Détection des clients à risque de défaut de paiement

Raphaël Pinault - Camille Pousse - Cindy Puech

Table des matières

1 Introduction **3**

2 Préparation des données **3**

2.1 Choix de la population . . . 3

2.2 Critère à expliquer . . . 4

2.3 Division entre bases d'entraînement et de test . . . 5

2.4 Sélection des variables explicatives . . . 5

2.5 Discrétisation . . . 7

2.5.1 Variables qualitatives . . . 8

2.5.2 Variables quantitatives . . . 9

2.5.3 Création de variables croisées . . . 10

2.5.4 Analyse des variables sélectionnées . . . 10

3 Modèle : régression logistique **14**

3.1 Côtes et rapports de côtes . . . 14

3.2 Présentation des modèles et interprétations . . . 15

3.2.1 Modèle 1 : Régression logistique . . . 15

3.2.2 Modèle 2 : Régression logistique résultant d'une approche stepwise . . . 16

3.3 Performance du modèle . . . 20

3.4 Application du modèle à la base de test . . . 22

4 Stabilité (IS) **22**

4.1 Stabilité des populations . . . 22

4.2 Stabilité des défauts . . . 23

5 Grille de score **23**

5.1 Application à la base de test . . . 24

5.1.1 A partir du score basé sur les coefficients . . . 24

5.1.2 A partir du score basé sur les probabilités . . . 25

5.2 Analyse des distributions conditionnelles . . . 26

5.3 Optimisation de l'utilisation du score de risque . . . 30

5.4 Limites du score . . . 31

6 Challenger le modèle **31**

6.1 Random Forest . . . 31

6.1.1 Evaluation du modèle de Random Forest . . . 32

6.2 Arbre de décision . . . 32

6.2.1 Evaluation du modèle d'arbre de décision . . . 33

6.3 Avantages et inconvénients . . . 34

7 Conclusion **35**

1 Introduction

Dans le secteur bancaire, la détection des "mauvais" clients représente un enjeu crucial pour la stabilité financière et la pérennité des institutions. La capacité à anticiper et à identifier les clients présentant un risque élevé de défaut de paiement revêt une importance stratégique majeure. En effet, cette démarche permet aux banques d'optimiser la gestion de leurs risques financiers en mettant en place des politiques adaptées. La maîtrise des risques contribue non seulement à protéger les intérêts financiers de l'institution, mais elle favorise également une utilisation judicieuse des ressources, prévient la fraude, et garantit la conformité aux réglementations en vigueur. Dans ce contexte, l'application de modèles économétriques ou de machine learning joue un rôle essentiel en fournissant des outils prédictifs permettant d'identifier les signaux avant-coureurs de défaut de paiement et d'adopter des mesures proactives.

Ce rapport est divisé en six sections principales, chacune répondant à des aspects spécifiques de notre étude. Nous débutons par la section concernant la préparation des données qui détaille les différentes étapes, notamment le choix de la population étudiée, le critère à expliquer, la division entre les bases d'entraînement et de test, la sélection des variables explicatives et la discrétisation des variables. La section suivante présente les modèles de régression logistique étudiés, avec une explication des concepts des côtes et rapports de côtes, une présentation des modèles et de leurs interprétations, ainsi qu'une évaluation de leur performance. Nous abordons ensuite la stabilité des modèles, en examinant la stabilité des populations et des défauts. La grille de score est ensuite développée, avec une explication de son application à la base de test, une analyse des distributions conditionnelles, une optimisation de son utilisation et l'identification de ses limites. Nous explorons ensuite des approches alternatives en challengeant le modèle avec des modèles de Random Forest et d'Arbre de décision, en évaluant leur performance. Enfin, nous résumons les principaux résultats obtenus à travers cette analyse.

2 Préparation des données

2.1 Choix de la population

Il est nécessaire de restreindre notre analyse à une sous-partie de la population des clients de la banque patrimoniale, afin de se concentrer exclusivement sur les clients les plus importants de la banque. En effet, se concentrer sur une population cible de la banque patrimoniale permet au modèle d'être plus précis dans ses prédictions afin de mieux anticiper les défauts de paiement au sein de cette sous-partie de la population. Ainsi, nous ne prenons en compte que les caractéristiques et les comportements spécifiques à cette sous population.

Pour ce faire, nous avons limité notre analyse aux clients ayant un statut juridique de personne physique, représentant 97% de la base, en excluant les clients dont le statut juridique n'est pas renseigné ou ceux de personne morale, c'est-à-dire les entreprises.

Dans le même objectif de ne prendre en compte que les clients cibles de la banque, notre analyse

est restreinte aux statuts économiques suivants : particulier, personnel et assimilé, profession libérale et entrepreneur individuel. Nous avons choisi d'inclure certains clients qui agissent sous le nom d'une entreprise tel que les professionnels libéraux et les entrepreneurs car ils peuvent souvent avoir des comportements financiers similaires à ceux des particuliers. Notament, ils gérent leurs finances de manière indépendante et peuvent avoir des besoins financiers personnels étroitement liés à leur activité professionnelle.

Enfin, nous avons filtré les données en fonction de l'âge pour ne conserver que les clients majeurs (âgés de 18 ans ou plus) et ayant moins de 80 ans. Cette sélection s'explique par le fait qu'il est peu probable qu'un prêt soit autorisé pour des enfants, et que certaines personnes d'un certain âge pourraient rencontrer des difficultés pour rembourser le prêt jusqu'à son terme. Etant donné que notre étude porte sur une agence patrimoniale, il est cohérent de conserver des clients jusqu'à l'âge de 80 ans, car même à un âge avancé, de nombreux individus continuent à gérer activement leurs finances et à rechercher des solutions d'investissement et de gestion patrimoniale pour assurer leur sécurité financière à long terme et celle de leur famille.

2.2 Critère à expliquer

Une fois la population sélectionnée, notre attention se porte sur le critère à expliquer, à savoir une indication permettant de déterminer si le client a fait au moins un défaut de paiement dans les 12 mois suivant la date d'observation. La proportion de clients ayant fait défaut s'élève à 1,62%. Ce taux est particulièrement préoccupant dans notre contexte, puisque nous souhaitons prédire la probabilité que le client fasse un défaut de paiement. Par conséquent, si les cas de défauts de paiement sont sous-représentés au sein la banque patrimoniale, notre modélisation risque d'être biaisée et de prédire majoritairement des non-défauts ce qui peut générer un nombre élevé de faux négatifs, c'est-à-dire que le modèle pourrait manquer des cas de défaut réel.

Étant donné que les cas de défaut sont sous-représentés, nous avons procédé à une technique d'up-sampling dans le but d'atteindre un taux de défaut de 20%. Pour cela, nous avons dupliqué de manière aléatoire avec remise 17 634 lignes de la base de données où le client présente un défaut de paiement (c'est à dire les lignes où la variable cible est égale à 1). Pour trouver le nombre de lignes à ajouter dans notre base dans le but d'obtenir un taux de défaut de 20%, nous avons effectué le calcul suivant :

$$\frac{0.2 \times \text{nombre de clients total} - \text{nombre de clients à défaut}}{0.8} = \frac{0.2 \times 76764 - 1246}{0.8} = 17634$$

Bien que cette méthode apporte un biais, elle est tout de même préférable à la méthode de down-sampling qui consiste à supprimer les clients n'ayant pas de défaut de paiement jusqu'à atteindre un taux de client ayant un défaut de paiement de 20%. Étant donné qu'il n'y a que 1 246 clients présentant un défaut de paiement, il aurait fallut réduire la base à 6 230 clients ($= \frac{1 \times 1246}{0.2}$) pour obtenir un taux de défaut de 20% ce qui aurait entraîné une perte d'informations bien plus significative. Cette méthode aurait introduit un biais beaucoup plus important dans l'analyse que celle de l'upsampling.

Une alternative à la méthode d'upsampling aurait été d'utiliser la méthode de proxy. Cette méthode consiste à utiliser une variable de substitution qui capture approximativement la même information que notre variable dépendante, en remplacement de celle-ci. En d'autres termes, au lieu de baser la détection du défaut de paiement sur les 12 derniers mois, nous aurions pu intégrer une variable externe visant à mieux représenter la population sous-représentée. Cependant, l'adoption d'une proxy complexifie le modèle, rendant ainsi sa généralisation plus difficile. De plus, parmi les variables disponibles, aucune ne semble réellement apte à capturer de manière significative les caractéristiques spécifiques de la population sous-représentée. C'est pourquoi nous avons jugé plus pertinente l'utilisation de la méthode d'upsampling pour pallier le déséquilibre dans les données.

2.3 Division entre bases d'entraînement et de test

Nous divisons ensuite notre base de données en deux ensembles distincts : la base d'apprentissage, sur laquelle le modèle sera entraîné, et la base de test, sur laquelle le modèle sera testé. Cette démarche permet d'évaluer notre modèle tout en évitant d'utiliser l'intégralité des données afin d'éviter un surapprentissage du modèle. Le surapprentissage se produit lorsque le modèle s'ajuste de manière trop spécifique aux données d'entraînement, compromettant ainsi sa capacité à se généraliser sur de nouvelles données. Dans cette optique, nous avons choisi de sélectionner 30% de la base pour la base de test et 70% de la base pour la base d'apprentissage. En procédant à cette répartition, nous avons veillé à maintenir le même pourcentage de défaut dans les deux bases, soit 20%, assurant ainsi la cohérence des critères explicatifs entre les deux ensembles de données.

2.4 Sélection des variables explicatives

Nous avons initialement identifié les variables explicatives qui semblaient les plus pertinentes et les plus susceptibles de présenter un lien de causalité avec la variable à expliquer.

Nous avons en suite créé des ratios dans le but de rassembler plus d'informations, de capter des relations complexes entre les variables et de normaliser certaines variables. Premièrement, nous avons crée un ratio de mouvement qui mesure le mouvement financier en calculant la différence entre la somme des crédits et la somme des débits, normalisée par la somme du solde du compte courant et du montant d'épargne. Ce ratio donne une indication sur la dynamique des flux monétaires du clients en fonction de son capital.

$$\text{mouvement} = \frac{\text{sum\_credit} - \text{sum\_debit}}{\text{solde\_cpte\_courant} + \text{mtn\_epargne} + 1}$$

Deuxièmement, nous avons élaboré un ratio qui évalue la capacité de couverture de liquidité en prenant le rapport entre le capital restant dû et la somme des liquidités. Une couverture insuffisante pourrait signaler des difficultés à honorer les obligations financières, tandis qu'une couverture solide indiquerait

une capacité à faire face aux échéances de dettes.

$$\text{liquidity\_coverage} = \frac{\text{capital\_restant\_du}}{\text{mtn\_liqui} + 1}$$

Si la somme de *solde_cpte_courant* + *mtn_epargne* ou *mtn_liqui* est égale à 1, et que certaines observations ont un dénominateur égal à 0, nous attribuons un ratio de 0 dans ces cas-là.

Nous avons en suite restreint notre sélection de variables explicatives en respectant deux critères. Premièrement, nous avons choisi uniquement les variables présentant une corrélation significative avec la variable à expliquer. Deuxièmement, nous réduisons notre ensemble de variables explicatives sélectionnées en examinant les corrélations entre elles. En effet, il est crucial que les variables explicatives ne présentent pas de corrélation élevée entre elles afin de réduire la colinéarité des variables pour minimiser le problème d'additivité. Une corrélation supérieure à 0.4 entre deux variables indique une forte corrélation, ce qui implique qu'il est nécessaire de conserver une seule variable pour éviter les biais dans l'estimation des effets. Lorsque deux variables explicatives sont fortement corrélées entre elles, nous sélectionnons celle qui présente la corrélation la plus élevée avec la variable à expliquer. Pour obtenir un premier aperçu des corrélations entre les variables, nous avons généré un graphique de corrélation. Cela nous permet de visualiser rapidement les relations entre les différentes variables quantitatives de notre ensemble de données.

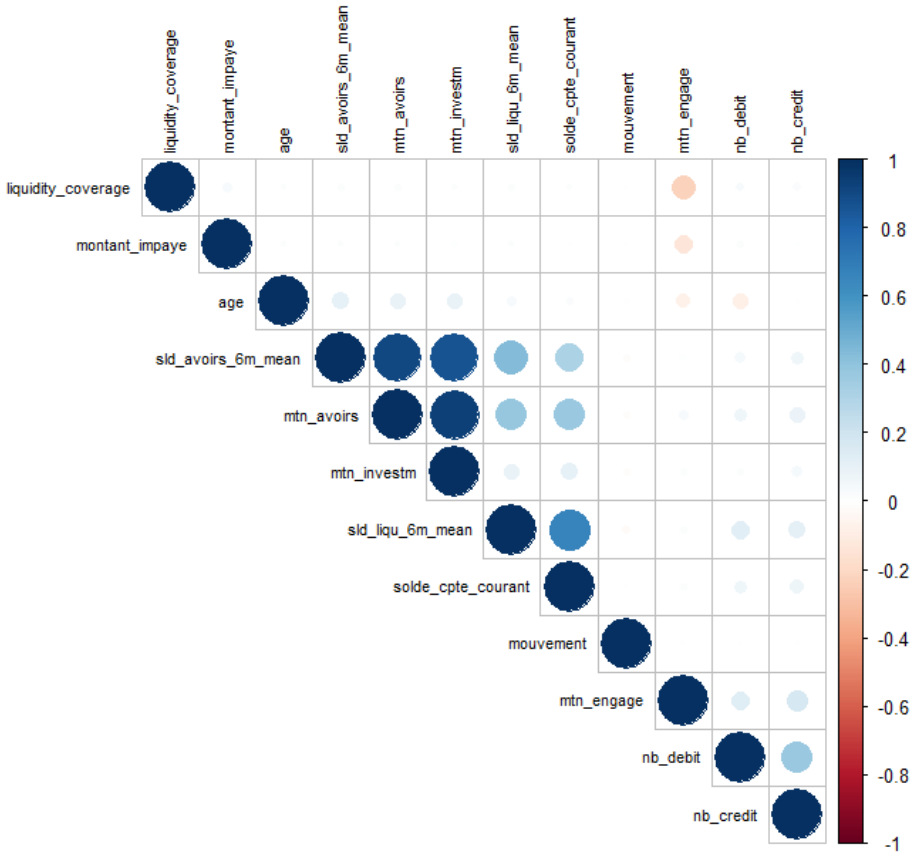

FIGURE 1 – Graphique de corrélation

Par exemple, nous avions préalablement sélectionné la variable correspondant au montant du solde du compte courant et celle correspondant au montant moyen des liquidités durant les 6 derniers mois. Ces deux variables présentant une corrélation positive de 0.66. Nous avons donc conservé une seule de ces deux variables, en choisissant celle qui affiche la corrélation la plus élevée avec la variable cible, à savoir le montant moyen des liquidités durant les 6 derniers mois. Ainsi, suite à cette sélection, nous avons choisi d'étudier les variables quantitatives suivantes :

- Ratio de la différence de la somme des crédits et des débits sur la somme du solde du compte courant et du montant d'épargne (*mouvement*)
- Ratio entre le capital restant dû et la somme des liquidités (*liquidity_coverage*)
- Age (*age*)
- Montant des investissements (*mtn_investm*)
- Nombre de flux débiteurs (*nb_debit*)
- Nombre de flux créditeurs (*nb_credit*)
- Montant moyen des liquidités durant les 6 derniers mois précédent la date d'arrêté (*sld_liqu_6m_mean*)

Après avoir transformé les variables *montant_impaye* et *mtn_engage* en dummy, nous obtenons cet ensemble de variables qualitatives :

- Indicatrice indiquant un montant des engagements non nul (*dummy_mtn_engage*)
- Indicatrice indiquant un impayé non nul (*dummy_impaye*)
- Catégorie sociaux professionnel (*CSP*)
- Situation familiale (*sit_familiale*)
- Détention ou non d'un prêt (*top_credit*)
- Détention ou non d'un crédit immobilier (*topCreditImmo*)
- Détention ou non d'un crédit (*topCred*)
- Détention ou non d'une facilité de paiement (*topFacilite*)
- Détention ou non d'un compte joint (*top_compte_joint*)
- Détention ou non d'une garantie CNP (*top_Gar_Cnp*)
- Détention ou non d'un prêt in fine (*top_pret_infine*)
- Détention ou non d'un découvert non autorisé (*topDecNonAuto*)
- Détention ou non d'une gestion de portefeuille (*topGestionPTF*)

En cas d'âge manquant, nous attribuons une valeur de -1 pour indiquer cette absence. Cette absence d'information peut être interprétée comme le fait que le banquier ou la personne enregistrant les informations du client n'a pas jugé utile de renseigner cette donnée, ou bien savait déjà pertinemment que cette information était associée à un risque potentiel.

2.5 Discrétisation

Nous avons en suite procédé à la discrétisation des variables dont le but est de réduire davantage la corrélation entre elles. La discrétisation des variables permet de rendre le modèle moins sensible aux valeurs atypiques. En regroupant les observations dans des catégories, le modèle est moins sensible aux

fortes variations. De plus, les variables discrétisées sont plus facilement interprétables et permettent de mettre en lumière des seuils critiques.

2.5.1 Variables qualitatives

Pour discrétiser les variables qualitatives, nous regroupons les modalités présentant des défauts similaires afin d'établir une distinction sur les défauts entre les groupes qui sont constitués d'au moins 5% de la population. Cette même méthode de découpage est appliquée à la base de test. Cela nous a mené à créer des classes pour les variables de la classe socio-professionnelle et de la situation familiale présentées dans les tableaux ci dessous.

Modalité	Classe	% de défaut
Ouvriers	Travailleurs indépendants	45%
Agriculteurs	Travailleurs indépendants	34%
Autre	Travailleurs indépendants	33%
Artisans, commerçants, chefs d'entreprise	Travailleurs indépendants	29%
Non renseigné	Divers	22%
Cadres, professions supérieures	Divers	21%
Professions intermédiaires	Divers	21%
Employés	Divers	20%
Sans activité	Divers	18%
Retraités	Retraités	12%

Table 1 – Classe de la variable CSP

Modalité	Classe	% de défaut
Séparé(e)	Séparé(e), divorcé(e), célibataire ou marié(e)	38%
Divorcé(e)	Séparé(e), divorcé(e), célibataire ou marié(e)	21%
Célibataire	Séparé(e), divorcé(e), célibataire ou marié(e)	20%
Marié(e)	Séparé(e), divorcé(e), célibataire ou marié(e)	20%
Veuf(ve)	Veuf(ve), NR ou pacs	15%
Non renseigné	Veuf(ve), NR ou pacs	0%
Pacs	Veuf(ve), NR ou pacs	0%

Table 2 – Classe de la variable situation familiale

En ce qui concerne la variable de la situation familiale, nous avions envisagé de séparer la modalité "séparé(e)" des autres modalités, étant donné le taux élevé de défaut associé à cette catégorie. Cependant, le nombre de clients dans cette modalité était trop faible (moins de 5%), ce qui aurait pu fausser les résultats des modèles. De même, pour la catégorie "ouvriers" de la variable indiquant la catégorie

sociale professionnelle qui indique un taux de défaut de 45% mais qui ne possède que 289 clients. De plus, nous avons également ajusté les catégories de la variable indiquant l'absence de découvert autorisé en attribuant la valeur 1 lorsque la variable *topDecNonAut* est égale à 1 ou 2, et 0 sinon.

2.5.2 Variables quantitatives

Pour la discrétisation des variables quantitatives continues, le processus commence en divisant la population en 20 classes, ce qui représente 5% de la population dans chaque classe. Pour chaque classe nous calculons le niveau de risque, c'est-à-dire la moyenne de défaut par classe. Ensuite, nous réduisons progressivement le nombre de classes en calculant, à chaque étape, le coefficient de Cramér, également connu sous le nom de V de Cramér. Ce coefficient est une mesure qui pondère les modalités d'une variable catégorielle et est défini comme suit :

$$V = \sqrt{\frac{\chi^2}{n \times \min(k-1, r-1)}} = \sqrt{\frac{\chi^2}{n \times \min(k-1, 2-1)}} = \sqrt{\frac{\chi^2}{n \times 1}} = \chi^2_{\text{normé}}$$

avec

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

où n est le nombre total d'observations, k est le nombre de classes pour la variable discrétisée, r est le nombre de catégories de la variable cible ; et O_{ij} est la fréquence observée dans la cellule i, j du tableau de contingence et E_{ij} est la fréquence attendue dans la cellule i, j.

Etant donné que la variable à expliquer n'a que deux modalités, le V de Cramér est égal au $\chi^2_{\text{normé}}$ dans ce cas là. Cependant, nous voulions une mesure qui pondère le nombre de modalités, c'est pourquoi nous avons légèrement modifié le V de Cramér en divisant par $n \times (k-1)$ au lieu de diviser par $n \times \min(k-1, r-1)$ dans le but de prendre en compte le nombre de modalités dans le calcul. Voici le coefficient de Cramér modifié que nous avons donc calculé :

$$V_{\text{mod}} = \sqrt{\frac{\chi^2}{n \times (k-1)}}$$

Le nombre optimal de classes obtenu pour chaque variable correspond à celui qui maximise le coefficient de Cramér modifié :

— 2 classes pour la variable *age*
— 2 classes pour la variable *mouvement*
— 9 classes pour la variable *liquidity_ coverage*
— 4 classes pour la variable *mtn_ investm*
— 2 classes pour la varibale *nb_ debit*
— 2 classes pour la varibale *nb_ crebit*
— 3 classes pour la variable *sld_ liqu_ 6m_ mean*

2.5.3 Création de variables croisées

Après l'obtention de ces classes, nous avons créé deux variables croisées. La première représente l'interaction entre la catégorie socio-professionnelle et l'âge (*int_ CSP_ age*), car nous prévoyons une corrélation entre ces deux variables. Cette variable pourrait être utile pour identifier des segments de la population susceptibles d'avoir des comportements différents par rapport à la variable cible. De la même manière, nous anticipons une corrélation entre les variables *mtn_ investm* et *topGestionPTF*, étant donné qu'elles concernent toutes deux des investissements. C'est pourquoi notre deuxième variable croisée, nommée *int_ mtn_ investm_ topGestionPTF*, est l'intéraction entre ces deux variables. La création de ces variables vise à fournir des informations plus riches sur la relation entre ces variables et la variable cible ainsi qu'à capturer le phénomène de colinéarité.

2.5.4 Analyse des variables sélectionnées

Nous recalculons le coefficient de Cramér pour chaque paire de variables discrétisées afin de détecter toute corrélation entre elles, et nous excluons celles dont la corrélation dépasse 0.4. De plus, nous examinons le V de Cramér entre les variables explicatives et la variable cible pour évaluer si elles maintiennent toujours une relation de causalité.

Tout d'abord, on note que la variable *mouvement* a un V de Cramér de 0.47 avec la variable *nb_ debit* et de 0.58 avec la variable *nb_ credit*. Le coefficient de Cramér avec la variable cible étant plus faible avec la variable *mouvement* qu'avec les variables *nb_ credit* ou *nb_ debit*, nous décidons de supprimer la variable *mouvement*. De même, la variable *topFacilite* affiche un coefficient de Cramér de 0.49 avec la variable *nb_ credit*. Etant donné que le coefficient de Cramér est plus élevé entre la variable cible et la variable *nb_ credit* qu'avec la variable *topFacilite*, nous avons supprimé la variable *topFacilite*. Ce processus a été répété pour chaque paire de variables explicatives, ce qui nous a conduit à supprimer les variables suivantes : *mouvement, topFacilite, nb_ debit, topCreditImmo, top_ Gar_ Cnp, age, CSP, topGestionPTF et mtn_ investm* et la dummy de la variable *mnt_ engage*.

Dans un second temps, nous examinons le coefficient de Cramér entre les variables explicatives discrétisées que nous avons sélectionnées et la variable cible. Les résultats sont présentés ci-dessous :

Nous observons que les variables les plus influentes sur la variable cible, en termes de coefficient de Cramér, sont la détention ou non d'un découvert autorisé par le client (*topDecNonAut*), le montant moyen de ses liquidités durant les 6 derniers mois (*sld_ liqui_ 6m_ mean*) et le nombre de ses flux créditeurs (*nb_ credit*). En revanche, l'information sur la situation familiale du client (*sit_ familiale*), le fait qu'il ait un prêt ou non (*top_ credit*), ainsi que le fait qu'il ait un prêt in fine ou non (*top_ pret_ infine*), semblent être étroitement liés au risque de faire défaut de paiement. La matrice des V de Cramer de l'ensemble des variables sélectionnées est présentée en annexe.

Variables	Coefficient de Cramér
topDecNonAut	0.56
sld_liqu_6m_mean	0.45
nb_credit	0.21
int_investm_topGestionPTF	0.16
liquidity_coverage	0.14
int_CSP_age	0.12
dummy_impaye	0.12
topCred	0.10
top_compte_joint	0.05
sit_familiale	0.02
top_credit	0.02
top_pret_infine	0.01

TABLE 3 – Coefficients de Cramér de la variable top_def_12m_90j

Ci-dessous sont présentés les tableaux indiquant le pourcentage d'observations dans chaque modalité des variables ainsi que le taux de défaut par modalité pour chaque variable.

Modalité	% observation	% défaut
0	86.6%	11.3%
1	13.4%	76.5%

TABLE 4 – Variable topDecNonAut

Environ 13% des clients ont un découvert non autorisé et affichent un taux de défaut de 76%, tandis que ceux qui n'ont pas de découvert non autorisé ont un taux de défaut de seulement 11%.

Modalité	% observation	% défaut
$[-55900, 639]$	33%	45.2%
$(639, 7860]$	33%	10.4%
$(7860, 4990000]$	33%	4.4%

TABLE 5 – Variable sld_liqu_6m_mean

La variable *sld_liqu_6m_mean* est discrétisée en trois classes égales, chacune contenant 33% des observations. Les taux de défaut diffèrent selon les modalités, avec le taux le plus élevé (45.2%) observé dans le premier intervalle, couvrant les valeurs de -55 900 à 639 pour la liquidité moyenne sur les six derniers mois. On remarque que le taux de défaut est décroissant avec la modalité des liquidités, c'ets à dire que plus un client a de liquidités, moins il semble à risque.

Modalité	% observation	% défaut
$[0, 1]$	59.1%	26.9%
$(1, 958]$	40.9%	10.0%

TABLE 6 – Variable nb_credit

Environ 59% des observations de la variable *nb_credit* sont comprises entre 0 et 1, tandis que le

reste (environ 41%) se trouve entre 1 et 958. Le taux de défaut est significativement plus élevé pour les clients n'ayant que 0 ou 1 flux créditeurs (27%) par rapport ceux qui en ont entre 1 et 958 (10%).

Modalité	**% observation**	**% défaut**
$[-6290, 0]_0$	51.4%	21.5%
$(0, 33000]_0$	0.0%	NaN
$(33000, 20500000]_0$	0.0%	NaN
$[-6290, 0]_1$	3.1%	31.5%
$(0, 33000]_1$	20.5%	27.1%
$(33000, 20500000]_1$	25.0%	9.7%

TABLE 7 – Variable int_investm_topGestionPTF

Pour la variable *int_ investm_ topGestionPTF*, nous avons six classes dont deux qui sont vides : $(0, 33000]_0$ et $(33000, 20500000]_0$. Ceci signifie qu'aucun client ne se trouve dans le cas où il possède un montant d'investissement compris entre 0 et 20 500 000 sans détenir une gestion de portefeuille. La moitié des clients (51%) sont dans la situation où leur investissement est inférieur à 0 sans avoir de gestion de portefeuille, avec un taux de défaut de 21.5%. Les clients se trouvant dans le même intervalle d'investissement mais avec une gestion de portefeuille représente 3% des observations avec un taux de défaut de 32%. Enfin, les clients possédant une gestion de portefeuille et ayant des investissements compris entre 0 et 33 000, ou entre 33 000 et 20 500 000, ont un taux de défaut de 27% et 10% respectivement.

Modalité	**% observation**	**% défaut**
$[-6950000, 0]$	98.5%	19.3%
$(0, 1140000]$	1.5%	64.8%

TABLE 8 – Variable liquidity_coverage

Le tableau pour la variable *liquidity_ coverage* présente deux classes. La première couvrant l'intervalle de -6 950 000 à 0 contient 98.5% des observations, avec un taux de défaut de 19%. La seconde classe, allant de 0 à 1 140 000, ne compte que 1.5% des observations, mais avec un taux de défaut notablement plus élevé de 65%. Bien que cela puisse sembler contre-intuitif, nous attribuons ce résultat au faible nombre d'observations dans cette classe.

Modalité	**% observation**	**% défaut**
Travailleurs indépendants [-1,58]	4.8%	33.1%
Divers [-1,58]	46.5%	21.8%
Retraites [-1,58]	0.2%	10.6%
Travailleurs indépendants (58,80)	4.6%	28.7%
Divers (58,80)	26.7%	17.8%
Retraites (58,80)	17.1%	12.4%

TABLE 9 – Variable int_CSP_age

La variable *int_ CSP_ age* est discrétisée en six classes, chacune associée à une catégorie socio-professionnelle (CSP) et à une tranche d'âge spécifique. Pour les travailleurs indépendants âgés de

moins de 58 ans, le taux de défaut est de 33%, tandis que pour ceux âgés de 58 à 80 ans, ce taux diminue à 29%. Les individus classés dans la catégorie Divers présentent des taux de défaut de 22% pour les moins de 58 ans et de 18% pour les 58 à 80 ans. On constate donc que pour ces deux catégories socio-professionnelles, le taux de défaut tend à diminuer avec l'âge. Cela peut s'expliquer par le fait que les personnes plus âgées ont souvent travaillé pendant plus longtemps et ont pu économiser davantage d'argent. De plus, elles ont moins de charges financières, comme le remboursement de prêts immobiliers, ce qui les expose moins au risque de défaut de paiement. Cependant, cette tendance s'inverse pour les clients retraités, où le taux de défaut est de 11% pour les moins de 58 ans (représentant seulement 0.2% des observations) et de 12% pour les 58 à 80 ans.

Modalité	**% observation**	**% défaut**
0	99.96%	6.2%
1	0.04%	93.8%

TABLE 10 – Variable dummy_impaye

La variable *dummy_impaye* est une variable binaire qui prend la valeur 1 en cas d'impayé par le client. Seulement 0.04% des clients ont un impayé, avec un taux de défaut très élevé à 94%, ce qui semble cohérent. Les clients n'ayant jamais eu d'impayé présentent un taux de défaut de 6%.

Modalité	**% observation**	**% défaut**
0	93.7%	19.0%
1	6.4%	35.2%

TABLE 11 – Variable topCred

Environ 94% des clients ne possèdent pas de crédit, avec un taux de défaut de 19%. En revanche, ceux qui en possèdent un présentent un taux de défaut plus élevé, à 35%.

Modalité	**% observation**	**% défaut**
0	54.5%	21.8%
1	45.5%	17.8%

TABLE 12 – Variable top_compte_joint

La variable *top_compte_joint* est une variable binaire où 1 signifie que le client détient un compte joint. Environ 54.5% des clients ne possèdent pas de compte joint, avec un taux de défaut de 22%. En revanche, ceux qui en possèdent un (représentant 45.5% des clients) ont un taux de défaut légèrement inférieur, à 18%. Cela suggère que les clients qui détiennent un compte joint ont une taux de défaut légèrement plus bas que ceux qui n'en ont pas.

Modalité	% observation	% défaut
Séparé(e), divorcé(e), célibataire ou marié(e)	97.4%	20.2%
Veuf(ve), NR ou pacs	2.6%	14.3%

Table 13 – Variable sit_familiale

Concernant la situation familiale des clients, la majorité (97%) sont séparées, divorcées, célibataires ou mariées, avec un taux de défaut de 20%. Les clients veufs, non renseignés (NR) ou en union civile (PACS), ne représente que 2.6% des observations, mais affiche un taux de défaut plus bas, à 14%, probablement en raison de responsabilités financières potentiellement moins importantes.

Modalité	% observation	% défaut
0	88.1%	20.3%
1	11.9%	17.8%

Table 14 – Variable top_credit

12% des clients détiennent un prêt avec un taux de défaut de 18% contre 20% pour les clients qui n'en détiennent pas. Ceci peut s'expliquer par le fait que les emprunteurs ont probablement été soumis à des critères de solvabilité plus stricts, ce qui reflète une capacité financière plus solide et une meilleure gestion des risques.

Modalité	% observation	% défaut
0	99.7%	20.0%
1	0.3%	23.8%

Table 15 – Variable top_pret_infine

La majorité des clients (99.7%) ne possèdent pas de prêt in fine, présentant un taux de défaut de 20%, tandis que ceux qui en détiennent un (0.3% des clients) affichent un taux de défaut plus élevé, à 24%.

3 Modèle : régression logistique

3.1 Côtes et rapports de côtes

Etant donné que la variable cible est une variable binaire (encodée en 0 ou 1), nous avons établi un premier modèle de régression logistique. Cette méthode, couramment utilisée dans le domaine de la classification et de la prédiction permet de modéliser la relation entre une variable binaire dépendante et une ou plusieurs variables indépendantes. Elle produit des probabilités comprises entre 0 et 1, permettant ainsi de classer les observations dans différentes catégories.

Les coefficients obtenus avec une régression logistique permettent d'estimer des rapports de côte (ODD ratio) pour chacune des variables indépendantes. Ils représentent la contribution de chaque variable indépendante à la probabilité d'observer l'évènement de la variable dépendante binaire. Plus

précisément, un coefficient positif indique une augmentation de la probabilité d'observer l'événement lié à la variable dépendante, tandis qu'un coefficient négatif indique une diminution de cette probabilité. En général, un coefficient plus grand en valeur absolue implique une influence plus importante sur la probabilité. Le rapport de côte (ODD ratio) est le rapport des chances de deux évènements où la chance (ODD) d'un évènement est défini comme suit (et correspond à la fonction logit) :

$$\text{Odds}(p) = \log\left(\frac{p}{1-p}\right)$$

où p représente la probabilité de l'évènement et $1-p$ la probabilité de l'absence de l'évènement. Le rapport de côte (ODD ratio) est donc :

$$\text{OR}(p_1, p_0) = \frac{\text{Odds}(p_1)}{\text{Odds}(p_0)} = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \frac{p_1}{1-p_1} \times \frac{1-p_0}{p_0}$$

où p_0 est la probabilité d'obersver l'évènement 0 et p_1 est la probabilité d'observer l'évènement 1.

Un rapport de côte de 1 indique des chances égales d'observer l'évènement, un rapport de côte supérieur à 1 indique une augmentation des chances, et un rapport de côte inférieur à 1 indique une diminution des chances.

3.2 Présentation des modèles et interprétations

Avant d'implémenter nos modèles, nous réévaluons d'abord les niveaux de chaque variable. Cela implique la sélection de la modalité qui servira de référence dans le modèle, en excluant celle présentant le plus haut niveau de risque, c'est-à-dire la modalité affichant le taux de défaut le plus élevé pour chaque variable. Ce choix nous permettra par la suite d'interpréter les coefficients, en analysant les modalités de risque décroissant par rapport à cette référence. Ainsi, nous anticipons que tous les coefficients de nos modèles seront négatifs et que leur amplitude augmentera avec le niveau de risque de la modalité. Les modalités les plus risquées par variables sont présentées ci dessous.

3.2.1 Modèle 1 : Régression logistique

Nous implémentons notre premier modèle de régression logistique en utilisant le code R suivant où nous spécifions la famille de distribution comme binomiale avec la fonction de lien logit présentée précédemment. Les résultas du modèles sont présentés en annexes.

```r
model <- glm(top_def_12m_90j ~ sit_familiale + top_credit +  top_
    compte_joint + top_pret_infine + topDecNonAut + topCred + liquidity
    _coverage + dummy_impaye + nb_credit + sld_liqu_6m_mean + int_CSP_
    age + int_investm_topGestionPTF, data = train_data, family = "
    binomial")
```

Nous observons que les variables *top_pret_infine0* et *int_CSP_ageTravailleurs indépendants_(58,80]* ne sont pas significatives dans le modèle. Toutes les autres variables sont significatives à un niveau de confiance de inférieur à 0.1%, hormis les variables *sit_familialeVeuf(ve), NR ou pacs* qui est significative à 1% et *int_CSP_ageRetraites_[-1,58]* qui est significative à 0.1%. De plus, à l'exception du coefficient de la variable *top_pret_infine0* qui n'est pas significatif, tous les coefficients sont négatifs, ce qui est cohérent étant donné que nous avons choisi comme référence la catégorie présentant le taux de défaut le plus élevé pour chaque variable. Ainsi, on s'attendait à ce que le modèle prédise des défauts moins importants pour les autres modalités.

Afin de pouvoir comparer ce modèle, nous nous intéressons aux critères AIC (Akaike Information Criterion) et BIC (Bayesian Information Criterion) qui sont des mesures utilisées pour évaluer la qualité d'un modèle statistique, et sont définis comme suit :

$$\text{AIC} = -2 \times \log(L) + 2 \times k$$

$$\text{BIC} = -2 \times \log(L) + k \times \log(n)$$

où L est la fonction de vraisemblance maximale du modèle, k est le nombre de paramètres estimés dans le modèle et n est la taille de l'échantillon. Ces critères permettent de comparer différents modèles en prenant en compte à la fois la qualité de l'ajustement du modèle aux données et sa complexité. Un modèle avec un AIC ou un BIC plus faible est considéré comme préférable. Les critères AIC et BIC de ce modèle sont respectivement de 42 403 et 42 585.

3.2.2 Modèle 2 : Régression logistique résultant d'une approche stepwise

Afin de garantir que seules les variables ayant une réelle causalité sur le fait de faire défaut ou non sont conservées dans le modèle, nous implémentons une approche stepwise. L'approche stepwise est une méthode qui consiste à ajouter ou retirer sélectivement des variables du modèle de manière itérative, en évaluant l'impact de chaque ajout ou retrait sur la qualité du modèle. Pour cela nous exécutons le code R suivant :

```r
stepwise_model <- stepAIC(model, direction = "both", trace = FALSE)
```

La fonction `stepAIC()` est utilisée pour effectuer la sélection de variables stepwise basée sur le critère d'information d'Akaike (AIC). Cette sélection est faite dans les deux sens comme l'indique l'argument `direction = "both"` indiquant à la fonction de considérer à la fois l'ajout et la suppression de variables lors du processus stepwise. Cela signifie qu'à chaque étape, la fonction peut ajouter une variable si elle améliore le modèle et/ou retirer une variable si elle diminue l'AIC. Les résultats du modèle sont présentés ci dessous.

	Variable dépendante :
	top_def_12m_90j
(Intercept)	6.9182***
	(0.3095)
sit_familialeVeuf(ve), NR ou pacs	−0.1795*
	(0.0907)
top_credit1	−0.4682***
	(0.0530)
top_compte_joint	−0.1861***
	(0.0270)
topDecNonAut0	−2.6101***
	(0.0317)
topCred0	−0.6950***
	(0.0525)
liquidity_coverage[-6.95e+05,0]	−0.5504***
	(0.1075)
dummy_impaye0	−2.9493***
	(0.2740)
nb_credit(1,958]	−0.5162***
	(0.0310)
sld_liqu_6m_mean(639,7.86e+03]	−1.3772***
	(0.0299)
sld_liqu_6m_mean(7.86e+03,4.99e+06]	−1.9671***
	(0.0398)
int_CSP_ageDivers_[-1,58]	−0.4070***
	(0.0526)
int_CSP_ageRetraites_[-1,58]	−0.9737**
	(0.3282)
int_CSP_ageTravailleurs indépendants_(58,80]	−0.0668
	(0.0719)
int_CSP_ageDivers_(58,80]	−0.4883***
	(0.0555)
int_CSP_ageRetraites_(58,80]	−0.6728***
	(0.0615)
int_investm_topGestionPTF[-6.29e+03,0]_0	−0.3776***
	(0.0663)
int_investm_topGestionPTF(0,3.3e+04]_1	−0.5476***
	(0.0691)
int_investm_topGestionPTF(3.3e+04,2.05e+07]_1	−1.1548***
	(0.0718)
Observations	66 077
Deviance nulle	66 132
Deviance résiduelle	42 363
AIC	42 401
Note :	*** p<0.001 ; ** p<0.01 ; * p<0.05

Table 16 – Résultats du modèle stepwise

L'approche stepwise a supprimé la variable *top_pret_infine0* qui n'était pas significative dans le modèle précédent. Cela indique qu'un client qui ne détient pas de prêt in fine n'est pas plus à risque qu'un client qui en détient un. Les coefficients sont donc désormais tous négatifs, ce qui est attendu puisque nous avons pris comme référence la catégorie présentant le taux de défaut le plus élevé pour

chaque variable. Cela nous a permis d'obtenir des critères AIC et BIC légèrement inférieurs par rapport au modèle précédent, s'établissant à 42 401 et 42 574 respectivement.

Dans un modèle de régression logistique, les coefficients estimés représentent les variations du rapport de côte de faire défaut associés à une unité d'augmentation dans la variable explicative correspondante, toutes choses étant égales par ailleurs. Tous les coefficients étant négatifs, cela signifie que si un client appartient à une classe différente de la classe de référence pour une variable explicative donnée, alors la probabilité de défaut diminue. Par exemple, un client ayant un solde de liquidité entre 639 et 7 860 présente moins de risque de faire défaut qu'un client ayant un solde inférieur à 639 (qui est la modalité de référence). De plus, en comparant les amplitudes des coefficients, on observe que le coefficient pour la modalité *sld_liqu_6m_mean(639,7.86e+03]* est de -1.3772, tandis que celui pour la modalité *sld_liqu_6m_mean(7.86e+03,4.99e+06]* est de -1.9671. Cela indique qu'un client ayant un solde moyen de liquité sur les 6 derniers mois entre 7 860 et 4 990 000 est moins à risque de faire défaut qu'un client ayant un solde entre 639 et 7 860.

Plus précisément, les coefficients sont interprétés à l'aide de l'exponentielle du coefficient. Par exemple, pour la variable concernant la situation familiale, nous avons un coefficient de -0.1795, cela correspond à $exp(-0.1795) = 0.835$. Ceci signifie que si un individu est veuf ou pacsé, le rapport de côte de défaut est 0.835 le rapport de côte de défaut d'un client séparé, divorcé, célibataire ou marié. Autrement dit, le rapport de côte diminue de 16% $(= 1 - 0.835)$ pour un client veuf ou pacsé par rapport à un client séparé, divorcé, célibataire ou marié.

On observe que le coefficient le plus élevé en valeur absolue est celui de la variable *dummy_impaye0*, ce qui suggère que cette variable a un impact plus important que les autres variables sur la probabilité de défaut, c'est à dire que la différenciation des bons ou mauvais clients à partir de cette variable est importante. Le coefficient de cette variable est de -2.9493, ce qui correspond à $exp(-2.9493) = 0.052$. Cela indique que le rapport de côtes est significativement plus bas pour un client sans aucun impayé par rapport à un client ayant des impayés. Plus précisément, le rapport de côte diminue de 95% $(= 1 - 0.052)$.

Le deuxième coefficient le plus élevé en valeur absolue est celui de la variable *topDecNonAut0* qui est de -2.6101 avec $exp(-2.6101) = 0.074$ ce qui signifie que les clients ne détenant pas de découvert non autorisé ont une côte multipliée par 0.074 par rapport aux clients qui détiennent un découvert autorisé. Autrement dit, le rapport de côte diminue de 93% $(= 1 - 0.074)$.

Le coefficient élevé de -1.9671 attribué à la modalité (7 860, 4 990 000] de la variable *sld_liqu_6m_mean* correspond à une valeur exponentielle de $exp(-1.9671) = 0.139$. Cela signifie qu'un client ayant un montant moyent des liquidités durant les 6 derniers mois supérieur à 7 860 sur les 6 derniers mois présente une côte de 86% $(1 - 0.139)$ par rapport à un client ayant un montant des liquidités inférieur à 639, selon le modèle.

En ce qui concerne les variables comportant plusieurs modalités dans le modèle, nous nous assurons que l'amplitude des coefficients reflète correctement le niveau de risque associé à chaque modalité. En d'autres termes, nous vérifions que l'amplitude des coefficients négatifs augmente avec le niveau de

risque de chaque modalité.

Modalité	Coefficients	Niveau de risque
sld_liqu_6m_mean[-5.59e+04,639]	Réf.	45%
sld_liqu_6m_mean(639,7.86e+03]	−1.3772	10%
sld_liqu_6m_mean(7.86e+03,4.99e+06]	−1.9671	4%
int_CSP_ageTravailleurs indépendants_[-1,58]	Réf.	33%
int_CSP_ageTravailleurs indépendants_(58,80]	−0.0668	29%
int_CSP_ageDivers_[-1,58]	−0.4070	22%
int_CSP_ageDivers_(58,80]	−0.4883	18%
int_CSP_ageRetraites_(58,80]	−0.6728	12%
int_CSP_ageRetraites_[-1,58]	−0.9737	11%
int_investm_topGestionPTF[-6.29e+03,0]_1	Réf.	31%
int_investm_topGestionPTF(0,3.3e+04]_1	−0.5476	27%
int_investm_topGestionPTF[-6.29e+03,0]_0	−0.3776	21%
int_investm_topGestionPTF(3.3e+04,2.05e+07]_1	−1.1548	10%

TABLE 17 – Comparaison des coefficients et des niveaux de risque

Le tableau indique que dans l'ensemble l'amplitude des coefficients est cohérent avec le risque de défaut associé à chaque variable, à l'excepetion des modalitées *int_ investm_ topGestionPTF(0,3.3e+04]_ 1* et *int_ investm_ topGestionPTF[-6.29e+03,0]_ 0*. Ces deux modalités présentent respectivement des niveaux de risque de 27% et 21%, mais les coefficients associés sont dans le sens inverse, étant de −0.5476 et −0.3776 respectivement. Ceci peut s'expliquer par le fait que d'autres variables incluses dans le modèle peuvent influencer de manière différente l'effet de ces modalités sur la variable dépendante, ce qui peut entraîner des coefficients qui ne correspondent pas toujours de manière linéaire au niveau de risque.

A partir de ce modèle, nous obtenons les probabilités que chaque individu présente un défaut de paiement à l'aide du code R suivant :

```
y_pred_prob <- predict(stepwise_model, newdata = test_data, type = "
    response")
```

Ce code calcule les probabilités prédites en appliquant la fonction sigmoïde à la somme pondérée par les coefficients obtenus dans le modèle des variables explicatives pour chaque individu où la fonction sigmoïde est définie par :

$$f(x) = \frac{1}{1 + \exp^{-\eta_i}}$$

où η_i est la somme pondérée par les coefficients des variables explicatives pour l'observation i :

$$\eta_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \ldots + \beta_k \cdot x_{ik}$$

avec $\beta_0, \beta_1, \ldots, \beta_k$ les coefficients estimés du modèle et $x_{i1}, x_{i2}, \ldots, x_{ik}$ les valeurs des variables explicatives pour l'observation i. Cette fonction permet d'obtenir des probabilités entre 0 et 1. Une fois les probabilités obtenus, nous déterminons un seuil à partir duquel nous considérons qu'un individu a fait défaut. Ce seuil est choisi de manière uniforme et indépendante pour tous les individus, ce qui signifie que tous les individus partagent le même seuil. Nous établissons ce seuil à 0.4 afin de minimiser le nombre de faux négatif, c'est à dire le nombre d'observation où le modèle va prédire que le client n'est pas à risque de faire défaut alors que c'est le cas.

3.3 Performance du modèle

Afin d'évaluer la performance de notre modèle, nous commençons par examiner la matrice de confusion qui est une table permettant de visualiser les performances du modèles en comparant les prédictions du modèle avec les valeurs réelles de l'ensemble de données à l'aide de différents indicateurs :

- Vrai Positif (TP) : Les observations réelles appartenant à la classe positive (1) et correctement prédites comme telles par le modèle.
- Vrai Négatif (TN) : Les observations réelles appartenant à la classe négative (0) et correctement prédites comme telles par le modèle.
- Faux Positif (FP) : Les observations réelles appartenant à la classe négative (0) mais incorrectement prédites comme appartenant à la classe positive (1) par le modèle (erreur de type I).
- Faux Négatif (FN) : Les observations réelles appartenant à la classe positive (1) mais incorrectement prédites comme appartenant à la classe négative (0) par le modèle (erreur de type II).

		Références	
		0	**1**
Prédictions	**0**	21 728 (77%)	2 670 (9%)
	1	920 (3%)	2 994 (11%)

TABLE 18 – Matrice de confusion du modèle stepwise

La matrice de confusion indique qu'au total nous avons environ 88% d'observations qui sont correctement prédites par le modèle. Nous avons 9% de faux négatifs et 3% de faux positifs.

Cette matrice nous permet de calculer deux métriques importantes que nous retrouvons par la suite dans la courbe de ROC :

- La sensibilité qui mesure la capacité du modèle à détecter les véritables positifs parmi toutes les observations réelles de la classe positive. Elle est particulièrement importante lorsque minimiser les faux négatifs est crucial, comme c'est le cas ici. En effet, notre objectif est de minimiser les

situations où le modèle ne prédit pas correctement qu'un individu va faire défaut alors qu'il va effectivement faire défaut. La sensibilité est définie comme suit :

$$\text{Sensibilité} = \frac{TP}{TP + FN} = \frac{2994}{2994 + 2670} = 0.528$$

Une sensibilité de 0.528 dans notre modèle indique que parmi les clients réellement en défaut, 52.8% d'entre eux sont correctement identifiés comme tels par la prédiction du modèle.

— La spécificité qui mesure la capacité du modèle à éviter les faux positifs parmi toutes les observations réelles de la classe négative. Elle est importante lorsque minimiser les faux positifs est crucial. La spécificité est définie comme suit :

$$\text{Spécificité} = \frac{TN}{TN + FP} = \frac{21728}{21728 + 920} = 0.959$$

La spécificité du modèle étant de 0.959, cela signifie que parmi les clients qui n'ont pas fait défaut, 95.9% d'entre eux sont correctement catégorisés comme tels dans le modèle.

Nous examinons en suite la courbe ROC, qui un graphique illustrant la performance d'un modèle de classification binaire à divers seuils de classification. Sur l'axe des abscisses de la courbe ROC, nous observons le taux de faux positifs, représentant la proportion d'observations de la classe négative incorrectement identifiées comme positives (1 - spécificité). L'axe des ordonnées de la courbe de ROC illustre le taux de vrais positifs, correspondant à la proportion d'observations de la classe positive correctement classées (sensibilité). Une courbe de ROC idéale se caractérise donc par une montée rapide vers le coin supérieur gauche du graphique où la sensibilité (taux de vrais positifs) est de 1 et la spécificité (taux de vrais négatifs) est de 1, reflétant ainsi une performance parfaite du modèle. La courbe ROC ci-dessous, représentant les résultats de notre modèle sur les données d'entraînement, démontre une performance satisfaisante, car elle tend vers le coin supérieur gauche.

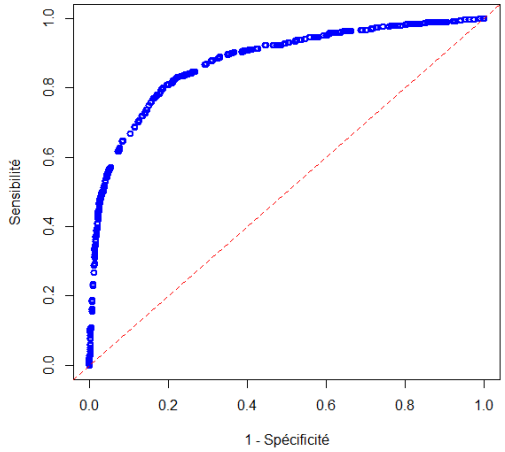

FIGURE 2 – Courbe de ROC

Nous calculons en suite l'AUC (Area Under the Curve), qui est l'aire sous cette courbe de ROC et qui mesure la capacité globale du modèle à classer correctement les observations. Une valeur proche de 1 indique une performance élevée, tandis qu'une valeur proche de 0.5 suggère une performance similaire à un modèle aléatoire. La performance d'un modèle alétoire est modélisé par la ligne rouge diagonale à 45 degrés sur le graphique. L'AUC est de 0.8771 dans notre modèle ce qui signifit que le modèle a une probabilité élevée de classer correctement les observations de la classe positive par rapport à la classe négative. Cela traduit donc une bonne performance du modèle de classification.

En se basant sur la courbe ROC, nous pouvons calculer le coefficient de Gini à partir de l'aire sous la courbe ROC de la manière suivante :

$$\text{Gini} = 2 \times \text{AUC} - 1 = 2 \times 0.8775 - 1 = 0.755$$

Cette formule exprime le coefficient de Gini comme une mesure de la capacité du modèle à classer correctement les observations. Avec un coefficient de Gini de 0.7542, notre modèle démontre une capacité robuste de discrimination entre les défauts de paiement et l'absence de défaut ce qui suggère une bonne fiabilité dans la prise de décision.

3.4 Application du modèle à la base de test

Nous appliquons ici notre modèle de régression logistique à la base de test. Nous optenons un AUC de 0.8745 et coefficient de Gini de 0.7491 ce qui représente une baisse de moins de 1% par rapport à la base d'apprentissage. Cette diminution est conforme à notre critère d'acceptation de 10% maximum de baisse du coefficient de Gini à l'application sur la base de test. De plus, une baisse de moins de 1% par rapport à la base d'apprentissage suggère que notre modèle généralise bien aux données non observées, ce qui est essentiel pour son utilisation. Cette performance élevée indique que notre modèle est capable de distinguer efficacement les clients ayant fait défaut de paiement et ceux n'ayant pas fait défaut, ce qui renforce sa fiabilité dans le secteur bancaire.

4 Stabilité (IS)

4.1 Stabilité des populations

Nous mesurons la stabilité des populations en comparant la proportion d'individu par modalité entre la base d'apprentissage et la base de test. Pour confirmer ceci, nous calculons l'Index de Stabilité (IS) suivant pour lequel nous obtenons des résultats inférieur à 0.15 (voir tableau en annexes) pour toutes les variables ce qui signifie que la population est stable.

$$SI = \sum_k (p_k - b_k) \ln\left(\frac{p_k}{b_k}\right)$$

où p_k est la proportion de population au sein la modalité k dans la base de train et b_k est la proportion de population au sein de la modalité k dans la base de test.

4.2 Stabilité des défauts

De même, nous mesurons la stabilité des défauts en comparant la proportion de défaut par modalité entre la base d'apprentissage et la base de test. Toutes les modalités présentent également un Index de Stabilité inférieur à 0.15, comme indiqué dans le tableau en annexe, ce qui suggère que la proportion de défaut reste stable.

5 Grille de score

L'objectif de la grille de score est de classer les clients, du plus risqué au moins risqué, afin de faciliter la prise de décision. Elle offre aux équipes informatiques une mesure de performance intégrable au sein d'un système d'implémentation. Choisi pour sa simplicité d'interprétation, ce tableau permet aux directeurs marketing, entre autres, de prendre des décisions stratégiques. En fonction de la situation spécifique d'un client, ils peuvent ainsi déclencher des campagnes marketing ciblées ou décider d'accorder ou de refuser un prêt.

Un score représente une mesure évaluant le risque qu'un consommateur fasse défaut à la banque, et il est défini par une combinaison de variables identifiées comme des facteurs expliquant ce risque.

La création de la grille de score, basée sur les coefficients obtenus dans la régression logistique, s'effectue de la manière suivante : nous débutons par conserver les coefficients des variables en valeur absolue. Ensuite, nous calculons la somme des coefficients des modalités présentant le risque de défaut le plus faible par variable, c'est à dire la modalité ayant le coeffcieint le plus élevé en valeur absolue parmi toutes les modalités de la variable.

Le score est ensuite calculé en divisant la valeur absolue du coefficient de chauqe modalité par la somme des coefficients associés aux modalités présentant le risque le plus faible par variable. Cette valeur est ensuite multipliée par 1000.

Le score de la modalité i est donc obtenu comme suit :

$$\text{Score coef}_i = \frac{|\text{coefficient}_i|}{\sum_{j=1}^{k} \max |\text{coefficient}_j|} \times 1000$$

où k est le nombre total de variable.

Les scores obtenus sont présentés ci-dessous. Un score élevé pour une modalité indique son impact sur l'augmentation du score global. Ainsi, plus un client obtient un score élevé, moins il est à risque.

Modalité	Score (coefficients)
dummy_impaye0	216.42
int_CSP_ageDivers_(58,80]	35.83
int_CSP_ageDivers_[-1,58]	29.87
int_CSP_ageRetraites_(58,80]	49.37
int_CSP_ageRetraites_[-1,58]	71.45
int_CSP_ageTravailleurs indépendants_(58,80]	4.90
int_investm_topGestionPTF(-6.29e+03,0)_0	27.71
int_investm_topGestionPTF(0,3.3e+04]_1	40.19
int_investm_topGestionPTF(3.3e+04,2.05e+07]_1	84.74
liquidity_coverage[-6.95e+05,0]	40.39
nb_credit(1,958]	37.88
sit_familialeVeuf(ve), NR ou pacs	13.17
sld_liqu_6m_mean(639,7.86e+03]	101.06
sld_liqu_6m_mean(7.86e+03,4.99e+06]	144.34
topCred0	51.00
top_compte_joint1	13.65
top_credit1	34.36
topDecNonAut0	191.53

TABLE 19 – Score calculé à partir des coefficients du modèle

On constate qu'un client qui n'a aucun impayé (score de 216), qui ne présente pas de découvert autorisé sur son compte (score de 191) ou qui affiche un solde de liquidité supérieur à 639 (score de 101 ou plus) est susceptible d'obtenir un score élevé et d'être classé comme un "bon client". être un travailleur indépendant âgé de 58 à 80 ans (score de 5), être veuf ou pacsé (score de 13) ou détenir un compte joint (score de 13) contribue peu à l'augmentation du score global, et donc à la probabilité d'être classé comme un "bon client".

5.1 Application à la base de test

Puisque les hypothèses de stabilité sont confirmées pour le défaut ainsi que pour les populations, nous sommes en mesure d'appliquer des scores à la base de test.

5.1.1 A partir du score basé sur les coefficients

En appliquant le score basé sur les coefficients à la base d'apprentissage, nous observons les deux distributions représentées sur le graphique ci-dessous. Les individus prédits comme étant à risque de défaut sont marqués en rouge, tandis que ceux considérés comme non risqués sont en bleu.

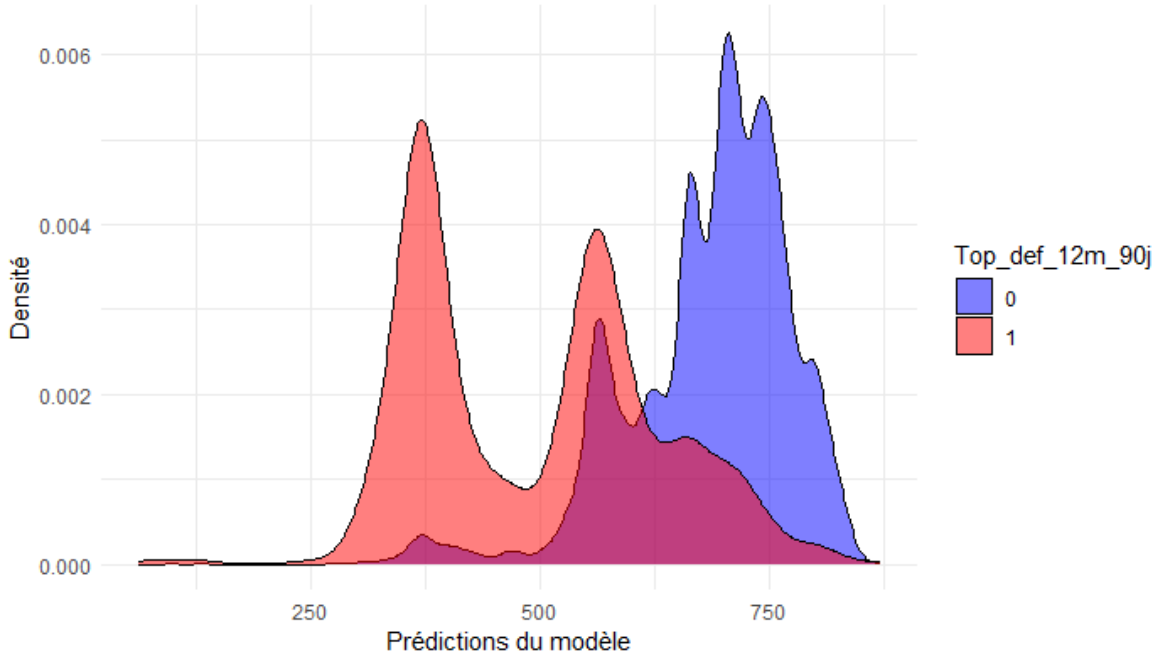

FIGURE 3 – Distribution des scores basés sur les coefficients

On observe une zone de chevauchement des deux distributions, où la prise de décision devient délicate. Nous tentons donc une approche complémentaire pour calculer un score basé sur les probabilités de chaque client d'être en défaut.

5.1.2 A partir du score basé sur les probabilités

Nous calculons un nouveau score en se basant sur les probabilités de défaut obtenues pour chaque client. Pour ce faire, nous calculons le score pour chaque client i comme suit afin d'obtenir un score faible pour les clients à risque :

$$\text{Score proba}_i = (1 - \mathbb{P}_i(\text{défaut})) \times 1000$$

De même que précédemment, plus un client obtient un score élevé, moins il est à risque. En appliquant ce score basé sur les probabilités de chaque client à la base d'apprentissage, nous examinons les deux distributions présentées sur le graphique ci-dessous.

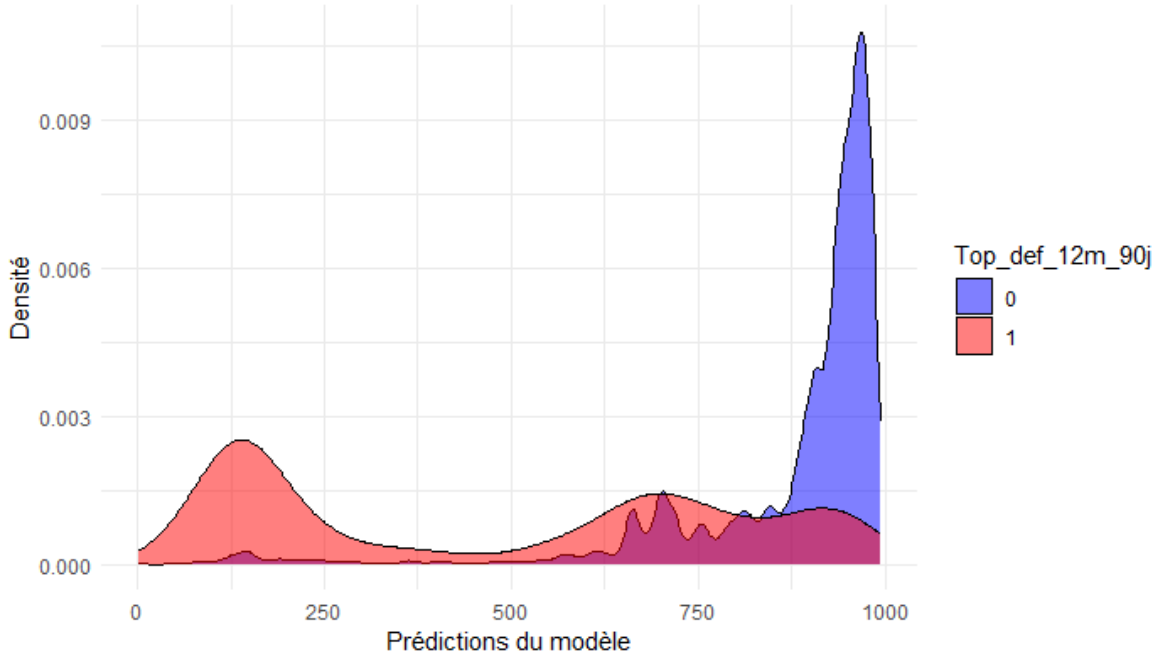

FIGURE 4 – Distribution des scores basés sur les probabilités

Nous observons de nouveau une zone de chevauchement entre les deux distributions. Pour surmonter cette difficulté, nous entreprenons une analyse de la population pour identifier une ou plusieurs variables permettant de mieux discriminer les deux populations. Nous effectuons cette analyse en nous appuyant sur les résultats obtenus à partir de la grille de score qui utilise les coefficients du modèle.

5.2 Analyse des distributions conditionnelles

Nous avons observé qu'au sein de la population dont le score basé sur les coefficients du modèle prédit se situe approximativement entre 500 et 700, le modèle présente une performance moyenne. En effet, les populations de défaut et de non-défaut de paiement se chevauchent, chacune présentant un pic dans la prédiction des scores, comme on peut le voir sur l'image centrée sur ces scores ci-dessous.

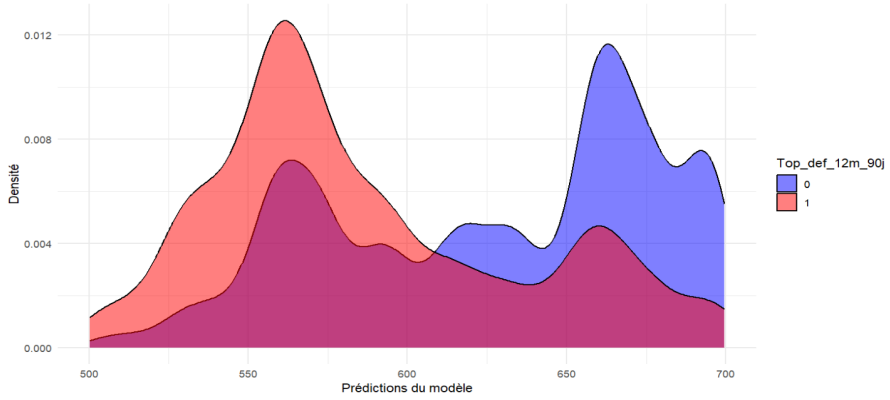

FIGURE 5 – Zoom sur la zone à différencier sur la distribution des scores basés sur les coefficients

Afin de mieux différencier les bons et les mauvais clients nous nous focalisons sur cette zone où le score se situe entre 500 et 700. Nous avons donc constitué une sous-base regroupant tous les clients ayant un score compris entre 500 et 700, qu'ils aient rencontré un défaut de paiement ou non. Cette population a été extraite des bases d'entraînement et de test, que nous avons nommées respectivement "train_coverage" et "test_coverage". Elles représentent respectivement 29 528 et 12 595 individus, soit environ 45% des bases d'entraînement et de test initiales. Parmi ces deux nouvelles bases, 5 853 et 2 491 clients se trouvent en défaut de paiement dans les bases d'entraînement et de test respectivement, ce qui équivaut à environ 20% de chaque base et à 44% des clients en défaut de paiement des bases initiales. En conséquence, nous estimons que le modèle parvient à bien discriminer les cas des clients dans environ 55% des cas. L'objectif est ainsi de trouver des variables permettant de les discriminer de manière plus efficace. La démarche adoptée consiste à explorer d'autres variables que celles prises en compte jusqu'à présent, mais qui pourraient se révéler plus pertinentes pour différencier les deux populations dans l'intervalle de score [500, 700].

Les variables les plus discriminantes entre les deux populations sont les suivantes :

— *sum_D_201704_201610* : Somme des mouvements débiteurs (en valeur absolue) sur les 6 derniers mois.

— *sld_courant_6m_min* : Montant minimal du compte courant durant les 6 derniers mois précédant la date d'arrêté.

— *sld_epargne_6m_min* : Montant minimal de l'épargne durant les 6 derniers mois précédant la date d'arrêté.

— *max_debit* : Montant maximal du flux débiteur.

— *max_credit* : Montant maximal du flux créditeur.

Concernant les variables continues telles que *sld_courant_6m_min* et *sld_epargne_6m_min*, nous les avons préalablement discrétisées en classes en fonction du nombre de classes qui maximisent le V_{mod} représentant le coefficient de Cramér modifié que nous avons précédemment défini. Par exemple, pour *sld_courant_6m_min*, nous obtenons 2 classes avec un V_{mod} de 0.213, et pour *sld_epargne_6m_min*, nous obtenons 3 classes avec un V_{mod} de 0.0392.

Modalité	**% observation**	**% défaut**
$[-3.76e+05, 44.8]$	50.0%	31.8%
$(44.8, 4.99e+06]$	50.0%	7.8%

TABLE 20 – Variable sld_courant_6m_min

Modalité	**% observation**	**% défaut**
$[-684, 0]$	77.2%	20.1%
$(0, 31.1]$	8.5%	26.0%
$(31.1, 7.26e+05]$	14.5%	10.4%

TABLE 21 – Variable sld_epargne_6m_min

Pour *sum_ D_ 201704_ 201610*, nous avons observé que parmi la population des clients présentant un défaut de paiement dans la base *train_ coverage*, 5 067 sur 5 853 (87%) ont une somme des débits inférieure ou égale à 0.

Modalité	**% observation**	**% défaut**
1	83.3%	20.6%
0	16.7%	16.1%

TABLE 22 – Variable sum_D_201704_201610

Pour *max_ debit*, nous avons remarqué que le premier quantile et la médiane sont tous deux égaux à 0, et que le troisième quartile est une valeur de 600. Nous avons donc divisé cette variable en deux modalités, où l'une prend la valeur 1 si le montant est inférieur à 600 et l'autre 0, ce qui a permis de maximiser l'écart entre les moyennes des défauts de ces deux modalités.

Modalité	**% observation**	**% défaut**
1	59.7%	21.9%
0	40.3%	16.7%

TABLE 23 – Variable max_debit

Enfin, pour *max_ credit*, nous avons observé que le premier quartile et la médiane sont également égaux à 0, tandis que le troisième quartile est de 280. Nous avons donc divisé cette variable en deux modalités, où les valeurs égales à 0 prennent la modalité 1 et les autres valeurs la modalité 0, maximisant ainsi l'écart des moyennes de défaut entre les modalités.

Modalité	**% observation**	**% défaut**
1	75.2%	21.5%
0	24.8%	14.6%

TABLE 24 – Variable max_credit

Après avoir calculé le V de Cramer entre les variables identifiées, nous avons remarqué une forte corrélation (0.56) entre les variables *max_ debit* et *max_ credit*. Par conséquent, nous ne pouvons pas les inclure toutes les deux dans le modèle. De plus, nous avons observé que la variable *max_ credit* est la plus corrélée avec la variable de défaut *top_ def_ 12m_ 90j*, avec une corrélation d'environ 0.08.

De plus, *sum_ D_ 201704_ 201610* et *max_ debit* sont également fortement corrélées (0.48). Cependant, étant donné que nous avons retiré la variable *max_ debit* et que la variable *sum_ D_ 201704_ 201610* n'est pas corrélée avec *max_ credit* (0.31), que nous conservons, nous décidons de la conserver également.

Nous avons par la suite défini les modalités de référence, suivant le même processus que précédemment, où la modalité de référence est celle présentant le plus haut risque (c'est-à-dire celle avec le taux de défaut le plus élevé). Ensuite, nous avons de nouveau implémenté un modèle logit sur cette

population de client restreinte, qui avait initialement un score compris entre 500 et 700, en utilisant uniquement ces variables. La formule du modèle est la suivante :

```
model <- glm(top_def_12m_90j ~ sld_epargne_6m_min + sld_courant_6m_min
    + max_credit + sum_D_201704_201610, data = train_coverage, family
   = "binomial")
```

Ensuite, nous avons recalculé une grille de score basé sur les coefficient du modèle pour cette sous population en suivant le même processus que précdemment. Ci-dessous, nous avons également représenté les distributions basées sur ce calcul de score.

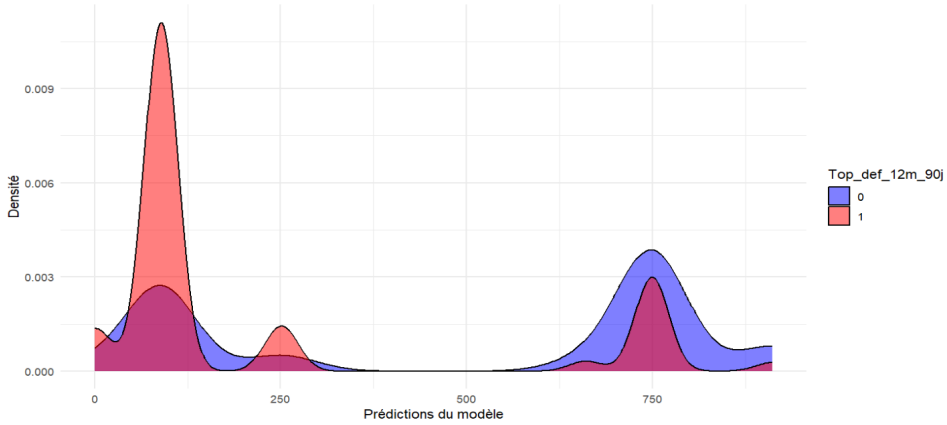

FIGURE 6 – Distribution des nouveaux scores basés sur les coefficients

Sur le graphique, on observe une certaine différenciation entre les deux populations, mais la présence de deux pics reste problématique pour la prise de décision. En effet, nous constatons un pic bleu autour de 50 correspondant aux clients qui n'ont jamais présenté de défaut de paiement (c'est à dire $top\_def\_12m\_90j = 0$), alors que ces clients devraient normalement avoir un score élevé pour être classés comme de bons clients. Cependant, la présence de ce pic pourrait être justifiée par le fait qu'il est possible que cette population soit extrêmement à risque mais n'ait pas encore connu de défaut jusqu'à présent. Par conséquent, cette population nécessite une surveillance particulièrement attentive. Le deuxième pic délicat se trouve autour de 750 et concerne la population ayant déjà fait défaut de paiement (c'est-à-dire $top\_def\_12m\_90j = 1$, ce qui se traduit par un pic rouge). Ce pic est particulièrement préoccupant car il attribue un score élevé à des clients ayant déjà fait un défaut de paiement, les classant ainsi comme de bons clients. Notre objectif était de minimiser ce pic, et bien que nous ayons réussi à le réduire, une minorité persistante demeure. Dans la population ayant présenté un défaut et obtenant un score élevé, compris entre 650 et 800, il y a 1 073 personnes dans la base d'entraînement et 471 dans la base de test. Cela représente 1.6% de la population initiale dans les bases d'entraînement et de test qui reste incorrectement prédite. Il serait pertinent d'organiser des rendez-vous avec les individus obtenant un deuxième score dans cette plage afin d'obtenir des informations supplémentaires sur ces clients ou de procéder à une vérification plus approfondie de leur compte.

5.3 Optimisation de l'utilisation du score de risque

En se basant sur les résultats du modèle stepwise et de l'analyse des distributions conditionnelles des scores, nous pouvons segmenter les scores en classes de risque pour mieux cibler les actions de prévention des défauts. En utilisant les coefficients de chaque variable obtenu dans le modèle stepwise, nous pouvons déterminer l'impact de chaque caractéristique sur le risque de défaut. Par exemple, les clients avec des scores plus élevés peuvent être ceux qui ont une situation familiale stable, peu de crédits ou une bonne couverture de liquidité. En revanche, ceux avec des scores plus bas peuvent avoir des caractéristiques telles qu'un historique de paiements en défaut, un nombre élevé de crédits ou une faible couverture de liquidité. En segmentant les scores en fonction de ces caractéristiques, nous pouvons adapter les stratégies de gestion du risque pour chaque groupe de clients, en mettant l'accent sur la surveillance intensive, les plans de remboursement ajustés ou les conseils financiers personnalisés, selon le niveau de risque estimé pour chaque client.

- Très faible risque (score de 800-1000) : Pour les clients de cette classe, caractérisés par des coefficients négatifs élevés dans des variables telles que la situation familiale stable, un faible endettement et une bonne couverture de liquidité, une surveillance minimale peut suffire, avec des contrôles réguliers pour détecter tout changement de comportement financier.
- Faible risque (score de 600-799) : Les clients de cette classe, avec des coefficients moins négatifs mais toujours significatifs dans ces mêmes variables, peuvent nécessiter une surveillance plus étroite. Des offres de produits ou services adaptés peuvent être proposées pour maintenir leur fidélité et les encourager à maintenir une santé financière stable.
- Risque modéré (score de 400-599) : Une surveillance plus attentive est recommandée pour les clients de cette classe, avec des coefficients moins négatifs mais significatifs dans ces mêmes variables. Des conseils financiers personnalisés peuvent être proposés pour les aider à gérer leurs dettes et à améliorer leur situation financière.
- Risque élevé (score de 200-399) : Pour cette classe, caractérisée par des coefficients moins négatifs mais toujours significatifs dans les variables de risque, des mesures préventives plus strictes sont nécessaires. Cela peut inclure des limites de crédit plus basses, des plans de remboursement ajustés et un suivi actif pour détecter les signes précoces de détérioration financière.
- Très haut risque (score de 0-199) : Les clients de cette classe, avec des coefficients positifs ou proches de zéro dans les variables de risque, exigent une intervention immédiate et proactive pour éviter les défauts. Cela peut impliquer des mesures telles que la réduction drastique des limites de crédit, des plans de remboursement stricts et un soutien financier intensif, voire des solutions de consolidation de dettes ou de restructuration financière.

En appliquant ces stratégies adaptées à chaque classe de score, nous visons à minimiser les défauts clients tout en préservant une relation client saine et durable.

5.4 Limites du score

Dans le cadre de notre analyse, il est important de prendre en considération certaines limites qui sont associées au score. Tout d'abord, le découpage des variables peut influencer la précision du modèle, car une mauvaise sélection des modalités peut entraîner une perte d'information significative. De plus, malgré nos efforts pour éliminer les problèmes de multicolinéarité entre les variables explicatives utilisées dans le modèle, il est possible qu'il persiste quelques résidus, ce qui pourrait compromettre la fiabilité des prédictions. Le ré-échantillonnage des données que nous avons effectué peut également introduire un biais dans les résultats, en particulier si les échantillons ne sont pas représentatifs de la population totale. De plus, la fiabilité des données peut être remise en question en raison de l'absence de données externes pour valider les résultats du modèle. En effet, le manque d'informations sur les clients qui possèdent plusieurs comptes bancaires peut limiter la précision du score, car certaines données pertinentes pourraient être stockées dans d'autres institutions financières.Etant donné que notre analyse est basée sur des données d'une banque patrimoniale, il est possible que des informations cruciales se trouvent dans d'autres banques du client, ce qui soulève la problématique du manque d'informations lorsque l'individu a des comptes multiples. En résumé, bien que le score puisse être un outil utile pour évaluer le risque, il est essentiel de prendre en compte ces limitations lors de l'interprétation des résultats et de les compléter par une analyse approfondie.

6 Challenger le modèle

Dans cette section, nous explorons d'autres méthodes statistiques pour concevoir des scores et identifier les individus à risque. Nous allons recourir à des techniques d'apprentissage automatisé reconnues telles que le Random Forest et les arbres de décision. Nous utiliserons les mêmes métriques d'évaluation des résultats afin de pouvoir comparer ces résultats à ce qui a été réalisé précédemment.

6.1 Random Forest

Random Forest est un algorithme d'apprentissage supervisé utilisé pour la classification. Il construit un ensemble de plusieurs arbres de décision, où chaque arbre est formé sur un sous-ensemble aléatoire des données et des caractéristiques. Ensuite, il combine les prédictions de chaque arbre pour fournir une prédiction finale. Ainsi les prédictions sont plus robustes. L'objectif est d'utiliser les mêmes variables que celles sélectionnées dans le modèle Stepwise précédemment présenté, mais cette fois-ci avec l'algorithme de Random Forest. Pour cela, nous avons utilisé la commande suivante :

```
rf <- randomForest(stepwise_model$formula , data=na.omit(train_data))
```

6.1.1 Evaluation du modèle de Random Forest

Nous obtenons un score AUC de 0.91, un indice de Gini de 0.81 et un niveau de précision de 0.81, ce qui représente une amélioration par rapport au modèle stepwise. Ces performances sont très encourageantes et prometteuses.

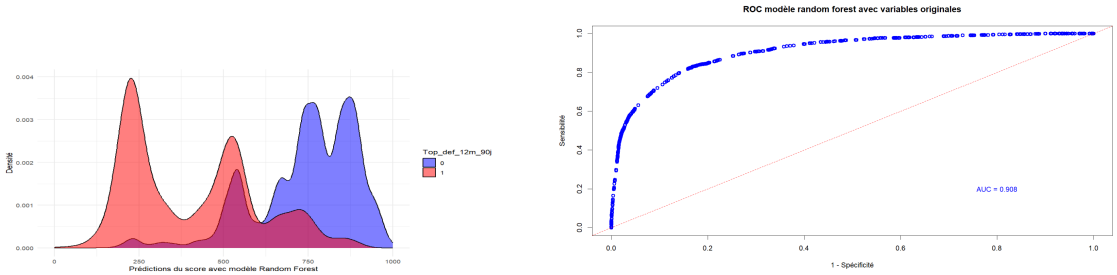

FIGURE 7 – Métrique du modèle random forest

La distribution des scores prédits par le modèle de Random Forest est satisfaisante, mais on remarque toujours une zone de chevauchement assez importante. En examinant la courbe ROC, nous constatons que le taux de faux négatifs semble être maîtrisé. Cependant, nous n'avons pas exploré davantage cette méthode en raison de la faible transparence et de l'interprétabilité limitée de ce modèle. En effet, le modèle de Random Forest est une agrégation de 500 modèles d'arbres de décision différents, chacun présentant une interprétation des données différente. C'est pourquoi nous avons décidé de mettre celui de côté et de s'intéresser à un modèle d'arbre de décision.

6.2 Arbre de décision

Pour l'implémentation des arbres de décisions, nous avons également utilisé les mêmes variables que celles présentes dans le modèle stepwise auxquelles nous avons ajouté les variables qui ont servies à la discrimination. Cependant, nous avons choisi de ne pas discrétiser les variables comme nous l'avons fait pour le modèle stepwise. Nous avons également choisi de ne pas regrouper les modalités des variables qualitatives. Les variables croisées sont remplacées par les deux variables primaires. Ces décisions reposent sur deux raisons principales. Premièrement, les modèles de type arbre de décision traitent efficacement les situations de colinéarité entre les variables. Un arbre peut également exclure une variable du modèle s'il estime qu'elle n'est pas pertinente. Ainsi, doubler l'information en incluant deux variables très similaires n'a pas d'impact négatif sur ce type de modèle. Deuxièmement, par sa nature, cet algorithme va créer des règles de décision pour faire ses prédictions sur les notes. Par conséquent, il générera lui-même les différentes classes au sein de chaque variable. Les classes créées indirectement à travers ces règles de décision seront beaucoup plus précises que celles que nous aurions créées par discrétisation.

Nous avons effectué une recherche des paramètres optimaux afin de perfectionner notre modèle d'arbre de décision et d'obtenir de meilleures perfromances.

```
param_grid_multi <- makeParamSet(
 makeDiscreteParam("maxdepth", values=c(1,2,5,15,25,30)),
 makeNumericParam("cp", lower = 0.001, upper = 0.01),
 makeDiscreteParam("minsplit", values=c(10,50,100,150,200)))
```

Les paramètres retenus pour le modèle final sont : 0.001 pour le paramètre de complexité (cp), 15 pour le paramètre de profondeur maximale des arbres et 10 pour le nombre minimal d'individus à chaque terminal. Le modèle estimé est le suivant :

```
dt_tune <- rpart(top_def_12m_90j ~ sit_familiale + top_credit + top_
    compte_joint + topDecNonAut + topCred + liquidity_coverage +
    montant_impaye + nb_credit + sld_liqu_6m_mean + age + CSP + mtn_
    investm + topGestionPTF +  sum_D_201704_201610 + sld_epargne_6m_min
     + sld_courant_6m_min + max_credit , data=na.omit(save_train),
    control = rpart.control(cp = 0.001,minsplit=10,maxdepth = 15))
```

6.2.1 Evaluation du modèle d'arbre de décision

Nous obtenons un modèle très satisfaisant avec un score AUC de 0.94, un indice de Gini de 0.87 et une précision de 0.79. La répartition des scores prédits est très satisfaisante, malgré quelques individus à risque avec un score élevé. Le tableau ci-dessous compare les performances des différents modèles, indiquant que l'arbre de décision semble être celui offrant les meilleures performances.

Modèle	AUC	Gini	Précision
Stepwise	0.88	0.75	0.76
Random Forest 1	0.91	0.81	0.81
Arbre de Décision	0.94	0.87	0.79

TABLE 25 – Comparaison des performances des modèles

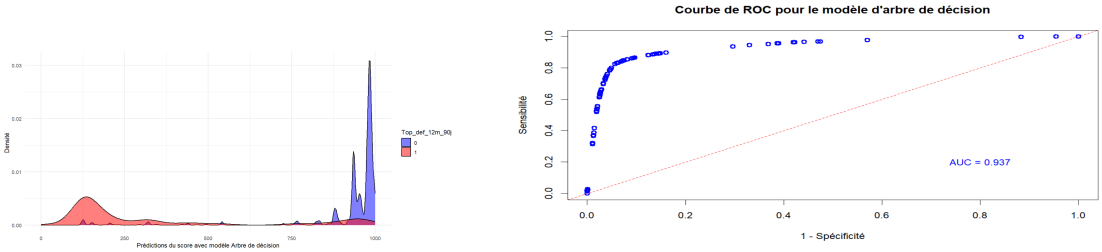

FIGURE 8 – Métrique du modèle d'arbre de décision

Par la suite, en utilisant la fonction `rpart.rules`, nous pouvons accéder aux différentes règles de décision prises par l'algorithme. Nous nous intéressons aux décisions prises par l'arbre pour les variables *mtn_investm* , *age* et *sld_liqu_6m_mean*.

— Pour la variable *mtn_investm* nous avons les seuils suivants : 9.2, 56, 100, 4300, 14 000 et 290 000 contre deux seuils à 0 et 33 000 dans la régression logistique.

— Pour la variable *age* nous avons les seuils suivants : 28, 29, 44, 48 et 72, tandis que lors de la discrétisation manuelle des variables, nous avions seulement un seuil à 58 ans.

— Pour la variable *sld_liqu_6m_mean* nous avons les seuils suivants : -1282.80, -1163.07, -313.41, -1.39, 12.47, 20.21, 43.51, 231.30, 373.55, 460.81, 609.90, 11 192.81 et 12 912.86. L'arbre de décision semble donc offrir une meilleure précision que la discrétisation manuelle des variables, où nous n'avions que deux seuils pour cette variable : 639 et 7 860.

On remarque donc que les découpages des variables obtenues avec le modèle d'arbre de décision est bien plus précis que ceux que nous avions définis à la main pour le modèle de régression logistique. Cette observation explique en partie les meilleurs résultats obtenus avec ce modèle.

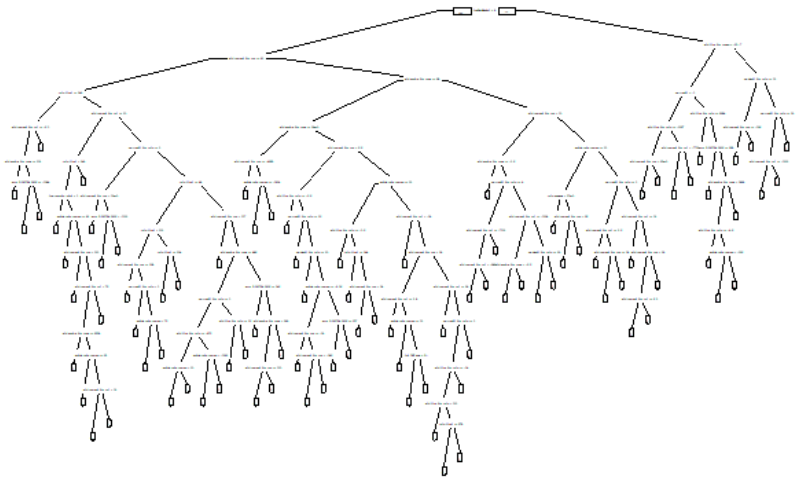

FIGURE 9 – Représentation de l'arbre de décision

Lorsque l'arbre de décision se divise, une même variable peut apparaître dans les deux branches résultantes. Ces branches représentent deux sous-groupes distincts, chacun ayant des règles associées à cette variable qui peuvent différer. Cependant, il est important de noter que des seuils très proches peuvent être observés pour une même variable (par exemple : 29 et 28 pour l'âge). Il est possible que ces seuils correspondent en fait à une même probabilité de risque.

6.3 Avantages et inconvénients

Le modèle Random Forest, tout comme l'arbre de décision, présente plusieurs avantages par rapport au modèle de régression logistique. En termes de flexibilité, ces modèles peuvent capturer des relations

non linéaires entre les variables, ce qui les rend potentiellement plus adaptés à des situations où les relations sont complexes ou mal comprises. De plus, ils peuvent gérer efficacement les variables catégorielles sans nécessiter de transformation supplémentaire, offrant ainsi une approche plus simple pour inclure des variables de différents types dans le modèle.

Cependant, les modèles d'arbres de décision et de Random Forest présentent des inconvénients. Ils sont sujets au surajustement, ce qui signifie qu'ils peuvent trop bien s'adapter aux données d'apprentissage spécifiques et ne pas généraliser correctement aux nouvelles données. Cette suradaptation peut être particulièrement préoccupante lorsque les modèles sont complexes ou lorsque les ensembles de données sont petits. En ce qui concerne le modèle de Random Forest, bien qu'il puisse fournir de bonnes performances prédictives, son interprétation peut être difficile car il repose sur un agrégat de nombreux arbres de décision individuels, ce qui peut rendre la compréhension des prédictions globales plus complexe que dans le cas de modèles linéaires comme la régression logistique. En conséquence, dans les contextes où la transparence et l'interprétabilité du modèle sont importantes, la régression logistique est généralement préférable à Random Forest.

7 Conclusion

Notre analyse souligne l'importance de la détection précoce des clients présentant un risque élevé de défaut de paiement dans le secteur bancaire. L'utilisation de modèles économétriques et de machine learning permet d'identifier les caractéristiques des clients ayant un impact important sur la probabilité de faire défaut afin de mettre en place des plans adaptés pour gérer les risques, en utilisant une grille de score, afin d'agir rapidement et de manière préventive.

L'analyse des résultats de notre modèle de régression logistique révèle des variables significatives telles que la situation familiale, les prêts et crédits détenus, la possession de comptes joints ou de découverts non autorisés, le niveau de liquidité, les impayés, la catégorie socio-professionnelle (CSP), l'âge, ainsi que la gestion de portefeuille et les investissements du client. Etant donné que toutes ces variables sont significatives, elles jouent un rôle déterminant dans la prédiction du risque de défaut de paiement des clients.

Nous avons par la suite mis en place une grille de score basée sur ces variables afin de permettre aux banques de classer les clients en fonction de leur niveau de risque, facilitant ainsi la prise de décision en matière d'octroi de crédit et de tarification.

En poursuivant notre exploration, nous nous sommes concentrés sur les scores compris entre 500 et 700, où un chevauchement entre les clients ayant déjà fait défaut et ceux n'en ayant jamais fait est observé, dans le but d'améliorer la distinction entre les bons et les mauvais clients. Malgré nos efforts pour réduire les chevauchements entre ces populations, nous avons remarqué la présence persistante de clients à haut risque, notamment ceux ayant déjà rencontré des défauts de paiement et obtenant des scores élevés. Pour ces clients, des mesures supplémentaires, telles que des rendez-vous ou des vérifications de compte approfondies, pourraient être nécessaires pour une évaluation plus précise de

leur risque financier.

En utilisant les mêmes variables sélectionnées par le modèle stepwise, nous avons entraîné un modèle d'arbre de décision et obtenau des performances remarquables avec un score AUC de 0.91, un indice de Gini de 0.83 et une précision de 0.96. Malgré quelques individus à risque avec des scores élevés, l'ensemble des résultats souligne une approche très prometteuse. La représentation graphique du modèle suggère qu'il sacrifie en partie son interprétabilité pour obtenir des performances supérieures. Il est judicieux pour le banquier de confronter ces deux modèles et de sélectionner celui qui correspond le mieux à ses objectifs et aux besoins de la banque. Cette approche favorise un équilibre entre la facilité d'interprétation et les capacités prédictives, en fonction des besoins particuliers.

Cependant, il convient de noter que notre analyse présente certaines limites, notamment en termes de généralisation des résultats. En effet, les données utilisées proviennent d'une banque patrimoniale où les clients sont souvent des particuliers aisés ayant des besoins de gestion de patrimoine plus complexes. Par conséquent, les conclusions tirées de ces données ne peuvent pas être directement étendues à d'autres types de banques qui ciblent des segments de clientèle différents ou qui proposent des services bancaires différents. Ainsi, il est recommandé de poursuivre la recherche dans ce domaine, en explorant des méthodologies plus avancées et en intégrant des données supplémentaires pour améliorer la précision des prédictions.

En définitive, notre étude met en lumière l'importance croissante de l'analyse prédictive dans le secteur bancaire, offrant aux institutions les moyens d'anticiper et de gérer efficacement les risques financiers, tout en garantissant une prise de décision éclairée.

Annexes

	Variable dépendante :
	top_def_12m_90j
liquidity_coverage[-6.95e+05,0]	−0.5509***
	(0.1082)
sit_familialeVeuf(ve), NR ou pacs	−0.1795*
	(0.0907)
top_credit1	−0.4681***
	(0.0531)
top_compte_joint	−0.1861***
	(0.0270)
top_pret_infine0	0.0119
	(0.2989)
topDecNonAut0	−2.6101***
	(0.0317)
topCred0	−0.6950***
	(0.0526)
liquidity_coverage[-6.95e+05,0]	−0.5509***
	(0.1082)
dummy_impaye0	−2.9491***
	(0.2741)
nb_credit(1,958]	−0.5162***
	(0.0310)
sld_liqu_6m_mean(639,7.86e+03]	−1.3772***
	(0.0299)
sld_liqu_6m_mean(7.86e+03,4.99e+06]	−1.9671***
	(0.0398)
int_CSP_ageDivers_[-1,58]	−0.4070***
	(0.0526)
int_CSP_ageRetraites_[-1,58]	−0.9737**
	(0.3282)
int_CSP_ageTravailleurs indépendants_(58,80]	−0.0668
	(0.0719)
int_CSP_ageDivers_(58,80]	−0.4883***
	(0.0555)
int_CSP_ageRetraites_(58,80]	−0.6728***
	(0.0615)
int_investm_topGestionPTF[-6.29e+03,0]_0	−0.3776***
	(0.0663)
int_investm_topGestionPTF(0,3.3e+04]_1	−0.5476***
	(0.0691)
int_investm_topGestionPTF(3.3e+04,2.05e+07]_1	−1.1547***
	(0.0718)
Observations	66 077
Deviance nulle	66 132
Deviance résiduelle	42 363
AIC	42 403
Note :	*** p<0.001 ; ** p<0.01 ; * p<0.05

Table 26 – Résultats du modèle de régression logistique

	top_def_12m_90j	sit_familiale	top_credit	top_compte_joint	top_pret_infine	topDecNonAut	topCred	liquidity_coverage	dummy_impaye	nb_credit	sld_liqu_6m_mean	int_CSP_age
top_def_12m_90j	1.000	0.023	0.020	0.050	0.005	0.556	0.099	0.137	0.116	0.207	0.450	0.125
sit_familiale	0.023	1.000	0.031	0.116	0.002	0.018	0.010	0.002	0.010	0.009	0.021	0.170
top_credit	0.020	0.031	1.000	0.230	0.151	0.015	0.326	0.332	0.170	0.149	0.031	0.164
top_compte_joint	0.050	0.116	0.230	1.000	0.034	0.025	0.110	0.084	0.033	0.093	0.066	0.097
top_pret_infine	0.005	0.002	0.151	0.034	1.000	0.018	0.069	0.119	0.003	0.021	0.011	0.015
topDecNonAut	0.556	0.018	0.015	0.025	0.018	1.000	0.079	0.200	0.066	0.140	0.380	0.095
topCred	0.099	0.010	0.326	0.110	0.069	0.079	1.000	0.256	0.240	0.177	0.067	0.063
liquidity_coverage	0.137	0.002	0.332	0.084	0.119	0.200	0.256	1.000	0.217	0.039	0.118	0.066
dummy_impaye	0.116	0.010	0.170	0.033	0.003	0.066	0.240	0.217	1.000	0.010	0.072	0.043
nb_credit	0.207	0.009	0.149	0.093	0.021	0.140	0.177	0.039	0.010	1.000	0.386	0.100
sld_liqu_6m_mean	0.450	0.021	0.031	0.066	0.011	0.380	0.067	0.118	0.072	0.386	1.000	0.078
int_CSP_age	0.125	0.170	0.164	0.097	0.015	0.095	0.063	0.066	0.043	0.100	0.078	1.000

TABLE 27 – Matrice des V de Cramer

Variable	Modalité	% observation train	% observation test	IS
dummy_impaye	dummy_impaye.1	0%	0%	0.00
dummy_impaye	dummy_impaye.0	100%	100%	0.00
int_CSP_age	int_CSP_age.Travailleurs indépendants_[-1,58]	5%	5%	0.00
int_CSP_age	int_CSP_age.Divers_[-1,58]	47%	46%	0.00
int_CSP_age	int_CSP_age.Retraites_[-1,58]	0%	0%	0.01
int_CSP_age	int_CSP_age.Travailleurs indépendants_(58,80]	5%	5%	0.00
int_CSP_age	int_CSP_age.Divers_(58,80]	27%	27%	0.00
int_CSP_age	int_CSP_age.Retraites_(58,80]	17%	17%	0.00
int_investm_topGestionPTF	int_investm_topGestionPTF.[-6.29e+03,0]_1	3%	3%	0.00
int_investm_topGestionPTF	int_investm_topGestionPTF.[-6.29e+03,0]_0	51%	51%	0.00
int_investm_topGestionPTF	int_investm_topGestionPTF.(0,3.3e+04]_1	21%	20%	0.00
int_investm_topGestionPTF	int_investm_topGestionPTF.(3.3e+04,2.05e+07]_1	25%	25%	0.00
liquidity_coverage	liquidity_coverage.(0,1.14e+05]	1%	1%	0.00
liquidity_coverage	liquidity_coverage.[-6.95e+05,0]	99%	99%	0.00
nb_credit	nb_credit.[0,1]	59%	60%	0.00
nb_credit	nb_credit.(1,958]	41%	40%	0.01
sit_familiale	sit_familiale.Séparé(e), divorcé(e), célibataire ou marié(e)	97%	98%	0.00
sit_familiale	sit_familiale.Veuf(ve), NR ou pacs	3%	2%	0.00
sld_liqu_6m_mean	sld_liqu_6m_mean.[-5.59e+04,639]	33%	33%	0.00
sld_liqu_6m_mean	sld_liqu_6m_mean.(639,7.86e+03]	33%	33%	0.00
sld_liqu_6m_mean	sld_liqu_6m_mean.(7.86e+03,4.99e+06]	33%	34%	0.00
top_credit	top_credit.0	88%	87%	0.00
top_credit	top_credit.1	12%	13%	0.03
topCred	topCred.1	6%	6%	0.00
topCred	topCred.0	94%	94%	0.00
topDecNonAut	topDecNonAut.1	13%	13%	0.00
topDecNonAut	topDecNonAut.0	87%	87%	0.00

TABLE 28 – Stabilité des populations

Variable	Modalité	% défaut train	% défaut test	IS
dummy_impaye	0	20%	20%	0.00
dummy_impaye	1	94%	93%	0.00
int_CSP_age	Divers_(58,80]	18%	18%	0.00
int_CSP_age	Divers_[-1,58]	22%	21%	0.00
int_CSP_age	Retraites_(58,80]	12%	13%	0.00
int_CSP_age	Retraites_[-1,58]	11%	12%	0.00
int_CSP_age	Travailleurs indépendants_(58,80]	29%	32%	0.00
int_CSP_age	Travailleurs indépendants_[-1,58]	33%	33%	0.00
int_investm_topGestionPTF	(0,3.3e+04]_1	27%	28%	0.00
int_investm_topGestionPTF	(3.3e+04,2.05e+07]_1	10%	10%	0.00
int_investm_topGestionPTF	[-6.29e+03,0]_0	0.21	0.21	0.00
int_investm_topGestionPTF	[-6.29e+03,0]_1	31%	31%	0.00
liquidity_coverage	(0,1.14e+05]	65%	61%	0.00
liquidity_coverage	[-6.95e+05,0]	19%	19%	0.00
nb_credit	(1,958]	10%	10%	0.00
nb_credit	[0,1]	27%	27%	0.00
sit_familiale	Séparé(e), divorcé(e), célibataire ou marié(e)	20%	20%	0.00
sit_familiale	Veuf(ve), NR ou pacs	14%	16%	0.00
top_compte_joint	0	22%	22%	0.00
top_compte_joint	1	18%	18%	0.00
top_credit	0	20%	20%	0.00
top_credit	1	18%	17%	0.00
topCred	0	19%	19%	0.00
topCred	1	35%	37%	0.00
topDecNonAut	0	11%	11%	0.00
topDecNonAut	1	76%	76%	0.00

TABLE 29 – Stabilité des défauts

	max_credit	max_debit	sum_D_201704_201610	sld_courant_6m_min	sld_epargne_6m_min	top_def_12m_90j
max_credit	1.00	0.56	0.31	0.07	0.11	0.08
max_debit	0.56	1.00	0.48	0.02	0.16	0.06
sum_D_201704_201610	0.31	0.48	1.00	0.00	0.10	0.04
sld_courant_6m_min	0.07	0.02	0.00	1.00	0.05	0.30
sld_epargne_6m_min	0.11	0.16	0.10	0.05	1.00	0.07
top_def_12m_90j	0.08	0.06	0.04	0.30	0.07	1.00

TABLE 30 – V de Cramer pour les variables du modèle de discimination

	Variable dépendante :
	top_def_12m_90j
sld_epargne_6m_min[-684,0]	-0.3081^{***}
sld_epargne_6m_min(31.1,7.26e+05]	-0.6824^{***}
sld_courant_6m_min(44.8,4.99e+06]	-1.7916^{***}
max_credit0	-0.6774^{***}
sum_D_201704_2016100	−0.0813
Note :	*** p<0.001 ; ** p<0.01 ; * p<0.05

TABLE 31 – Résultats du modèle de la régression logistique pour la discrimination des populations