1.317, inf]	0.0000	9.2	0.000	0	19.15	12.53
BORROWER_AGE (Référence)	(-inf, 30.5]	0.0000	17.2	0.000	0	11.58	10.94
BORROWER_AGE	(30.5, 38.5]	-0.0779	21.2	0.003	20	11.58	9.06
BORROWER_AGE	(38.5, 52.5]	-0.2929	34.4	0.000	76	11.58	7.18
BORROWER_AGE	(52.5, inf]	-0.4591	27.3	0.000	120	11.58	5.03
BORROWER_SENIORITY	(-inf, 2.5]	0.0000	29.1	0.000	0	12.66	10.52
BORROWER_SENIORITY	(10.5, inf]	-0.4539	32.3	0.000	118	12.66	4.90
BORROWER_SENIORITY	(2.5, 4.5]	-0.1048	15.2	0.000	27	12.66	9.18
BORROWER_SENIORITY	(4.5, 10.5]	-0.3358	23.4	0.000	88	12.66	6.83
CB_DAYS_CREDIT	(-254.5, -47.5]	-0.3522	29.8	0.000	92	17.18	8.92
CB_DAYS_CREDIT	(-2921.5, -254.5]	-0.7457	35.6	0.000	195	17.18	6.04
CB_DAYS_CREDIT (Référence)	(-47.5, inf]	0.0000	5.2	0.000	0	17.18	12.42
CB_DAYS_CREDIT	(-inf, -2921.5]	-0.7503	29.4	0.000	196	17.18	7.43
CB_NB_CREDIT_CLOSED (Référence)	(-inf, 0.5]	0.0000	25.0	0.000	0	13.55	10.11
CB_NB_CREDIT_CLOSED	(0.5, 1.5]	-0.2592	17.4	0.000	68	13.55	8.04
CB_NB_CREDIT_CLOSED	(1.5, inf]	-0.4585	57.5	0.000	120	13.55	6.44
NAME_EDUCATION_TYPE	Graduated	-0.4997	24.4	0.000	130	16.50	5.00
NAME_EDUCATION_TYPE (Référence)	Non graduated	0.0000	75.6	0.000	0	16.50	8.49
OCCUPATION_TYPE	0	-0.9269	7.1	0.000	151	9.38	5.34
OCCUPATION_TYPE	1	-0.7332	51.5	0.000	100	9.38	6.06
OCCUPATION_TYPE	2	-0.5365	34.2	0.000	49	9.38	9.68
OCCUPATION_TYPE (Référence)	3	-0.3485	7.2	0.000	0	9.38	11.49

TABLE 2 – Grille de score

Après avoir construit la grille de score, il faut attribuer à chaque individu une note en fonction de ses caractéristiques. On vérifie que la somme des plus grosses notes fasse 1000, que la somme des contributions fasse 100% et que la somme des fréquences parmi chaque modalité des variables fasse 100%. Une fois cela fait, on peut être certains que la grille ne comporte pas de problème majeur.

FIGURE 9 – Distribution des notes sur l'ensemble d'entraînement

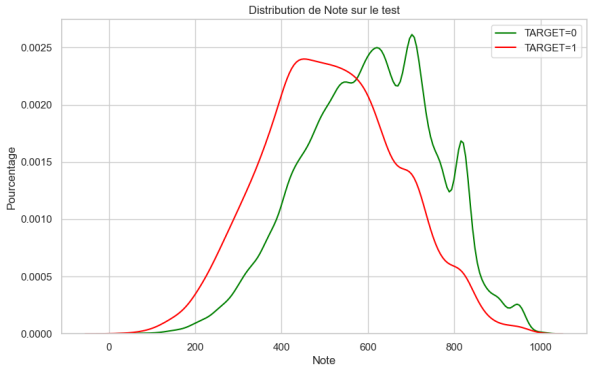

FIGURE 10 – Distribution des notes sur l'ensemble de test

Après avoir observé le graphique des densités conditionnelles, on voit que nos deux courbes ne sont pas superposées ce qui est une bonne chose. Cependant le score n'est pas très discriminant car elles ne sont pas très éloignées les unes des autres. En revanche les courbes font sens. La distribution des défauts (1) est centrée sur des notes faibles tandis que les clients sains ont une distribution de notes plus élevée ce qui est attendu.

8 Quantification du risque

8.1 Segmentation en classe de risques

Après avoir réalisé la grille de score, nous avons pu attribuer une note à chaque individu selon ses caractéristiques données par la discrétisation. On peut donc créer les Classes Homogènes de Risques (CHR) qui sont au nombre de 7 et qui nous permettent d'obtenir le taux d'observations par segment ainsi que le taux de défaut par segment.

Nous devions respecter les hypothèses suivantes :

1. Homogénéité au sein de chaque classe
2. Hétérogénéité entre les classes
3. Éviter 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

Nous avons choisi d'utiliser la méthode *Jenks Natural Breaks Optimization* pour créer des classes homogènes :

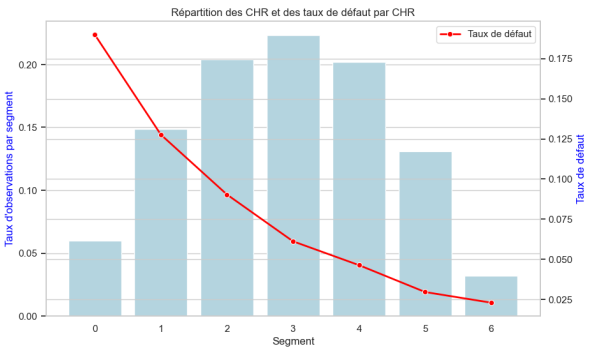

FIGURE 11 – Répartition des CHR et des taux de défauts CHR sur les données train

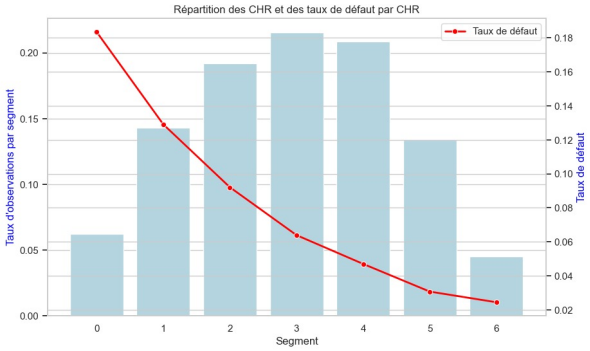

FIGURE 12 – Répartition des CHR et des taux de défauts CHR sur les données test

Le graphique ci-après montre l'évolution des taux de défauts des CHR dans le temps. Nous pouvons remarquer que les courbes sont bien distinctes et éloignées. Toutefois, nous avons constaté un resserrement sur les classes les plus saines montrant des similitudes.

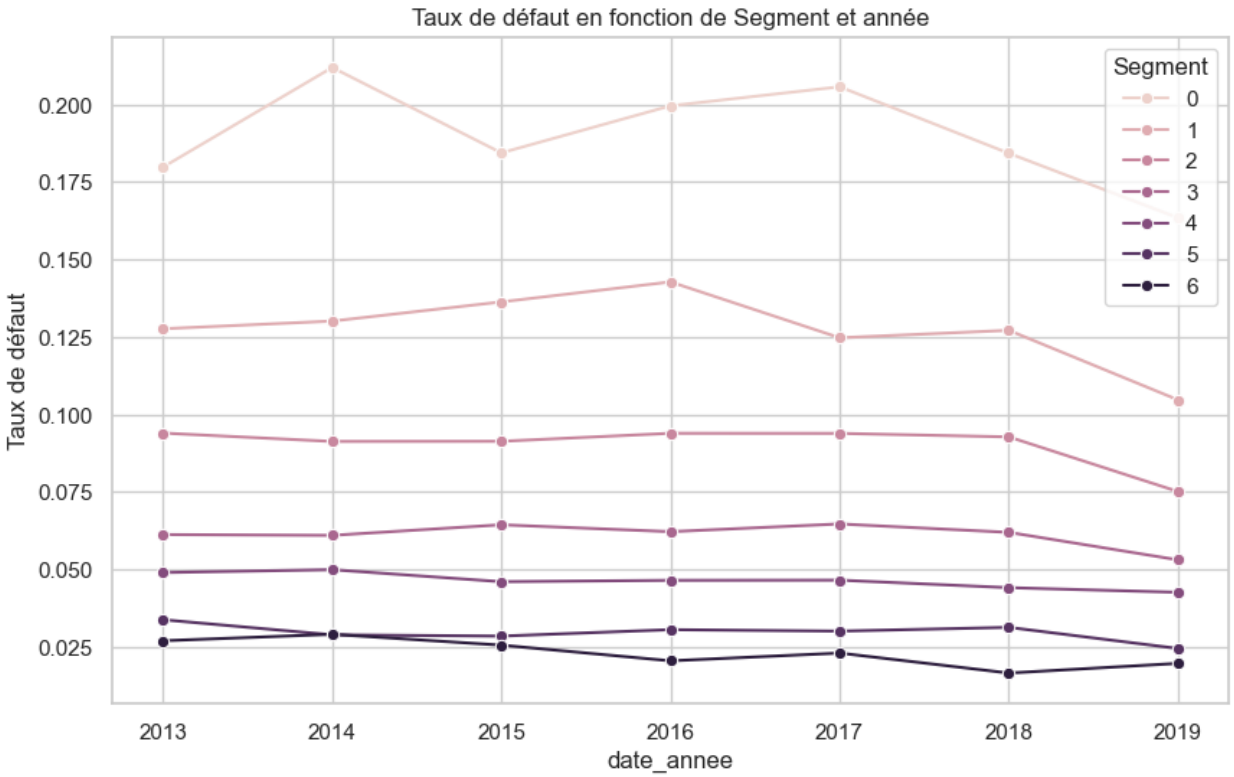

FIGURE 13 – Taux de défaut en fonction des CHR dans le temps

8.1.1 LRA

La LRA est donnée par :

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

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

8.2 Marge de Conservatisme (MoC)

Dans cette section, après avoir estimé le taux de défaut et créé les CHR, nous allons nous intéresser aux marges de conservatisme qui correspondent à différents types d'incertitudes.

8.2.1 MoC C

La MoC C est une marge de prudence permettant de couvrir les erreurs d'estimation générales du modèle.

Parmi les méthodes proposées dans la littérature, nous avons choisi la méthode de l'inférence non-paramétrique. Cette méthode consiste à ré-échantillonner les individus

des classes de risque un grand nombre de fois assurant la convergence (ici 1000) et de déduire la marge à partir de la distribution des taux de défauts TTC observés sur ces échantillons bootstraps. La MoC C est l'écart entre le 90e percentile de cette distribution et la moyenne. Le calcul est effectué au moyen de l'échantillon de calibrage dit test.

8.2.2 MoC A

La MoC A est une marge de prudence ajoutée à une estimation de risque pour tenir compte des faiblesses des données et des faiblesses méthodologiques.

Dans notre cas, il a été observée une baisse significative des taux de défauts sur l'année 2019. Ce qui induit un impact sur les taux de défauts globaux.

On utilise la même méthode que pour la MoC C, à savoir l'inférence non paramétrique. L'ajustement est alors, la différence entre le taux de défaut TTC observé hors 2019 et le taux de défaut TTC sur toute la période d'estimation. Le calcul est effectué au moyen de l'échantillon de calibrage dit test.

8.3 PD

On définit la probabilité de défaut comme suit :

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

Dans notre contexte, nous avons calculé les MoC A et C.

9 Résultats

Après avoir calculé par segment les résultats on peut observer pour chaque CHR la probabilité de défaut

Segment	Note	LRA	MOC_A	MOC_C	PD
0	0-351	0.189927	0.002949	0.007160	0.200036
1	352-467	0.127641	0.001573	0.003748	0.132963
2	468-565	0.090319	0.001076	0.002930	0.094325
3	566-658	0.061191	0.000917	0.002443	0.064551
4	659-752	0.046346	0.000738	0.002046	0.049131
5	753-854	0.029610	0.000715	0.002136	0.032460
6	855-1000	0.023005	0.001384	0.003977	0.028366

TABLE 3 – Résultats PD par segment