Projet de fin d'étude

Segmentation et dépréciation des actifs financiers

Gestion Des Risques et Ingénierie Financière

Etudiants
Wassim Amara
Malcom Smeralda
Sidi Traore

Encadrants
Maxime Debbah
Adnane Benhajja
Adam Ouahab

Responsable
Samuel Danthine

Année : 2020-2021

Remerciements

La réalisation de ce projet a été possible grâce à la direction de l'ENSAI à qui nous voulons témoigner toute notre gratitude.

Nous souhaitons ensuite adresser toute notre reconnaissance à nos tuteurs de stage Mr. **Maxime DEBBAH**, Mr. **Adnane BENHAJJA** et Mr. **Adam OUAHAB** pour leur patience, leur disponibilité et surtout leur conseils très instructifs qui ont contribué au bon déroulement de ce projet. Les séances de suivi et leur point de vu d'expert nous a énormément aidé dans les démarches que nous avons pu entreprendre lors de ce projet.

Abstract

In order to anticipate and minimize losses related to financial assets, the International Accounting Standards Board has introduced IFRS9 in January 2018 to replace former accounting standard IAS39. This new standard is implemented through three key phases: **Classification, impairment and hedge accounting**. Impairment mainly consists of introducing criteria in order to quantify credit risk level. According to these criteria, financial institutions classify their contracts into three different stages: S1 for safe contracts, S2 for contracts with significant risk deterioration of probability of default and S3 for conrtacts already at default. Within banks' modeling departments, several statistical methods are implemented in-house in order to properly segment contracts according to their risk level. The main goal of our study will be to set up a regression model that allows to accomplish the second phase of the implementation of the new standard.

Table des matières

Introduction **1**

1 Présentation des données et problématique **2**

2 De la norme IAS 39 à la norme IFRS 9 **3**
2.1 Norme IAS 39 . . . 3
2.2 Norme IFRS 9 . . . 4

3 Risque de crédit **5**
3.1 Exposure At Default . . . 5
3.2 Loss Given Default . . . 5
3.3 Probability of Default . . . 5
3.4 Expected Credit Loss . . . 5
3.5 Scoring de crédit . . . 6

4 Statistiques descriptives et nettoyage des données **7**
4.1 Description de la population d'étude . . . 7
4.2 Nettoyage des données . . . 8

5 Modélisation de la classification en S1 ou en S2 **9**
5.1 Le modèle logistique . . . 9
5.1.1 Spécification du modèle . . . 9
5.1.2 L'odds ratio . . . 9
5.1.3 Régression logistique . . . 10
5.1.4 Attribution des scores . . . 10
5.2 Validation du modèle par bootstrap . . . 12
5.3 Détermination du seuil de distinction des contrats en S1 & S2 . . . 12
5.3.1 Approche graphique par densité conditionnelle du score . . . 13
5.3.2 Approche par maximisation du V de Cramer . . . 14
5.4 Analyse des résultats . . . 14
5.4.1 Matrice de confusion . . . 14
5.4.2 Distinctions des contrats BP des contrats CE . . . 15

Conclusion **17**

A Annexe **18**
A.1 Tests statistiques utilisés . . . 18

A.1.1 Le coefficient de Pearson 18
A.1.2 Test d'indépendance du χ^2 18
A.1.3 Test V de Cramer : liaison entre variables qualitatives 19
A.1.4 Test de Kruskal Wallis 19
A.2 Test de normalité 20

B Corrélations des variables **21**
B.1 V de Cramer des variables retenues 21
B.2 Matrice de corrélation des variables finales 22

C Mise en place du modèle de régression logistique **23**

D Dictionnaire des variables **24**

Introduction

Le secteur bancaire a toujours eu un rôle économique crucial. En effet, les banques ont accès à des fonds monétaires générés par des épargnants qui vont permettre de financer les projets de différents acteurs économiques (des particuliers, des entreprises ou des sociétés gouvernementales) en leur octroyant des crédits. L'ajout d'un taux d'intérêt associé à ces crédits va alors permettre aux banques de dégager un surplus monétaire. Cependant, lorsqu'un créancier accorde un crédit à une contrepartie, il court le risque que ce dernier ne le rembourse pas aux échéances fixées lors de la signature du contrat. C'est ce qu'on appelle le risque de crédit. La maîtrise de ce risque est au coeur du métier du banquier car il détermine la rentabilité des opérations effectuées. Si la banque sous-évalue ce risque, le montant prêté et les intérêts fixés ne permettront pas de tirer de bénéfices mais seront plutôt comptabilisés en pertes.

Le risque de crédit est conditionné par ce qu'on appelle trois facteurs de risque : l'Exposure At Default qui représente le montant auquel la banque sera exposée, la Loss Given Default qui est le pourcentage de la dette irrécupérable et la Probability of Default qui représente la probabilité qu'un débiteur ne puisse faire face à ses obligations de remboursement sans avoir recours à une autre banque. C'est pourquoi, une étude profonde du profil de la contre-partie sera effectuée afin d'évaluer s'il est profitable d'accorder ce crédit. Dans la suite de cette étude, nous nous focaliserons sur le facteur le plus étudié : la Probability of Default.

Suite à la crise financière, le domaine comptable a été touché comme pour beaucoup d'autres, par des changements réglementaires. En effet, différentes normes comptables ont été suspectées d'avoir contribué à la crise financière. C'est par exemple le cas de la norme IAS 39 intitulée "Instruments financiers : comptabilisation et évaluation". Cette norme a été remise en question en raison de sa complexité menant à une application incohérente avec les activités des entreprises, d'une trop grande latitude de classification des actifs financiers laissée aux banques, d'une reconnaissance tardive des pertes et d'une provision des pertes insuffisantes. C'est pour ces principales critiques que l'IASB[1] a publié en juillet 2014 la version finale de la norme IFRS 9 qui devra être appliquée par tous les organismes de gestion comptable à partir du 1^{er} janvier 2018.

IFRS 9 devrait d'abord permettre de simplifier l'évaluation et la comptabilisation des actifs financiers. De plus, le délai avant constatation d'une perte sur créance était trop long sous IAS 39. C'est pourquoi, une entreprise doit désormais évaluer et reconnaître en tout temps les pertes anticipées et publier régulièrement le montant que représentent ces pertes. Ceci lui permettra de rendre compte de l'évolution du risque de crédit des instruments financiers détenus dans son portefeuille. Pour finir, il s'agira pour les entreprises de comptabiliser la couverture qui permettra de compenser les éventuelles pertes futures associées aux divers instruments financiers.

[1]International Accounting Standard Boards est l'organisme international chargé de l'élaboration des normes comptables internationales IAS et IFRS

1 Présentation des données et problématique

Le Groupe BPCE est la 2^{nde} institution bancaire coopérative en France. Il comprend l'ensemble des entreprises qui composaient les deux groupes bancaires de la Banque Populaire et de la Caisse d'Epargne, ainsi que leurs filiales propres et communes. La BPCE agit alors pour l'intérêt de ses 36 millions de clients dans le monde, qu'il s'agisse de particuliers, de professionnels, d'entreprises ou d'investisseurs.
Afin de répondre aux différentes problématiques bancaires, il est devenu commun pour les banques d'utiliser la Data Science afin de satisfaire au mieux les besoins des clients et d'améliorer la gestion des risques. La BPCE a donc mis en place un processus permettant de collecter les données de ses clients. Dans l'objectif de répondre à une problématique précise. Ces données sont à caractère confidentiel et ne seront utilisées que pour les seuls besoins des travaux faisant l'objet de notre problématique.

Comme nous l'avons vus précédemment, IFRS 9 est une nouvelle norme entrée en vigueur le 1^er janvier 2018 remplaçant la précédente comptable IAS 39. L'application de la norme IFRS 9 se décompose en trois phases. La première phase vise à identifier, évaluer et classer les actifs selon leurs caractéristiques. La seconde phase consiste en la dépréciation des actifs. La troisième phase vient conclure le procédé en comptabilisant la couverture nécessaire pour les différents actifs détenus. On s'attardera dans le cadre de ce projet sur la phase de dépréciation afin d'observer les dégradations du risque de crédit des différents actifs. Au cours de cette phase, on cherche particulièrement à classer les actifs selon trois catégories. Les actifs sains, classés en "S1" sont ceux n'ayant pas subit de dégradation du risque de crédit, on provisionne donc les pertes attendues sur les douze prochains mois. Les actifs ayant subi une dégradation significative du risque de crédit sont classés en "S2" et sont provisionnés à hauteur des pertes attendues jusqu'à maturité. Les actifs classés en "S3" sont les actifs dont les pertes sont déjà avérées et ne seront pas étudiés dans la suite de ce projet.

Il s'agira dans le cadre de ce projet de trouver des critères permettant de faciliter la classification d'un actif en tant qu'actif sain (S1) ou en tant qu'actif risqué ayant subit une dégradation significative du crédit (S2). Pour ce faire, après avoir nettoyé la base de données, nous présenterons les quelques résultats théoriques nécessaires à la bonne compréhension des modèles statistiques que nous étudierons. Puis, nous appliquerons les méthodes de modélisation de la régression logistique. Il s'agira alors d'interpréter le modèle retenu puis d'en étudier les performances.

2 De la norme IAS 39 à la norme IFRS 9

Comme bien d'autres professions, la banque est fortement réglementée. La réglementation bancaire est destinée à soutenir la solidité et l'intégrité des établissements de crédits. En effet, les règles encadrant l'activité bancaire ont des conséquences majeures sur le fonctionnement et la gestion des établissements de crédit, c'est pourquoi, il est nécessaire de connaître comment a évolué la réglementation bancaire. Etant donné que notre sujet résulte de l'introduction d'une nouvelle norme comptable, nous avons jugé nécessaire d'effectuer un bref rappel historique de la transition entre la norme IAS 39 et la norme IFRS 9.

2.1 Norme IAS 39

En 2002, l'Union européenne a adopté une réglementation imposant aux sociétés européennes cotées d'appliquer les normes internationales d'information financière (les IFRS, et par conséquent la norme IAS 39) à compter de 2005 pour l'établissement de leurs états financiers consolidés. Il s'agit d'une évolution importante en matière de communication financière. La norme IAS 39 porte sur les instruments financiers et révolutionne donc la gestion des différents instruments financiers. En effet, pour les entreprises qui investissent, empruntent ou utilisent les instruments dérivés de façon significative, la norme soulève des questions plus vastes que toute autre norme antérieure dans tous les domaines de la comptabilité.

Pour les premiers utilisateurs, la norme modifie la façon dont les instruments financiers sont comptabilisés et implique des changements importants dans les systèmes, les procédures et la documentation. Certaines modifications de la norme IAS 39 affectent les préparateurs de comptes en IFRS, dans des domaines tels que la classification des actifs, les règles de décomptabilisation et d'autres domaines clés.

L'objectif principal de la norme IAS 39 est d'établir les principes de comptabilisation et d'évaluation des actifs financiers, des passifs financiers et de certains contrats d'achat ou de vente d'éléments non financiers. Cependant, la crise mondiale de 2007-2008 a souligné la fragilité de la norme IAS 39. En effet, cette norme est vite jugé non seulement trop complexe pour les établissements de crédit, mais également, elle n'est pas en adéquation avec la façon dont les institutions financières gèrent leurs instruments financiers.

2.2 Norme IFRS 9

Les normes IFRS (International Financial Reporting Standard) s'appliquent aux institutions financières cotées dans l'UE depuis le 1^er janvier 2005. IFRS 9 « Instruments Financiers » vient compléter et remplacer la norme IAS 39 « Instruments financiers : comptabilisation et évaluation », afin d'améliorer l'information financière.

Le principe de cette norme est d'employer une approche plus prospective pour la comptabilisation des pertes attendues sur des actifs financiers. Le projet a été lancé suite à la crise financière de 2008 et s'appuie sur une méthode de provisionnement sur les pertes attendues plutôt que les pertes subies. Auparavant, les établissements de crédit ne constatait le défaut que lorsque le client avait trois impayés ou plus de 90 jours de retard de paiement par exemple. Sous IFRS 9, on étudie l'évolution de la probabilité de défaut dans le temps afin de prévenir une éventuelle insolvabilité. Après avoir classé les actifs selon leurs différentes caractéristiques, on applique la phase de dépréciation que l'on peut résumer à l'aide du schéma suivant :

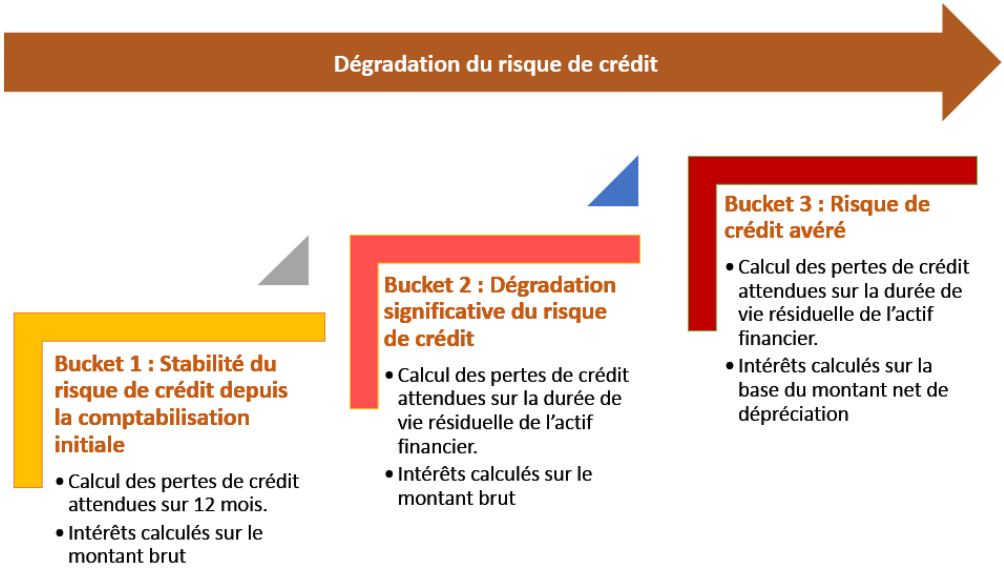

Figure 2.1: Provisionnement sous IFRS 9. Source "*L'essentiel de la norme IFRS 9*" - Conix

Ainsi, selon la dégradation du risque de crédit de l'actif financier concerné, l'établissement de crédit sera amené à comptabiliser les pertes attendues sur 12 mois pour les crédits sains ou jusqu'à maturité pour les actifs risqués. Le Bucket 3 regroupe les actifs dont les pertes sont déjà avérées. Pour finir, on comptabilise la couverture des instruments financiers qui ont été classés.

La norme IFRS 9 est donc moins complexe que la norme IAS 39. De plus, IFRS 9 est plus en adéquation avec la façon dont les établissements gèrent le risque de crédit. Enfin, l'atout principal de la norme IFRS 9 est qu'elle permet aux institutions financières de prévoir et d'avoir une vision de leurs dépenses futures.

3 Risque de crédit

Le risque de crédit représente la probabilité de perte financière liée au défaut de remboursement d'un emprunteur auprès d'un établissement de crédit. Dans toutes les activités économiques, lorsque la banque attend un paiement d'un client, elle s'expose à la possibilité que ce dernier ne le rembourse pas aux échéances attendues par le contrat initial. Il est donc nécessaire pour les institutions financières de maîtriser le risque de crédit afin de minimiser ses pertes. Il existe trois paramètres que l'on appelle paramètres de risques qui vont permettre de quantifier le risque de crédit. Il s'agit de la Probability of Default, la Loss Given Default et l'Exposure At Default.

3.1 Exposure At Default

L'Exposure At Default (EAD) ou l'exposition en cas de défaut représente l'exposition encourue par un créancier en cas de défaut de son débiteur. Autrement dit, il s'agit du montant total que l'établissement de crédit sera susceptible de perdre si jamais le client venait à faire défaut.

3.2 Loss Given Default

La Loss Given Default correspond à l'évaluation de la perte encourue en cas de défaut du client. La LGD s'exprime en pourcentage. Par exemple, si : $LGD = 40\%$, alors en cas de défaut de la contrepartie, le créancier perdra 40% de son montant initial.

3.3 Probability of Default

La Probabilité de Défaut est le paramètre de risque le plus étudié par les établissements de crédits. Elle détermine la probabilité qu'un client fasse défaut c'est-à-dire qu'il ne puisse pas faire face à ses obligations de remboursements.

3.4 Expected Credit Loss

Ces trois paramètres de risques permettent en particulier de calculer les pertes que les établissements s'attendent à perdre lors de l'octroi d'un crédit. Nous résumons la gestion des provisions et du capital réglementaire que les établissements doivent impérativement respecter sur la figure suivante.

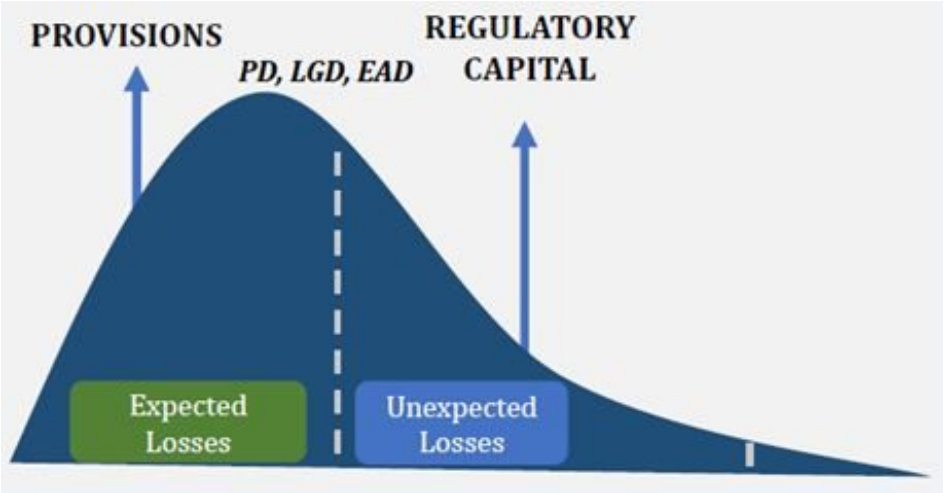

Figure 3.1: Provision et capital réglementaire

Comme nous l'observons sur cette figure, on distingue trois grandeurs :

- Expected Credit Loss : il s'agit du coût normal auquel les établissements sont exposés en accordant des crédit. On calcule cette grandeur par la formule $ECL = PD \cdot LGD \cdot EAD$.
- Unexpected Credit Loss : il s'agit du capital réglementaire que les établissements devront fournir afin de compenser les pertes qui sont impossibles d'anticiper.
- Les pertes extrêmes en cas de crise.

3.5 Scoring de crédit

Après avoir présenté les trois paramètres de risques permettant de quantifier le risque de crédit, il est difficile pour les institutions financières de déterminer à qui octroyer un crédit. C'est pour répondre à cette problématique que le Scoring de crédit a été introduit. L'objectif de cette méthode est d'associé un score à chaque client et en fonction de ce score, l'établissement de crédit pourra déterminer s'il est préférable ou non d'accorder un crédit à un client.

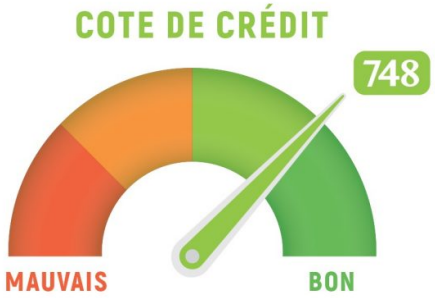

Figure 3.2: Exemple d'échelle de score de crédit

Dans le cadre de cette étude, nous allons appliquer une méthode de Scoring sur une régression logistique afin de déterminer si un crédit sera plus à même d'être catégorisé comme sain (S1) ou comme risqué (S2).

4 Statistiques descriptives et nettoyage des données

Avant d'entamer toute étude statistique, il est indispensable de s'assurer de la qualité des données. Après avoir décrit quelques caractéristiques de notre population d'étude, nous allons présenter le nettoyage des données que nous avons effectué sur notre base de données.

4.1 Description de la population d'étude

Afin d'étudier la répartition des clients selon les variables que nous avons retenu après le nettoyage de données, nous avons appliqué le test non paramétrique[1] de Wilcoxon sur quelques variables quantitatives que nous avons trouvé pertinentes pour la problématique de cette étude.

	Ensemble de la population N = 2 242 784	**Clients tombés en défaut** N=54 747 (2.4%)	**Clients non tombés en défaut** N= 2 188 037 (97.6%)	**p-valeur** du test
PD à l'octroi	0.03 [0.01 ; 0.06]	0.04 [0.02 ; 0.10]	0.03 [0.01 ; 0.06]	<0.01
PD à reporting	0.01 [0.004 ; 0.04]	0.08 [0.03 ; 0.16]	0.01 [0.004 ; 0.03]	<0.01
Maturité résiduelle (en années)	1.9 [1 ; 12]	1 [1 ; 4.8]	2 [1 ; 12.2]	<0.01
Nominal de départ (en milliers d'€)	25 [0 ; 200]	2.2 [0 ; 87.6]	26 [0 ; 200]	<0.01

Table 4.1: **Caractéristiques des clients selon différentes variables**. Médiane [intervalle interquantille]

Comme nous l'observons dans le tableau 4.1, seulement 2.44% des clients de la base de données sont tombés en défaut. Que ce soi à l'octroi ou à la date du reporting, la probabilité de défaut est globalement plus élevé pour les clients qui sont tombés en défaut plutôt que pour les autres clients. De même, les clients ayant en leur possession un nominal de départ plus élevé auront moins tendance à faire défaut. Ces chiffres et premières analyses sont plutôt rassurantes d'un point de vu métier. Nous présentons en annexe le dictionnaire des variables retenues.

[1] Ce test est dit non-paramétrique car il ne fait aucune hypothèse sur la distribution des échantillons.

4.2 Nettoyage des données

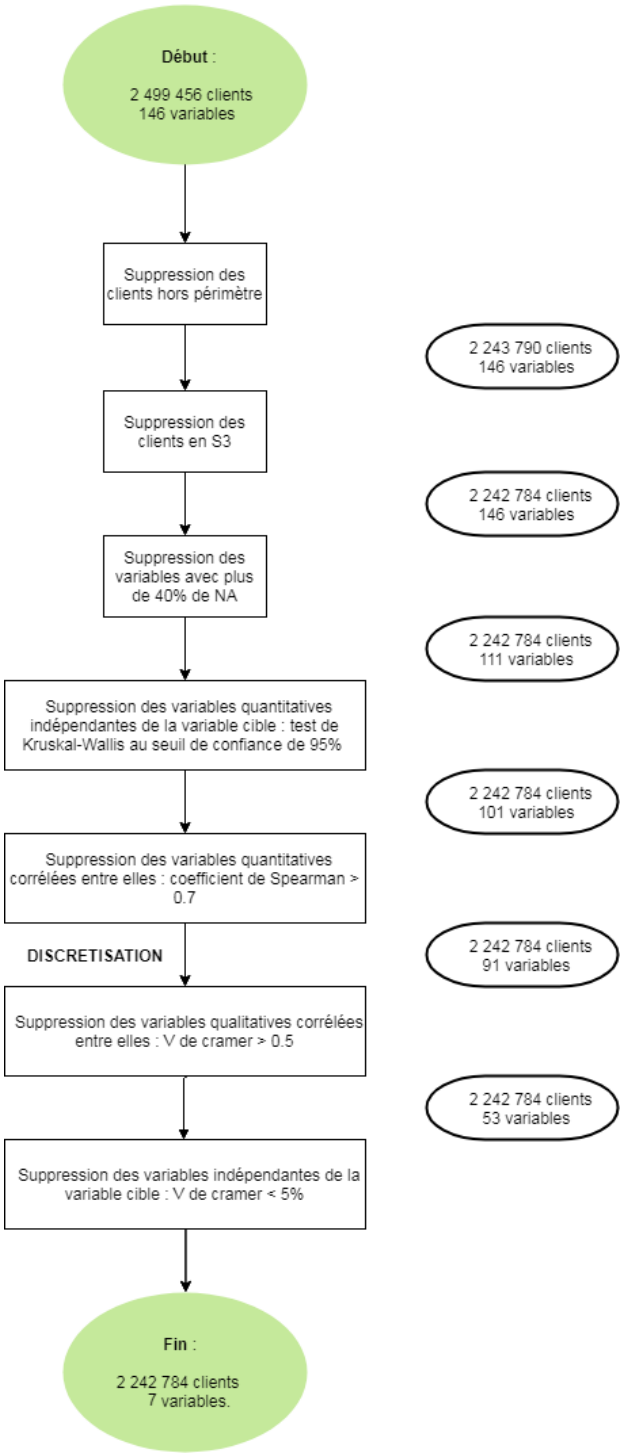

Figure 4.1: Nettoyage de données effectué sur notre base de données

Parmi ces 15 variables que nous avons retenu, nous avons testé différents modèles. Nous avons choisi de retenir un modèle à sept variables. Nous présentons en annexe ce choix de modèle.

5 Modélisation de la classification en S1 ou en S2

Après avoir introduit les outils nécessaires à la compréhension de notre problématique, nous allons présenter les différents modèles que nous avons pu implémenter sur la régression logistique. Cela nous permettra d'en comparer les performances et éventuellement de privilégier un modèle plutôt qu'un notre.

5.1 Le modèle logistique

On considère une phénomène d'intérêt (le défaut, par exemple) pouvant être représenté par une variable aléatoire binaire $Y \in \{0, 1\}$. Ici, la variable cible Y que l'on cherche à étudier est une variable de défaut sur un horizon de douze mois nommée "defaut_hoz_cible_12". On considère également k variables explicatives regroupées dans une matrice $X = (1, X_{\cdot 1} \dots X_{\cdot k})$. On cherche à formaliser le lien existant entre les caractéristiques observées $(X = x)$ et la réalisation du phénomène cible Y. L'approche de la régression logistique consiste à étudier l'espérance conditionnelle
$\mathbb{E}(Y|X = x) = \mathbb{P}(Y|X = x) = p(x)$

5.1.1 Spécification du modèle

Le modèle logistique introduit une fonction de lien *logit* pour estimer la probabilité $p(x)$. La spécification du modèle est la suivante :

$$logit\left(p(x)\right) = x\beta = \sum_{j=1}^{k} x_{\cdot j}\beta_j \iff p(x) = \frac{\exp(x\beta)}{1 + \exp(x\beta)}$$

où $\beta = (\beta_0, \dots, \beta_k)^T$ est le vecteur des paramètres qu'on estimera par maximum de vraisemblance. Des résultats théoriques permettent de démontrer que l'estimation par maximum de vraisemblance fournira un résultat efficace, convergent et asymptotiquement gaussien (voir Hosmer & Lemeshow, 2013).
L'application de la fonction de lien "logit" permet alors de traiter les variables cibles binaires et on peut ainsi calculer la probabilité de défaut $p(x)$ pour chaque individu i.

5.1.2 L'odds ratio

Outil de base utilisé en régression logistique, l'odds de $Y = 1$ sachant $X = x$ pour un individu A est donné par la relation $odds(A) = \frac{p(x)}{1-p(x)} = e^{x\beta}$. Pour comparer deux individus A et B de

caractéristiques respectives $x_{1.}$ et $x_{2.}$, on utilise le rapport de côtes ou odds ratio définit par :

$$OR(A,B) = \frac{odds(A)}{odds(B)} = e^{(x_{1.} - x_{2.})\beta}$$

Ainsi, si on veut comparer deux individus qui ne diffèrent que par la caractéristique j, alors on obtient $OR(A,B) = e^{(x_{1j} - x_{2j})\beta_j}$. De plus, si le régresseur est binaire, alors $OR(A,B) = e^{\beta_j}$. On dit qu'on mesure l'influence de la variable $X_{.j}$ toutes choses égales par ailleurs. Enfin, il est indispensable d'avoir toutes les odds ratios entre les différentes modalités de nos variables du modèle stricte supérieures à 1 (intervalle de confiance à 95% ne contenant pas 1) afin de garantir la significativité du score obtenu dans la partie suivante.

5.1.3 Régression logistique

Avant de lancer notre régression logistique sur notre base de données, nous avons procédé à un échantillonnage des données. Ainsi, nous avons séparé la base en une base d'apprentissage représentant 80% des données initiales qui nous permettra de créer notre modèle, puis une base de test contenant les 20% restant sur laquelle nous testerons les différentes performances du modèle créé. On rappelle que l'on cherche à expliquer la variable "defaut_hoz_cible_12" en fonction des sept variables que nous avons retenu.

5.1.4 Attribution des scores

Pour chaque client de notre base de données, nous avons attribué un score compris entre 0 et 1000 permettant de quantifier le risque qu'un client fasse défaut. Ce score est calculé à partir de la contribution de chaque modalité des variables que nous avons retenu dans notre modèle finale. Plus la modalité est risquée, moins sa contribution sera élevée. Ainsi, plus le score du contrat est faible, plus il sera risqué pour l'établissement d'octroyer un crédit à ce client.

Implémentation de la formule du calcul des scores

Tout d'abord, la discrétisation et l'exploration des données nous a permis de montrer que les modalités des différentes variables n'ont pas forcément la même intensité d'influence sur la variable cible. Pour prendre en compte ce phénomène, nous utilisons une formule de calcul d'un score qui permettra d'associé un poids proportionnel à l'importance de la modalité. Pour chaque variable j et chaque modalité i, on note $c(j,i)$ et $SC(j,i)$ respectivement le coefficient estimé par le modèle et le score associé à la modalité j.On note également $alpha_j = max(c(j,i))$.

Par la suite, on calcule le score de la modalité i d'une variable j à l'aide de la formule suivante :

$$SC(j,i) = 1000 * \frac{|c(j,i) - \alpha_j|}{\sum_j \max_i c(j,i)}$$

Cette formule permet en particulier de comparer chacune des modalités par rapport au client le plus risqué. La multiplication de la fraction par 1000 permet d'obtenir un score sur une échelle de 0 à 1000. Notons que cette formule prend sens lorsque toutes les estimations des coefficients du modèle sont positives ce qui est bien le cas pour nos résultats.

Table des scores

On retrouvera ci-dessous un tableau récapitulatif des variables retenues pour la régression logistique, les modalités associées à ces variables, les estimations des coefficients obtenues par la régression logistique ainsi que les scores calculés à partir de la formule énoncée précédemment. En outre, on remarque qu'aucune variable contribuait à plus de 50% dans la construction du score ce qui est rassurant pour l'interprétation et l'analyse des résultats. Nous nous sommes assurés que chacune des modalités était significative au seuil de 1%.

Variables	Modalités	Estimations	Scores
Intercept		-6,6877	
note_date	1 - 3	0	443,813779
Note de la PD à date du	4 - 6	1,3732	236,519534
porteur du risque	7	2,3074	95,4954411
	8	2,94	0
note_ecart	3 – 7	0	54,7068414
Différence entre la note	0 – 2	0,1478	32,3953868
à l'octroi et la note à date	(-2) – (-1)	0,2857	11,5784071
	(-7) – (-3)	0,3624	0
cod_seg_garant	Manquants	0,5928	70,6177163
Le segment du garant de	1600	0	160,105066
la garantie principale	2010	1,0606	0
	2020	0,7714	43,656784
	3200	0,6583	60,7300284
cod_cat_juridique_ne	1	0,3176	85,5322746
Code de la catégorie juridique	2	0	133,476239
du porteur de risque	3	0,8842	0
c_oc_taux_pd_octr_fi	[0 : 0,03]	0	65,6965159
PD à l'octroi	] 0,03 : 0,1]	0,3031	19,9414287
	] 0,1 : 0,89]	0,4352	0
c_fair_value	[0 : 70]	0	72,5197754
Valeur de marché de	] 70 : 345]	0,1085	56,1409335
l'exposition	] 345 : 1439885]	0,4804	0
c_mnt_cpta_8	0	0	69,6817825
Montant comptable	[0.1 :700000]	0,4616	0

Figure 5.1: Tableau récapitulatif des scores des modalités

Une fois les scores attribués aux différentes modalités, on calcule le score des différents clients de notre base de données en sommant les scores associés à chacune des modalités du client. Cela nous a donc permis d'ajouter une variable représentant le score de chaque client de la base de données.

Il nous est désormais possible de classer les contrats en S1 ou en S2 en se basant sur les différents scores obtenus par les clients.

5.2 Validation du modèle par bootstrap

La validation par bootstrap consiste tout d'abord à tirer N (ici, N=100) échantillons avec remise de notre échantillon de test. Ensuite, nous estimons les paramètres du modèle sur ces différents échantillons. Pour finir, nous estimons des statistiques sur ces paramètres et nous déduisons un intervalle de confiance des paramètres à 95 %. La table 5.1 présente les résultats obtenus. On remarque que les paramètres du modèle appartiennent à l'intervalle de confiance des paramètres estimés par bootstrap. En effet, ces paramètres sont très proches de la moyenne des échantillons bootstrapés. On peut donc conclure que notre modèle est plutôt robuste car il semble stable par rapport aux différents échantillons bootsrap générés.

Num	Variable	Cat	Estimation	Min	Max	Moyenne	Ecart-type	Borne Inf	Borne Sup	Estimate in IC ?
0	**Intercept**		-6,688	-7,437	-5,672	-6,642	0,341	-7,310	-5,973	VRAI
1	**note_date**	**2**	1,373	0,911	1,759	1,376	0,169	1,046	1,706	VRAI
2	**note_date**	**3**	2,307	1,897	2,804	2,337	0,203	1,938	2,735	VRAI
3	**note_date**	**4**	2,940	2,437	3,583	3,031	0,219	2,601	3,461	VRAI
4	**cod_seg_garant**	**2010**	0,593	0,221	1,814	1,005	0,350	0,319	1,692	VRAI
5	**cod_seg_garant**	**2020**	1,061	-0,273	1,533	0,688	0,389	-0,074	1,451	VRAI
6	**cod_seg_garant**	**3200**	0,771	-0,141	1,571	0,637	0,352	-0,053	1,327	VRAI
7	**cod_seg_garant**	**404**	0,658	-0,094	1,402	0,563	0,326	-0,077	1,203	VRAI
8	**cod_cat_juridique_ne**	**1**	0,318	0,036	0,621	0,313	0,130	0,058	0,569	VRAI
9	**cod_cat_juridique_ne**	**3**	0,884	0,627	1,208	0,886	0,120	0,650	1,122	VRAI
10	**note_ecart**	**2**	0,148	-0,331	0,412	0,091	0,159	-0,220	0,403	VRAI
11	**note_ecart**	**3**	0,286	-0,180	0,458	0,174	0,132	-0,084	0,431	VRAI
12	**note_ecart**	**4**	0,362	-0,336	0,645	0,293	0,191	-0,081	0,668	VRAI
13	**c_oc_taux_pd_octr_fi**	**2**	0,303	0,023	0,627	0,302	0,127	0,052	0,551	VRAI
14	**c_oc_taux_pd_octr_fi**	**3**	0,435	0,034	0,786	0,434	0,160	0,120	0,748	VRAI
15	**c_fair_value**	**2**	0,109	-0,460	0,401	0,037	0,158	-0,273	0,346	VRAI
16	**c_fair_value**	**3**	0,480	0,065	0,804	0,480	0,156	0,174	0,787	VRAI
17	**c_mnt_cpta_8**	**2**	0,462	-0,172	1,016	0,414	0,211	0,000	0,827	VRAI

Table 5.1: Résultats de la validation par bootstrap

On remarque aussi que le D de Sommers (cf Table 5.2) obtenu à partir de l'estimation du modèle (0.652) est très proche de la moyenne obtenue sur les échantillons par bootsrap (0.654). Cela nous permet donc de conclure que le modèle est stable.

N	value	Moyenne	Ecart-type	Minimum	Maximum
100	0,652	0,654	0,017	0,603	0,696

Table 5.2: Résultats du D de Sommers la validation par bootstrap

5.3 Détermination du seuil de distinction des contrats en S1 & S2

Une fois notre modèle validé, on peut entamer la phase de détermination du seuil de score qui permet de distinguer les contrats en S1 de ceux en S2, ceci étant l'objectif principal de notre étude. Pour ce faire, nous avons testé deux approches différentes.

5.3.1 Approche graphique par densité conditionnelle du score

Une façon de déterminer le seuil qui sépare les contrats S1 des contrats S2 est la représentation graphique de la densité du score conditionnelle au défaut du client. On représente d'une part la courbe densité des scores clients ayant fait défaut (en rouge), d'autre part la courbe de densité des scores clients n'ayant pas fait défaut (en noir).

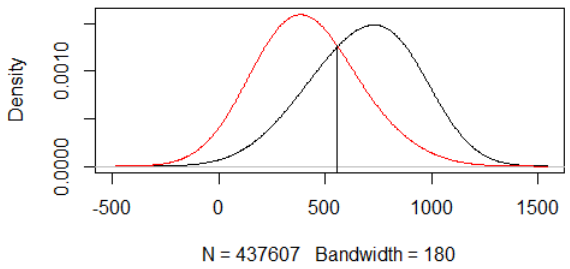

Figure 5.2: Approche graphique par densité conditionnelle du score

On observe qu'il y a un unique point d'intersection entre les deux courbes. Afin d'obtenir l'abscisse du score à l'intersection de ces deux courbes, il est possible d'employer deux méthodes :

- Une méthode graphique.
- Une méthode mathématique où on égalise les deux équations des densités des scores.

Afin de gagner du temps, nous avons opté pour la méthode graphique qui nous permet de fixer un score optimal égal à **530**. Ainsi, pour cette méthode, les clients dont le score sera strictement inférieur à 530 seront classés en S2 et les clients dont le score sera supérieur à 530 seront classés en S1.

Cependant, cette approche repose sur la normalité des échantillons étudiés ce qui n'est pas le cas ici. En effet, on observe les QQplot suivants :

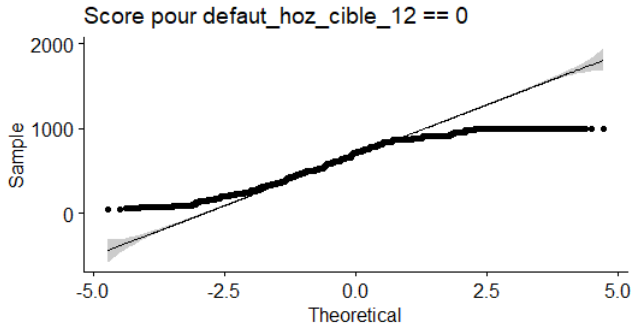

Figure 5.3: QQplot pour les clients non tombés en défaut

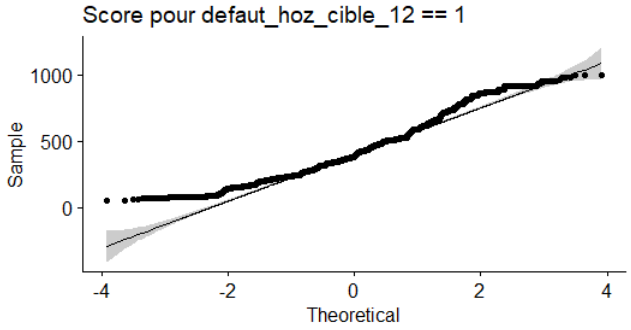

Figure 5.4: QQplot pour les clients en défaut

Le test de Shapiro-Wilk indique des p-valeurs pour chacun de ces échantillons inférieure à 5%, nous avons donc écarté cette approche de staging pour la suite de cette étude car les densités Gaussiennes représentées ne reflètent pas réellement les distributions des différents scores étudiés.

5.3.2 Approche par maximisation du V de Cramer

Il existe d'autres façons de choisir un score pour la classification en S1 ou en S2. Il s'agit de faire varier le score (on classera alors les clients à gauche du score en S2 car risqué et les clients à droite du score en S1).
Pour chaque score, on calculera alors le V de Cramer entre la classification en S1 ou en S2 selon le score et la variable de défaut. On cherchera ainsi le score qui va maximiser le V de Cramer.
Nous avons donc implémenté une boucle qui évalue le V de Cramer entre la variable cible de défaut et les stages (S1 & S2) en faisant varier de 0 à 1000 le score par pas de 10 à chaque itération. On trouvera ci-dessous une représentation graphique de la mesure du V de Cramer à chaque itération.

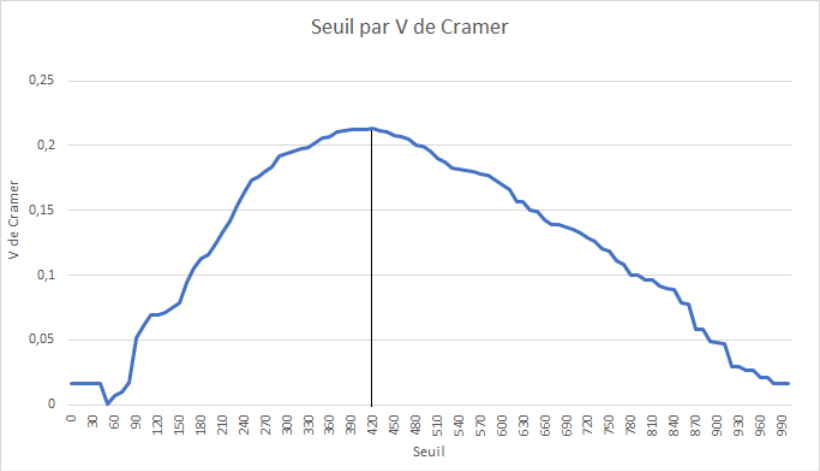

Figure 5.5: Approche par maximisation du V de Cramer

On remarque que le score optimal selon le test du V de Cramer est maximale pour une valeur de **420** avec une valeur du V de Cramer égale à **21%**. Pour la suite de notre étude, nous allons considérer ce seuil de staging de 420 pour étudier la performance de notre classification.

5.4 Analyse des résultats

5.4.1 Matrice de confusion

Afin de vérifier la qualité de notre classification et le choix du seuil de staging à 420, nous avons construit une matrice de confusion sur l'échantillon de test.

Contrats	**S1**	**S2**
Contrats sains	392 791 **87,6%**	44816 **10%**
Contrats en défaut	5062 **1,1%**	5887 **1,3%**

Table 5.3: Matrice de confusion sur l'echantillon de test

On retrouve les 2.4% de contrats en défaut cités précédemment. Cette matrice nous permet de calculer quelques indicateurs statistiques de performance de notre modèle de classification.

Précision

La précision du modèle se calcule à l'aide de la formule suivante :

$$Précision = \frac{VP}{VP + FP} = \frac{392791}{392791 + 44816} = \mathbf{0{,}89}$$

Ainsi, notre modèle possède une précision de **89%**. Cela signifie que l'on prédit avec une certitude de 89% qu'un contrat devrait être classé en S1.

Rappel (sensibilité)

Le rappel du modèle se calcule à l'aide de la formule suivante :

$$Rappel = \frac{VP}{VP + FN} = \frac{392791}{392791 + 5062} = \mathbf{0{,}98}$$

Notre modèle possède un rappel de **98%**. En d'autres termes, il associe correctement 98% des contrats à leurs stages adéquats. En vu des indicateurs de performance élevés, on peut conclure que notre modèle de classification est sensible et précis. On note que ces deux indicateurs doivent être analysés simultanément et qu'un bon compromis entre les deux assurera un modèle davantage performant.

Pour finir, nous avons obtenu selon notre modèle un taux de contrats classés en S2 égal à 11,3%. Le modèle créé par nos tuteurs de stage indiquait un taux de contrats classés en S2 environ égal à 10%. Il est donc rassurant que les résultats que nous avons obtenu en fixant un score de 420 est en adéquation avec les résultats obtenus par les modélisateurs de la BPCE.

5.4.2 Distinctions des contrats BP des contrats CE

Comme nous l'avons remarqué lors de la partie d'analyse des données, le groupe BPCE est issu de la fusion entre la Banque populaire et la Caisse d'épargne. Dans cette partie, on présentera quelques résultats que nous avons jugé intéressants selon l'appartenance du contrat. Cette distinction est réalisée à l'aide de la variable **Lib_group_contrib** qui associe RBP pour les contrats associé à la Banque populaire et RCE pour ceux de la Caisse d'épargne.

Stages	**BP**	**CE**
S1	1 317 680 **87,5%**	670 966 **91%**
S2	188 023 **12,5%**	66 115 **9%**

Table 5.4: Effectif des deux stages selon la banque d'origine

D'après cette table, on remarque que la base de données qui nous a été fourni est constituée d'environ 65% de contrats issus de la Banque Populaire et 35% de contrats issus de la Caisse d'Epargne.
Selon le seuil de staging de 420, on trouve environ 12.5% de contrats en S2 pour la Banque Populaire contre 9% en S1 pour la Caisse d'Epargne. Ces résultats sont en adéquation le modèle interne de la BPCE : 15% en S2 pour les contrats BP, 12% pour les contrats CE.

Contrats	S1	S2
Contrats sains	259 889	33 005
Contrats en défaut	3814	4509

Table 5.5: Matrice de confusion pour données de la BP

Contrats	S1	S2
Contrats sains	132 902	11 811
Contrats en défaut	1248	1378

Table 5.6: Matrice de confusion pour données de la CE

Précision = 89% & Rappel = 99%

Précision = 92% & Rappel = 99%

On remarque que les deux indicateurs de performance étudiés pour chacune des deux banques sont du même ordre de grandeur, cependant, notre modèle semble plus précis pour la Caisse d'Epargne avec une précision de 92%. Il est important de constater cette précision supplémentaire car cela pourrait permettre à la banque de mieux gérer les provisionnements. Néanmoins, ces chiffres viennent valider le choix de notre seuil de score pour la classification en S1 ou en S2 et cela permet également de souligner la cohérence de notre modèle pour des contrats traités par deux organismes différents.

Exposure at Default

Il est toujours intéressant d'avoir une idée de l'ordre de grandeur de l'EAD au sein d'un établissement de crédit. On représente donc dans les deux tableaux ci-dessous les montants des expositions des deux banques selon la classification en S1 ou en S2. Notons que ces chiffres ont été calculé à partir de l'échantillon test contenant 20% de la base de données initiale.

Stages	EAD
S1	38,404 Md€
S2	5,546 Md€

Table 5.7: EAD selon les stages pour le groupement BP

Stages	EAD
S1	15,796 Md€
S2	1,633 Md€

Table 5.8: EAD selon les stages pour le groupement CE

Ces résultats sont plutôt rassurants. Tout d'abord, il est logique que le montant auquel est exposé la Caisse d'Epargne est beaucoup plus faible que le montant est exposé la Banque Populaire en raison des différences d'effectifs des contrats entre les deux banques.
De plus, les montants des expositions auxquels sont exposés les banques sont du même ordre de grandeur que les montants obtenus par nos tuteurs.

Conclusion

La base de données mise à disposition par la BPCE nous a permis de découvrir d'un point opérationnel la modélisation du risque de crédit. Comme nous l'avons déjà précisé, les normes réglementaires évoluent dans le temps et en particulier l'introduction de la nouvelle norme IFRS 9 permet d'obtenir un point de vu prospectif du risque de crédit. Dans ce rapport, nous avons eu l'opportunité d'appliquer la seconde phase de la norme IFRS 9 pour la classification des contrats sains (S1) ou risqués (S2).

Il nous a été nécessaire et primordiale de nettoyer avec efficacité la base de données qui nous a été fournis afin de pouvoir appliquer des méthodes de modélisation dans des conditions optimales.

Ensuite, nous avons appliqué une régression logistique qui est une méthode de modélisation réputée comme interprétable et plutôt simple à implémenter. Cela nous a permis de créer une échelle de score afin de quantifier le risque de défaut d'un client. Ensuite, nous avons calibrer différents seuils de score afin de déterminer le score optimal permettant de classifier les contrats en S1 ou en S2. Il a été très intéressant pour nous d'appliquer la régression logistique sur les données réelles de la BPCE. Cela nous a permis non seulement d'approfondir nos connaissances statistiques sur la régression logistique ainsi que sur les méthodes de classification mais également d'avoir une première approche avec les techniques de modélisations du risque de crédit.

Le travail que nous avons effectué présente cependant quelques limites. En effet, bien que nous ayons pu comparer nos résultats de modélisations aux résultats obtenus par le modèle actuel de la BPCE, nous aurions apprécier de pouvoir utiliser d'autres modèles statistiques que la régression logistique. Ainsi, si nous avions eu le temps, nous aurions pu appliquer des méthodes de sélection de variables en utilisant des méthodes de régression pénalisée (Lasso, Ridge, Elastic net ...) ou encore des méthodes d'arbres de classification (Random Forest, Bagging, Boosting). Il aurait été également valorisant d'utiliser les SVM que nous avions cité dans notre projet bibliographique pour appliquer la seconde phase de la norme IFRS 9. Ainsi, nous aurions pu comparer les différentes performances associés aux modèles créés. Il aurait également été intéressant de stress-tester notre modèle, c'est-à-dire d'observer et d'analyser l'évolution des probabilités de défaut en période de crise.

Pour finir, il nous a été très instructif de découvrir des variables contenues dans une vraie base de données bancaire. Et de pouvoir appliquer les différentes méthodes d'analyse et de modélisation que nous avons appris lors de notre parcours universitaire afin de répondre à une problématique ayant de réelles enjeux sur le portefeuille de la BPCE.

A Annexe

A.1 Tests statistiques utilisés

A.1.1 Le coefficient de Pearson

Le ρ de Spearman est un cas particulier du coefficient de corrélation linéaire (r) de Pearson. Il s'agit d'un test non paramétrique se basant sur les rangs des observations plutôt que les valeurs. Soit R_i et S_i les rangs respectifs de x_i et y_i

$$\rho = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \overline{S})^2}}$$

Il a l'avantage de pouvoir décelé les relations non linéaires entre les variables.

A.1.2 Test d'indépendance du χ^2

Le test du χ^2 permet de tester l'indépendance entre deux variables qualitatives (à plusieurs modalités) que l'on notera X et Y d'un N-échantillon. La statistique associée à ce test est basée sur le calcul des distances entre les valeurs théoriques et les valeurs observées de l'échantillon. Les observations sont réparties dans un tableau de contingence contenant en ligne et en colonne les modalités de chaque variable. Les hypothèses du test sont les suivantes:

$$\begin{cases} H_0 : & \text{Les deux variables sont indépendantes} \\ H_1 : & \text{Les deux variables sont liées} \end{cases}$$

Sous l'hypothèse nulle, la statistique du test est donnée par:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij} - e_{ij})^2}{e_{ij}} \underset{N\to\infty}{\longrightarrow} \chi^2((r-1)(c-1))$$

Où :
- o_{ij} est l'effectif observé des clients pour lesquels X prend la modalité i et Y la modalité j.
- $e_{ij} = \frac{r_i \times c_j}{n}$ est l'effectif théorique attendu.
- $r_i = \sum_{j=1}^{c} o_{ij}$ et $c_j = \sum_{i=1}^{r} o_{ij}$.

Ainsi pour un seuil α donné, l'hypothèse nulle du test est rejetée dès que $\chi^2 \geq \chi^2_\alpha((r-1)(c-1))$. La condition d'application du test est que tous les effectifs attendus e_{ij} soient supérieurs ou égaux à 5. Dans le cas contraire, il est usuel de regrouper certaines classes (de variable X par exemple).

A.1.3 Test V de Cramer : liaison entre variables qualitatives

Le V de Cramer permet de mesurer l'intensité de la liaison entre deux variables qualitatives indépendamment du nombre de leur modalité et de l'effectif de la population. Il est donné par la racine carrée du χ^2 divisé par le χ^2_{max}. Ce χ^2_{max} théorique est égal à l'effectif multiplié par le plus petit côté du tableau (nombre de lignes ou de colonnes) moins 1. Formellement, le V est de Cramer est donnée par:

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{N \times [\min(r,c) - 1]}}$$

Où χ^2 est la statistique du test de χ^2 de liaison entre les deux variables qualitatives.
Ainsi plus la statistique V est proche de zéro, plus les variables étudiées sont indépendantes. Au contraire, donc, il vaudra 1 lorsque les deux variables sont complètement dépendantes, puisque le χ^2 est alors égal au χ^2_{max} (dans un tableau 2×2, il prend une valeur comprise entre -1 et 1). Donc, plus V est proche de 1, plus la liaison entre les deux variables étudiées est forte.

A.1.4 Test de Kruskal Wallis

Le test de Kruskal Wallis est un test non paramétrique qui permet de mesurer la liaison entre une variable quantitative X et une variable qualitative Y à $k \geq 2$ modalités. La variable X est ainsi observée sur les k groupes, représentant les modalités de la variable qualitative. Il s'agit d'un test bilatéral dont l'hypothèse nulle est que les groupes proviennent de la même loi et leurs rangs ne diffèrent pas (égalité des médianes). Si on désigne par n_i l'effectif du groupe i, R_i le rang moyen des observations dans le groupe i et $N = n_1 + \ldots n_k$ l'effectif total, la statistique KW du test vérifie :

$$KW = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \times R_i^2 - 3(N+1) \underset{n_i \to \infty}{\longrightarrow} \chi^2(k-1)$$

Le rejet de l'hypothèse nulle indique donc une liaison entre X et Y et cette relation est d'autant plus forte que la la statistique KW est grande.

A.2 Test de normalité

Le test de Shapiro-Wilk permet de tester la normalité d'une distribution. On a les hypothèses suivantes:

$$\begin{cases} H_0 : & \text{L'échantillon étudié est gaussien} \\ H_1 : & \text{L'échantillon étudié n'est pas gaussien} \end{cases}$$

Après avoir implémenté ce test sur R, nous avons obtenu les résultats suivants :

- Clients tombés en défaut : p-value $<$ 2.2e-16 $<$ 5%. Cela signifie l'absence de normalité.
- Clients qui ne sont pas tombés en défaut : p-value $<$ 2.2e-16 $<$ 5% . Cela signifie l'absence de normalité.

Nous avons également représenté les qqplot afin de confirmer que les échantillons étudiés n'étaient pas gaussien.

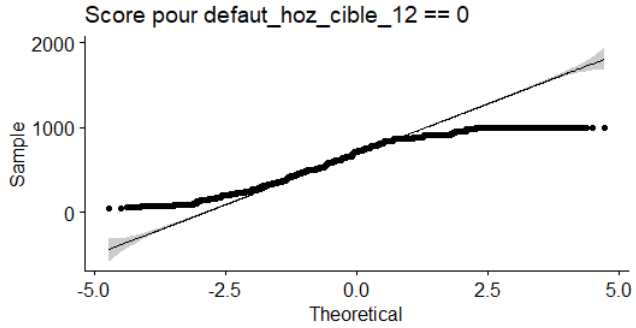

Figure A.1: Test de normalité du score pour les clients n'ayant pas fait défaut

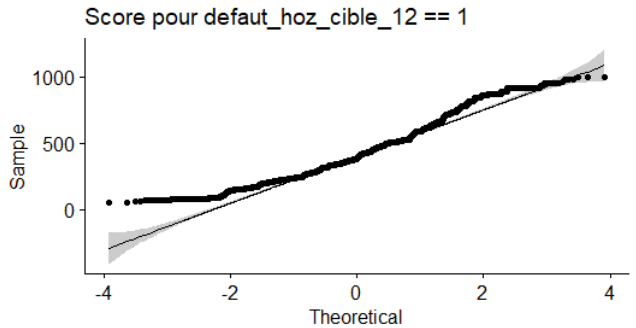

Figure A.2: Test de normalité du score pour les clients ayant fait défaut

B Corrélations des variables

B.1 V de Cramer des variables retenues

	defaut_hoz_cible_12
c_pd_new	0,166053532
c_pd_ndod	0,165136307
c_pd_externe_new	0,14867922
note_ecart	0,11326711
c_pd_ecart	0,103800262
class_pd_n3_final	0,078388135
cod_cat_juridique_new	0,077795252
c_mnt_cpta_8	0,073192853
c_oc_taux_pd_octr_final	0,070181941
c_fair_value	0,060443372
cod_nat_garantie	0,058816513
c_mnt_imp	0,047799019
c_nominal_de_depart	0,044528321
c_tau_lgd	0,044253833
c_nb_j_impayes	0,044009333
asset_class_new	0,043539635
c_valr_tx	0,024072849
c_tx_intr_effc_tie	-0,02942674
lib_typ_remboursement	-0,030548292
lib_famille_prdt	-0,044779664

Figure B.1: Corrélations des variables retenues à la variable cible

Ce tableau nous permet donc de conserver onze variables. Ensuite, nous avons utilisé la matrice de corrélation afin d'éliminer les variables explicatives corrélées entre elles. Cela nous a permis de ne conserver que sept variables. Notons que nous avons privilégié l'utilisation de variables qui sont utilisés en pratique (Pd_ndod plutôt que pd_new par exemple).

B.2 Matrice de corrélation des variables finales

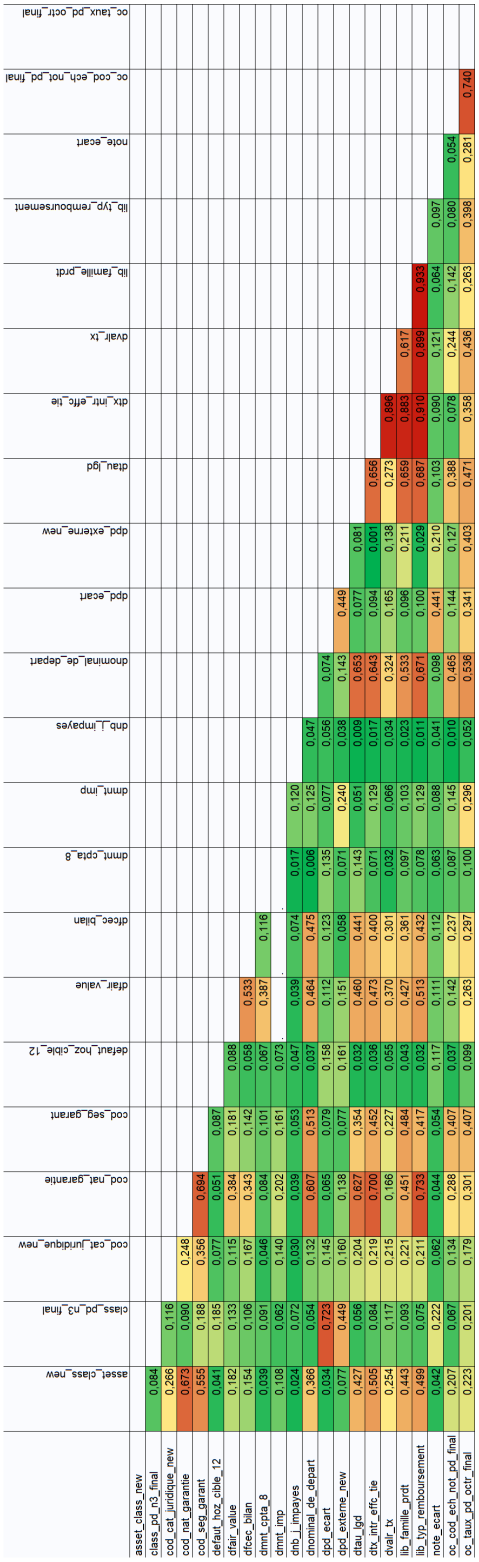

Figure B.2: Matrice de corrélation des variables finales (V de Cramer)

C Mise en place du modèle de régression logistique

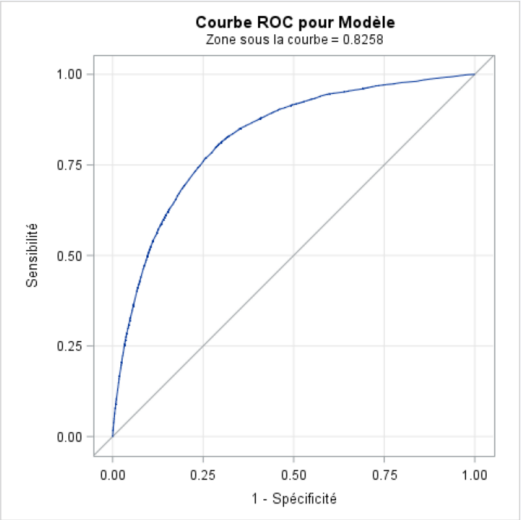

Figure C.1: Courbe ROC et AUC du modèle retenu

Estimations des rapports de cotes			
Effet	Valeur estimée du point	95% Intervalle de confiance de Wald	
note_date 2 vs 1	3.880	3.706	4.062
note_date 3 vs 1	9.867	9.348	10.414
note_date 4 vs 1	18.552	17.502	19.664
cod_seg_garant 404 vs 1060	1.809	1.683	1.944
cod_seg_garant 2010 vs 1060	2.887	2.670	3.123
cod_seg_garant 2020 vs 1060	2.162	1.994	2.343
cod_seg_garant 3200 vs 1060	1.931	1.785	2.089
cod_cat_juridique_ne 1 vs 2	1.374	1.334	1.415
cod_cat_juridique_ne 3 vs 2	2.420	2.355	2.488
note_ecart 200 vs 100	1.052	0.983	1.127
note_ecart 300 vs 100	1.217	1.131	1.309
note_ecart 400 vs 100	1.400	1.296	1.513
note_ecart 500 vs 100	1.516	1.392	1.650
c_oc_taux_pd_octr_fi 2 vs 1	1.360	1.321	1.401
c_oc_taux_pd_octr_fi 3 vs 1	1.556	1.502	1.612
c_fair_value 2 vs 1	1.115	1.071	1.162
c_fair_value 3 vs 1	1.617	1.567	1.668
c_mnt_cpta_8 2 vs 1	1.586	1.512	1.664

Figure C.2: Odds ratios des modalités retenues

D Dictionnaire des variables

Libellé de la variable	**Signification de la variable**
note_ecart	Différence entre la note à l'octroi et la note à date
note_date	Classe de PD final
cod_cat_juridique_new	Code de la catégorie juridique du porteur de risque
c_mnt_cpta_8	Montant comptable associé aux Valeurs Nettes Investies
c_oc_taux_pd_octr_final	Probabilité de défaut à l'octroi
c_fair_value	Valeur de marché de l'exposition
cod_nat_garantie	Nature de la garantie

Table D.1: Dictionnaire des variables retenues

References

[1] K. K. L. Gang Donga, "Credit scorecard based on logistic regression with random coefficients," *International Conference on Computational Science, ICCS 2010*, 2012.

[2] L. C. T. Katarzyna Bijak, "Does segmentation always improve model performance in credit scoring," *Expert Systems with Applications*, vol. 39, pp. 2434–2442, 2012.

[3] T. M. Lkhagvadorj Munkhdalai, "An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments," *Sustainability*, 2019.

[4] A. Al-Aradi, "Credit Scoring via Logistic Regression," *Department of Statistical Sciences, University of Toronto, Toronto, Canada*, 2014.

[5] Y. Liu, "A Strategy on Selecting Performance Metrics for Classifier Evaluation," *International Journal of Mobile Computing and Multimedia Communications*, 2014.