Rapport de synthèse des résultats

Développement d'un score d'octroi

Réalisé par :

MEUNIER Karl (contribution 33%)

RAOUL-JOURDE Thibaut (contribution 33%)

YE Victor (contribution 33%)

Professeur :

Mr Christophe RAULT

Référent RCi :

Mme Giulia MORETTI

Master Économétrie et Statistique Appliquée, Université d'Orléans 2019-2020 (Décembre 2019)

Table des matières

I. Résumé

A. Résumé en français

Il est important pour tout établissement financier de s'assurer de la solvabilité de son client et de sa capacité à respecter son engagement et c'est pourquoi le prêteur doit évaluer tout risque de défaut de paiement encouru avant tout accord. Les méthodes statistiques ainsi que les techniques d'apprentissage automatique (Machine Learning) permettent d'analyser les données mises à disposition pour développer une cotation pour chaque client.

C'est ainsi que le score d'octroi est introduit, il correspond à une note qu'un établissement financier attribue à un individu en fonction de ses caractéristiques personnelles et cette note va aider l'établissement financier pour la prise de décision de manière rapide et efficace : accorder ou refuser un prêt.

Le crédit scoring est aujourd'hui devenu une pierre angulaire pour la gestion du risque au sein des établissements financiers mais sa construction reste fastidieuse puisque cela demande la répétition de plusieurs tâches pour chacune des variables explicatives et il y a une vraie nécessité d'arbitrage dans le choix des informations à conserver ou non et cela dépends du niveau d'expertise dans le domaine du risque de crédit.

Nous avons eu à disposition une base de données regroupant des informations concernant des cas réels d'acceptation ou de refus du prêt par l'entreprise et notre rôle a été de concurrencer le score en y apportant les connaissances dont nous disposions à la suite du cours de méthode de scoring, tout en essayant de traiter les problématiques métiers liées au risque de crédit.

B. Overview

It is important for any financial institution to ensure the solvability of its customer and its ability to meet its commitment and that is why the lender must assess any risk of default before any agreement is reached. Statistical methods as well as Machine Learning techniques allow to analyze the data provided to develop a rating for each customer.

This is how the credit score is introduced, it corresponds to a score that a financial institution assigns to an individual based on his or her personal characteristics and this score will help the financial institution to make a quick and efficient decision: to approve or reject a loan.

Credit scoring has now become a fundamental component of risk management in financial institutions, but its construction remains laborious since it requires the repetition of several tasks for each of the explanatory variables and there is a real need for arbitration in the choice of which information to keep or to not keep and this depends on the level of expertise in the field of credit risk.

We had at our disposal a database containing information on real cases of acceptance or refusal on the loan by the company and our role was to compete with the score by adding the knowledge we had after the scoring method course, while trying to deal with the business challenges related to credit risk.

II. Introduction

Lorsqu'un emprunteur souhaite financer un projet, il peut demander l'aide d'un établissement de crédit pour obtenir un prêt s'étalant sur une durée prédéfinie, en contrepartie de l'instauration d'un taux d'intérêt de la part du prêteur. Cependant, il est évident que ce dernier doit s'assurer de la solvabilité ainsi que de la capacité à rembourser de son client avant de s'engager afin de minimiser ses pertes.

Ainsi, ces établissements financiers vont se fier à la note calculée pour chaque client sur la base des informations déclaratives fournies par celui-ci pour se donner le droit d'accepter ou de refuser une demande de crédit. Cette note correspond au score d'octroi et se base donc sur toutes les informations que va déclarer le client (précisons que l'établissement ne possède aucune information comportementale sur ce-dernier).

RCI Bank & Services (Renault Crédit International) est spécialisée dans le crédit à la consommation dans le secteur automobile ; grâce à la base de données réelle fournie par l'établissement, notre objectif est de créer une grille de score pertinente permettant d'évaluer de façon juste la capacité du client à respecter ses engagements.

III. Construction du score

La construction du score passe tout d'abord par l'étape indispensable que représente le traitement des données. Il nous faut premièrement analyser les données mises à disposition afin d'en faire une sélection pour éviter les problèmes que peut engendrer un mauvais tri : forte corrélation, pas de significativité des variables explicatives, multi-colinéarité…

Cette étape représente la plus grosse partie du travail de construction d'une grille de score puisque la procédure LOGISTIC sous le logiciel SAS nous permettra aisément de calculer un score pour chaque caractéristique d'un individu.

A. La base de données

Nous avons à notre disposition un échantillon fourni par RCI Bank & Services nous permettant d'établir un score d'octroi. Elle regroupe les contrats liés aux crédits automobiles accordés aux professionnels et particuliers pour le financement de leurs véhicules.

Cette base comporte 43 variables pour un total de 68 538 observations. La période d'observation [1] s'étend du 21 mars 2011 jusqu'au 21 décembre 2018. Ce qui correspond à 7 ans et 9 mois.

La variable dépendante « we18 » indique « 1 » si le contrat est tombé en défaut dans les 18 mois qui suivent l'entrée en gestion et « 0 » si ce n'est pas le cas.

Pour chaque individu, on dispose ainsi de 42 autres variables pouvant expliquer cette variable de réponse, il y a des informations diverses allant de la qualité du véhicule acheté (neuf ou occasion) jusqu'aux montants des loyers perçus par chacun d'entre eux par exemple.

De premier abord, la table semble incomplète puisque nous pouvons aisément remarquer, à la suite de simples statistiques descriptives, qu'il existe des valeurs manquantes sur certaines variables, concernant un nombre (très) important d'observations.

C'est pourquoi, après avoir déduit qu'il existait une délimitation nette chez RCI Bank & Services pour les contrats particuliers et les contrats professionnels, nous préférons directement scinder la table pour regrouper les modalités « PRI » de la variable « ty_pp »[2] entre-elles et les modalités « PRO » de l'autre côté. Le traitement ainsi que les étapes de modélisations et de construction du score sera totalement indépendant entre les deux tables.

Nous utilisons pour la suite du rapport l'appellation PRI pour la table regroupant uniquement les contrats privés et l'appellation PRO pour la table regroupant les contrats professionnels. Nous analysons tout d'abord les caractéristiques propres à la table PRI pour ensuite s'occuper de la table PRO.

[1] D'après la variable correspondant à la date d'observation de chaque contrat : dt_dmd
[2] Concerne le code d'usage du véhicule de la personne physique (privé ou professionnel)

B. La table PRI

Nom de la variable	Définition de la variable
Ty_pp	Code usage du véhicule
No_cnt_crypte	Identifiant du contrat crypté
No_par_crypte	Identifiant du client crypté
Date_gest	Mois de la date d'entrée en gestion
Dt_dmd	Date de la demande
Duree	Durée prévisionnelle du financement
Genre_veh	Genre du véhicule
Mt_dmd	Montant du financement (cts)
Pc_appo	Pourcentage d'apport (%)
Produit	Type du produit
Qual_veh	Qualité du véhicule
Age_veh	Age du véhicule (en mois)
Ind_cli_rnva	Indicateur client renouvelant
Nb_imp_tot	Nombre d'impayés total
Nb_imp_an_0	Nombre d'impayés des 12 derniers mois
Age	Age du client (en années)
Csp	Catégorie socio-professionnelle
Etat_civil	Code état-civil
Mode_habi	Code mode d'habitation
Mt_sal_men	Montant du salaire annuel (cts)
Rev_men_autr	Autre montant mensuel (cts)
Mt_alloc_men	Montant de l'allocation mensuelle (cts)
Nb_pers_chg	Nombre de personnes à charge
Mt_rev	Montant revenu mensuel (cts)
Mt_loy_men_mena	Montant du loyer mensuel du ménage (cts)
Mt_men_pre_immo	Montant mensuel du prêt immobilier (cts)
Mt_men_eng_mena	Montant mensuel divers du ménage (cts)
Mt_charges	Montant des charges (cts)
Ind_fch_fcc	Indicateur de fichage
Mt_ech	Montant de l'échéance (cts)
Mt_ttc_veh	Prix du véhicule à financer (cts)
Part_loyer	Part de l'échéance (%)
Anc_emp	Ancienneté de l'emploi (en mois)
Copot_	Comportement de paiement
Pan_dir_	Qualité du dirigeant à rembourser
Secteur_	Secteur d'activité
Diag_fch_cli	Indicateur de fichage par les banques
Ind_imp_regu	Nombre d'impayés régularisés (sur les 12 derniers mois)

Les variables[3] listées ci-dessus correspondent à la table brute PRI sur lequel on va commencer notre tri. Nous avons supprimé les variables n'ayant aucun lien avec les clients privés (comme par exemple la variable « cote_bdf » correspondant à la cotation Banque de France pour les clients professionnels).

Ainsi, on se retrouve avec 38 variables explicatives pouvant potentiellement expliquer la variable de réponse « we18 » : nous remarquons que l'évènement « 1 » (contrat en défaut) pour cette variable dépendante est rare puisqu'il représente seulement (et heureusement pour la banque !) 1,60% de l'échantillon total (contre logiquement 98,40% de contrats sains).

[3] Les lignes avec l'arrière-plan couleur bleu ciel correspondent aux variables qualitatives, les autres sont les variables quantitatives.

Nous savons donc que nous sommes dans un cas d'évènement rare et qu'il faut sûrement apporter des modifications sur la table afin de permettre au modèle de détecter de façon plus optimale la possibilité que le client tombe en défaut.

1. Modification de la base de données

Pour cette première étape concernant la table PRI, nous avons effectué quelques procédures afin de lister les statistiques descriptives et nous n'avons trouvé aucune valeurs aberrantes pour aucune des variables. Nous décidons de nous séparer de la variable « ty_pp » puisqu'elle dispose uniquement d'une modalité : PRI. Il en est de même pour la variable « genre_veh » car celle-ci renvoie toujours la modalité « VP » (pour véhicule privé). La variable « pan_dir_ » a aussi été éjectée du modèle pour la raison similaire : les modalités sont trop déséquilibrées (99,93% pour la modalité « bon » et 0,07% pour « mauvais »).

Nous avons ensuite jugé pertinent de créer deux nouvelles variables supplémentaire à partir de deux variables existantes de la table, pouvant nous apporter autant, voir plus d'informations que d'origine :

- « date_dmd_month » : elle renvoie simplement le mois de la date de demande de chaque observation, on peut observer une certaine saisonnalité des évènements (baisse de la demande de crédit en juillet-août due aux congés d'été).
- « periode » : elle renvoie le nombre de mois écoulé entre la date de demande du crédit par le client jusqu'à la date de l'entrée en gestion du contrat.

L'étape suivante désormais est de séparer la table PRI en deux échantillons distincts et toujours représentatifs de la globalité : 70% pour l'échantillon d'apprentissage (*training sample*) et 30% pour l'échantillon de test (*test sample*) pour nous permettre de valider le modèle. A ce stade, la proportion des défauts et non-défauts est inchangée. Bien entendu, il est important de préciser que pour tout traitement des variables concernant la table d'apprentissage, les mêmes actions seront réalisées sur l'échantillon test (Annexe 1).

Cette division, faite de manière aléatoire, nous permet d'utiliser l'échantillon d'apprentissage pour le choix des variables à inclure dans le modèle final. L'échantillon de validation nous servira, comme son nom l'indique, à valider le modèle sélectionné. Ce processus nous permet de tester le modèle sur des données différentes que celles utilisées pour l'apprentissage du modèle afin d'éviter des taux de performances anormalement élevées. Cela va renforcer la robustesse du modèle pour la prédiction sur de nouvelles données.

A présent, nous nous concentrons sur le ré-échantillonnage de la table d'apprentissage de manière à réduire la proportion de « 0 » pour la variable de réponse afin de permettre un meilleur apprentissage du modèle sur les cas de défaut, en utilisant donc la technique de *l'undersampling* (grâce à la procédure SURVEYSELECT sur SAS), de manière à obtenir une proportion de 90% de « 0 » et 10% de « 1 ».

2. Quelles variables qualitatives conserver ?

L'une des règles principales pour inclure toute variable dans le modèle est conditionnelle à la significativité de celle-ci. En effet, elle doit pouvoir être capable d'expliquer de façon relativement importante la variable de réponse « we18 » auquel cas il est inutile de compliquer le modèle avec des variables non-pertinentes. Les tests statistiques que nous réalisons pour tester la significativité des variables diffèrent du type de variable mais il s'agit du test du khi-deux pour les variables qualitatives.

Une fois que la significativité de la variable est confirmée, il ne s'agit pas de l'inclure directement dans le modèle car d'autres facteurs peuvent entrer en compte, notamment le fait que nous pouvons encore améliorer le niveau d'information apporté par cette variable au sein du modèle en utilisant le recodage des modalités.

Pour finir, il faut veiller à ne pas inclure dans le modèle des variables qui seraient trop corrélées entre-elles puisque deux variables apportant la même information sur la variable de réponse pourrait créer un problème de multi-colinéarité, faussant l'apport d'information de la part des autres variables. L'exemple type d'un problème de corrélation serait une variable donnant le salaire mensuel du client en euros, et une variable donnant la même information à l'exception que le salaire est affiché en centimes d'euros.

a. Étude de la significativité : test du khi-deux

Le test d'indépendance du khi-deux permet de tester l'hypothèse nulle correspondant à l'absence de relation entre deux variables catégorielles, face à l'hypothèse alternative indiquant l'existence d'au moins une relation entre les deux variables. L'indépendance (non rejet de l'hypothèse nulle) nous indiquerait que la valeur d'une variable ne nous donne aucune information sur la valeur possible de l'autre. On dit que les variables sont indépendantes l'une de l'autre, à ne pas confondre avec le terme de « variable indépendante ».

Dans notre cas, le test du khi-deux va comparer la distribution observée et la distribution théorique de la variable explicative dans le cas où celle-ci suivrait l'hypothèse d'indépendance avec la variable de réponse « we18 ».

Les tableaux obtenus nous indique la statistique du khi-deux, qu'il faut comparer à la table de distribution du khi-deux paramétrée par le de degré de liberté. Notre choix pour le risque de premier espèce sera de 5% et nous sommes en mesure de rejeter ou non l'hypothèse nulle d'absence de relation.

Les résultats nous permettent donc de dire qu'il est possible d'éliminer les variables « ind_fch_fcc » ainsi que « diag_fch_cli ». En réalité, nous avons également lancé le test pour les variables[4] que nous avons évoqué comme supprimées dans la section 1 et notre intuition semble avoir été bonne puisque ces variables ne doivent pas être dans le modèle. Cela s'explique par le fait que les modalités sont trop déséquilibrées et donc que l'information qu'elle est censée apporter n'est pas suffisante pour la conserver dans le modèle.

La variable que nous avions créée « dt_dmd_month » n'est pas significative, « produit » et « ind_cli_rnva » également. Nous ne les conservons pas pour la suite du modèle.

Les variables « periode » ainsi que « secteur_ » présentent un potentiel important mais malheureusement, elles disposent de certaines modalités avec un effectif théorique trop faible pour pouvoir interpréter le khi-deux, nous pouvons les recoder afin de pouvoir possiblement exploiter ce potentiel (Annexe 2).

Nous gardons pour la suite les variables : « qual_veh », « csp », « etat_civil », « mode_habi », « copot_ ».

Le recodage des variables « periode » et « secteur_ » [5] nous permettent d'obtenir des groupes contenant une taille d'échantillon satisfaisante pour chaque modalité, et en faisant à nouveau le test du khi-deux, les deux variables sont significatives donc nous les ajoutons à la liste des variables qualitatives à conserver pour la suite (Annexe 3).

[4] « genre_veh » et « pan_dir_ »

[5] Les nouvelles variables créées après recodage se nomment respectivement « periodes » et « secteur ».

b. *Étude de la corrélation : test du V de Cramer*

Dans cette partie, nous allons reprendre uniquement les variables que nous avons jugés significatives avec le test du khi-deux. A présent, il faut regarder la corrélation des variables retenues entre-elles.

Le V de Cramer permet de mesurer le degré d'association entre deux variables catégorielles et est basé sur le χ^2 de Pearson. Le V de Cramer varie dans l'intervalle [0,1] et le tableau ci-dessous représente les interprétations que la communauté scientifique s'accorde à dire :

Valeur	Force du lien statistique
0	Absence de relation
Entre 0,05 et 0,10	Très faible
Entre 0,10 et 0,20	Faible
Entre 0,20 et 0,40	Modérée
Entre 0,40 et 0,80	Forte
Entre 0,80 et 1	Colinéarité

Le seuil que nous nous sommes fixé pour choisir d'inclure ou non les variables est de 0,3. Les variables présentant une corrélation supérieure à 0,3 ne seront pas incluses dans le modèle.

Les résultats indiquent une forte corrélation entre la variable « secteur » et « csp » (Annexe 4) : après avoir approfondi les recherches sur ces deux variables, nous pouvons noter quelques points intrigants comme par exemple le fait que des individus possèdent une catégorie socio-professionnelle totalement décalée de ce que indique la variable secteur d'activité. En effet, il est noté des clients travaillant comme agriculteur dans le secteur auto-école (?) (Annexe 5). A la suite de cet exemple, nous choisissons de ne plus continuer avec la variable « secteur », qui se retire de notre modèle.

« qual_veh » et « periodes » sont également corrélées et notons qu'il existe qu'une très faible minorité où il s'agit de véhicules d'occasion avec une période supérieure à 1 (mois), nous verrons plus tard dans l'analyse s'il est possible de croiser ces deux variables.

Nous faisons maintenant à nouveau le test du V de Cramer des variables retenues (n'ayant pas de relations forte entre-elles) avec cette fois-ci la variable de réponse « we18 » afin de choisir les variables qualitatives à conserver.

Les valeurs obtenues semblent raisonnable et cela nous indique qu'il est judicieux de conserver les variables « qual_veh », « csp », « periode », « etat_civil », « mode_habi », « copot_ » (Annexe 6).

3. Quelles variables quantitatives conserver ?

A présent, une fois que nous avons une idée des variables qualitatives candidates à l'intégration dans le modèle, nous nous concentrons sur les variables quantitatives.

a. *Étude de la distribution de la variable quantitative*

Tout d'abord, il est important d'avoir une idée de la loi suivie par la variable quantitative analysée, afin de réaliser le test adéquat. Pour cela, nous pouvons réaliser une première approche avec la procédure UNIVARIATE sous SAS, la table des résultats pour chaque variable quantitative traitée nous donne différents tests de qualité d'ajustement pour une distribution selon une loi normale.

Le premier test de Kolmogorov-Smirnov permet de comparer la fonction de répartition uniforme avec la fonction de répartition de l'échantillon empirique. Le test de Cramér-von Mises est une version plus « puissante » du test de Kolmogorov-Smirnov, ce dernier étant sensible aux *outliers* tandis que le test

de Cramér-von Mises ne l'est pas du tout. Enfin, le test d'Anderson-Darling est un test de normalité de l'échantillon statistique et il permet de détecter l'écart par rapport à la normalité des valeurs minimales et maximales d'une distribution.

Nous retenons simplement que pour notre cas, aucune des variables ne suit une loi normale pour un risque de premier espèce de 5%, en raison de la p-value des trois tests ayant une valeur inférieur à 0,05 (Annexe 7).

b. Test non-paramétrique de Kruskal-Wallis

Étant donné qu'aucune variable ne suit une distribution normale, le test non-paramétrique de Kruskal-Wallis peut convenir. Comme dans chaque test non-paramétrique, le calcul porte sur les rangs attribués suite au classement des valeurs par ordre croissant. On s'affranchit ainsi des conditions de normalité des distributions et d'homogénéité des variances (indispensables à la fiabilité des tests paramétriques). Le test de Kruskal-Wallis nous permet de prouver si, pour chaque échantillon, il existe au moins une moyenne différente des autres.

Ainsi, les résultats obtenus nous permettent de dire qu'on ne conserve pas les variables « no_par_crypte », « rev_men_autr », « mt_men_eng_mena ».

On choisit ensuite de supprimer la variable « no_cnt_crypte », correspondant au numéro du contrat crypté de chaque client, de manière purement arbitraire car il est évident que le numéro du contrat de chaque client ne devrait en aucun cas influencer la probabilité de survenue de l'évènement « 1 », d'autant plus que les modalités ont tous été cryptés par RCI Bank & Services.

Enfin, nous pensons que certaines variables telles que « nb_pers_chg » et « mt_charges » disposent d'un potentiel pour la prédiction dans le modèle, nous choisissons de les mettre de côté pour les recoder plus tard, afin de voir s'il est possible de récupérer les informations de ces deux variables.

Au final, les variables suivantes vont être conservées pour la suite de la construction à la suite du test de Kruskal-Wallis : « duree », « mt_dmd », « pc_appo », « age_veh », « age », « mt_sal_men », « mt_alloc_men », « mt_rev », « mt_loy_men_mena », « mt_men_pre_immo », « mt_ech », « mt_ttc_veh », « part_loyer », « anc_emp », « nb_imp_tot » ainsi que « nb_imp_an_0 ».

4. Traitement des variables candidates

a. Discrétisation des variables continues

Pour les variables que nous avons retenus à la suite du test de Kruskal-Wallis, nous souhaitons dorénavant les discrétiser afin de tester si cette étape pourrait éventuellement améliorer la qualité du modèle. La discrétisation des variables quantitatives implique de changer la nature de celles-ci, en variables catégorielles.

La discrétisation des variables quantitatives est généralement possible selon 2 méthodes, nous allons effectuer les deux techniques afin de s'assurer de la pertinence des variables à conserver : tout d'abord avec la discrétisation des variables sans contrainte sur les déciles et ensuite, sous contrainte des déciles sur la variable de réponse « we18 » égale à « 1 ».

Il est bien entendu évident qu'on ne peut pas créer des classes de manière aléatoire, en réalité, il faut respecter certaines règles lors de la discrétisation des variables. La première est que chaque modalité devra être bien représentée et donc contenir un nombre suffisant d'observation. Nous définissons notre seuil « plancher » dans la construction des groupes de manière à ce que chaque modalité comporte à minima 15% des observations totales. La seconde condition à la discrétisation est que la représentativité des évènements de la variable de réponse doit être plus ou moins équivalente pour chaque modalité.

Pour exemple, la discrétisation[6] de la variable « age » nous permet de créer une variable catégorielle à 5 modalités :

- L'âge est inférieur à 38 ans
- L'âge est entre 38 et 45 ans
- L'âge est entre 45 et 51 ans
- L'âge est entre 51 et 57 ans
- L'âge est supérieur à 57 ans

Il est ensuite important de valider les regroupements effectués, grâce à un macro-programme, nous testons la significativité de chaque variable _cl1 grâce au test du khi-deux, de la même manière que pour le choix des variables qualitatives plus tôt dans notre analyse.

Il en ressort que la variables « mt_men_eng_mena_cl1 » n'est pas significative dans notre modèle, on peut donc la supprimer de la suite de l'analyse.

La deuxième méthode comme évoqué sera d'appliquer des contraintes de déciles sur la variable de réponse. Toujours en respectant la constitution de groupes représentatifs et comportant au moins 15% des observations totales, les modalités construites pour chaque variable restent sensiblement identiques, sauf que les intervalles peuvent varier légèrement d'une méthode à l'autre.

Ainsi, reprenons l'exemple de la discrétisation[7] de la variable « age », appliquée avec la deuxième méthode, nous obtenons la variable catégorielle ordonnée suivante :

- L'âge est inférieur à 36 ans
- L'âge est entre 36 et 45 ans
- L'âge est entre 45 et 51 ans
- L'âge est entre 51 et 59 ans
- L'âge est supérieur à 59 ans

A nouveau, nous devons valider les regroupements effectués et seul la variable « mt_men_eng_mena_cl2 » n'est pas significative (comme lors de la première méthode de discrétisation…). Nous la supprimons également du modèle (Annexe 8).

b. Colinéarité entre les variables qualitatives

Durant cette étape, il faut recommencer ce qui a été fait pour tester la corrélation entre les variables qualitatives dans la section 2.b. à la différence que nous ajoutons dans la liste les variables qui ont été discrétisées (selon les deux méthodes). En revanche, nous précisons que pour vérifier la colinéarité entre ces variables, nous avons logiquement supprimé les doublons de variables (_cl1 et _cl2) ayant le même recodage (comme « duree », « mt_alloc_men », etc…).

On peut noter par exemple que les variables « nb_imp_an_0_cl1 » et « nb_imp_tot_cl2 » sont fortement corrélées avec une valeur pour le V de Cramer de 0,85281. Logiquement puisque l'une regroupe le nombre d'impayés des 12 derniers mois et l'autre, le nombre d'impayés total. Il en est de même pour les variables « qual_veh » avec « age_veh_cl1 », on peut affirmer facilement par exemple qu'un véhicule qualifié d'occasion ne va pas avoir l'âge du véhicule noté comme 0 (mois) (V de Cramer : 0,77520).

En faisant un arbitrage avec le tableau regroupant les valeurs du V de Cramer entre les variables discrétisées et la variable de réponse « we18 », nous pouvons déjà nous séparer des variables suivantes :

[6] Pour simplifier la lecture, les variables discrétisées sans contrainte sur les déciles portent le même nom avec le suffixe _cl1.

[7] Pour simplifier la lecture, les variables discrétisées avec contrainte sur les déciles portent le même nom avec le suffixe _cl2.

« nb_imp_tot_cl1 », « pc_appo_cl2 », « mt_rev_cl2 », « anc_emp_cl2 », « mt_dmd_cl2 », « age_cl1 », « mt_loy_men_mena_cl1 », « mt_sal_men_cl2 », « age_veh_cl2 », « mt_ttc_veh_cl2 » et « mt_charges_cl1 ».

Nous trouvons que les variables « qual_veh » et « periode » peuvent apporter de l'information utile à la variable de réponse, cependant, elles sont trop corrélées pour les conserver les deux. Une solution possible est de les croiser afin d'éliminer cette corrélation. Nous créons les modalités croisées entre chacune des deux variables comme « Véhicule neuf avec période égale à 0 », « Véhicule neuf avec période égale à 1 », … (Annexe 9).

En relançant l'opération de vérification des corrélations, combinée à la mesure du niveau d'implication de chaque variable dans l'explication de la variable de réponse, toujours grâce au V de Cramer, nous décidons à ce stade de supprimer certaines variables trop corrélées à une autre, et celles apportant moins d'information parmi toutes celles mises en cause.

C'est-à-dire que les variables suivantes qui ont été supprimées du modèle avaient une forte corrélation avec une ou plusieurs autres variables et/ou elles étaient parmi celles expliquant le moins « we18 » : « mt_loy_men_mena_cl2 », « nb_imp_tot_cl2 », « copot_ », « mt_men_pre_immo », « mt_sal_men_cl1 », « age_veh_cl1 », « mt_dmd_cl1 », « mt_ttc_veh_cl1 », « mt_ech_cl2 ».

c. Arbitrage entre les variables quantitatives et modifiées

La préoccupation à la suite de la discrétisation des variables quantitatives a été de décider quelle variable conserver (celle initiale, la discrétisée _cl1, la discrétisée _cl2). Nous avons pour cela encore utilisé la mesure statistique du V de Cramer pour mesurer l'intensité de l'indépendance de la variable cible par rapport à la variable explicative. Nous choisissons de privilégier l'interprétabilité plus forte des variables modifiées pour une faible différence au niveau du V de Cramer.

d. Sélection des variables pour le modèle final

Nous avons à présent effectué la grande partie du nettoyage des données pour la construction de notre modèle, nous réalisons une procédure LOGISTIC sous SAS (plus de détails dans la section suivante) principalement pour tester l'importance des variables restantes dans le modèle. Nous appliquons l'option de sélection automatique des variables via la méthode *STEPWISE*. L'algorithme consiste tout d'abord à estimer le modèle avec seulement la constante et une variable, en essayant tous les modèles possibles (constante et une variable explicative), il sélectionne le meilleure modèle au sens du test LM. Ensuite, l'algorithme vérifie s'il peut éliminer des variables à l'aide d'un test Wald et il répète l'opération jusqu'à ce qu'on ne peut plus introduire de variables dans le modèle, ou que lorsqu'une variable rentre dans le modèle, il en ressort immédiatement. Ceci est conditionné selon un critère d'arrêt à définir, dans notre cas, nous le définissons à 5%.

Ainsi, la sélection des variables automatique via la méthode *STEPWISE* nous permet de conserver les variables suivantes : « duree_cl1 », « part_loyer_cl1 », « csp », « etat_civil », « mode_habi », « nb_pers_chg_cl1 », « anc_emp_cl1 », « nb_imp_an_0_cl1 ».

En faisant une dernière vérification au niveau de la corrélation entre les variables restantes, on s'aperçoit que les résultats sont satisfaisants et montrent une corrélation raisonnée entre chacune d'entre-elles. Le niveau de corrélation maximum étant de 0,19182 pour les variables « etat_civil » et « nb_pers_chg_cl1 » (Annexe 10).

L'étape de nettoyage des données et de sélection des variables est donc terminée pour la table PRI. Ainsi, à partir de 38 variables explicatives de départ, il en résulte 8, dont certaines ont été créées à la suite de recodage de variables quantitatives :

Variables conservées pour la grille de score de la table PRI	
Duree_cl1	Durée prévisionnelle du financement
Part_loyer_cl1	Part de l'échéance (%)
Csp	Catégorie socio-professionnelle
Etat_civil	Code état-civil
Mode_habi	Code mode d'habitation
Nb_pers_chg_cl1	Nombre de personnes à charge
Anc_emp_cl1	Ancienneté de l'emploi (en mois)
Nb_imp_an_0_cl1	Nombre d'impayés des 12 derniers mois

C. La table PRO

A partir de maintenant, nous allons décrire les modifications effectuées pour la table PRO, le principe étant la même que pour la table PRI, nous allons moins détailler les significations de chaque test effectués, comme cela a pu être le cas pour la table PRI, afin d'améliorer le confort de lecture.

D'un point de vue général, la structure de la base de données pour la table PRO étant la même que pour celle de PRI, à la différence que nous retranchons uniquement les observations ayant comme modalité « PRO » pour la variable « ty_pp ».

Ci-dessous le tableau regroupant la liste des variables[8] conservées pour la table PRI :

Nom de la variable	Définition de la variable
Ty_pp	Code usage du véhicule
No_cnt_crypte	Identifiant du contrat crypté
No_par_crypte	Identifiant du client crypté
Date_gest	Mois de la date d'entrée en gestion
Dt_dmd	Date de la demande
Duree	Durée prévisionnelle du financement
Genre_veh	Genre du véhicule
Mt_dmd	Montant du financement (cts)
Pc_appo	Pourcentage d'apport (%)
Produit	Type du produit
Qual_veh	Qualité du véhicule
Age_veh	Age du véhicule (en mois)
Ind_cli_rnva	Indicateur client renouvelant
Evpm_copot_pai_glb	Évaluation du comportement de paiement
Evpa_prtc	Indicateur de fichage pour les PRO
Evpm_cote	Évaluation cotation
Cote_bdf	Cotation de la Banque de France
Age	Age du client (en années)
Csp	Catégorie socio-professionnelle
Etat_civil	Code état-civil
Mode_habi	Code mode d'habitation
Mt_rev	Montant revenu mensuel (cts)
Mt_charges	Montant des charges (cts)
Ind_fch_fcc	Indicateur de fichage
Part_loyer	Part de l'échéance (%)
Anc_emp	Ancienneté de l'emploi (en mois)
Copot_	Comportement de paiement
Pan_dir_	Qualité du dirigeant à rembourser
Secteur_	Secteur d'activité
Ind_imp_regu	Nombre d'impayés régularisés (sur les 12 derniers mois)

[8] Les lignes avec l'arrière-plan couleur bleu ciel correspondent aux variables qualitatives, les autres sont les variables quantitatives.

La liste des variables implique désormais des variables propres aux clients professionnels (comme « evpa_prtc » étant l'indicateur de fichage pour les professionnels) et si l'on s'en souvient, des variables comme le montant du salaire mensuel « mt_sal_men », qui était dans la table PRI, a disparu puisque introduire cette variable aurait été inutile, elle aurait disposé de 100% de valeurs manquantes.

Nous avons à disposition 29 variables explicatives (en ne comptant pas « ty_pp ») pour tenter d'expliquer la variable de réponse « we18 ». Celle-ci a la même signification que pour la table PRI : « 1 » si le contrat est en défaut et « 0 » si le contrat est sain.

Également, nous sommes en présence d'un évènement rare : 2,17% de contrats en défaut et 97,83% de contrats sains, sur un total de 36 972 observations.

1. Première modification de la table PRO

Tout d'abord, nous avons jugé qu'il serait sûrement pertinent de réintroduire les variables [9]« periode » et « dt_dmd_month » que nous avons déjà créée en amont pour la table PRI.

Ensuite, le travail est de séparer notre table PRO en deux échantillons, *training sample* et *test sample*, toujours pour nous permettre de réaliser l'apprentissage du modèle sur 70% des données et de pouvoir tester la robustesse de celui-ci sur l'échantillon test (30%).

De la même manière, comme il s'agit également d'un cas d'évènement rare, nous faisant un ré-échantillonnage des « 1 » et des « 0 » dans l'échantillon d'apprentissage pour améliorer la capacité prédictive du modèle construit. Nous répartissons les « 1 » à 10% de l'échantillon d'apprentissage total et donc les 90% restants sont des modalités « 0 » pour la variable de réponse « we18 ».

En effectuant quelques vérifications globales des données, nous avons trouvé quelques *outliers* comme par exemple un pourcentage d'apport pour un prêt supérieur à 100% du montant, ou encore une part du loyer qui est négative. Pour la plupart de ces cas, nous avons pris la décision de remplacer ces valeurs aberrantes par la médiane, et non de supprimer l'observation complète, au risque de perdre de l'information utile.

2. Quels variables qualitatives conserver pour PRO ?

a. Étude de la significativité : test du khi-deux

Le test du khi-deux va nous permettre d'avoir un aperçu global des variables qualitatives à conserver pour la suite du modèle, et au vu des résultats, les premières variables qui peuvent être éliminées sont « dt_dmd_month », « ind_fch_fcc » et « evpa_prtc ». Elles ne sont pas significatives selon le test du khi-deux, pour un risque de première espèce de 5%.

Comme pour le premier cas, il nous faut recoder certaines variables pour pouvoir exploiter l'interprétation du test, en raison du nombre d'occurrences trop faible dans certaines modalités. Les variables concernées sont donc « periode », « csp », « evpm_cote » et « cote_bdf ».

Les 3 premières variables évoqués ont été recodées avec succès et nous a permis d'obtenir des variables significatives pour un risque de premier espèce égal à 5%. Cependant, pour « cote_bdf », nous n'avons malheureusement pas pu trouver de regroupements cohérents possibles, étant donné la signification

- [9] « date_dmd_month » : elle renvoie simplement le mois de la date de demande de chaque observation, on peut observer une certaine saisonnalité des évènements (baisse de la demande de crédit en juillet-août due aux congés d'été).
- « periode » : elle renvoie le nombre de mois écoulé entre la date de demande du crédit par le client jusqu'à la date de l'entrée en gestion du contrat.

claire de chaque modalité, émises par la Banque de France. On ne peut pas généraliser aucune modalité entre-elles, c'est pourquoi nous décidons de nous séparer également de cette variable (Annexe 11 et 12).

b. Corrélation des variables PRO avec le V de Cramer

Rappelons que le V de Cramer permet de mesurer le degré d'association entre deux variables catégorielles et le seuil que nous nous sommes fixés pour inclure ou non les variables dans le modèle est de 0,3.

Les variables les plus corrélées mutuellement sont « evpm_cote » et « pan_dir_ », à 0,84053 pour le V de Cramer (cf. tableau des interprétations des valeurs du V de Cramer, p.9). C'est un résultat logique puisque la première correspond à la cotation de l'entreprise et la seconde, à la capacité du dirigeant à rembourser, ce sont des variables apportant une même information.

Nous regardons ensuite la corrélation entre les variables qualitatives et la variable de réponse « we18 » pour savoir comment gérer les situations de colinéarité.

Il en ressort que, suite à un arbitrage, nous allons supprimer « csp », « pan_dir_ », « produit », « evpm_copot_pai_glb » et « genre_veh » du modèle. Ce choix a été fait puisque ces variables étaient trop corrélées à une autre dans le modèle, et qu'elles n'apportaient pas une information relativement supérieure pour la variable dépendante.

Les variables qualitatives candidates au modèle sont donc les suivantes à ce stade : « periode », « genre_veh », « qual_veh », « ind_cli_rnva », « evpm_cote », « etat_civil », « mode_habi », « copot_ » et « secteur_ ».

3. Quels variables quantitatives conserver ?

Comme pour la partie PRI, nous avons réalisé un test de normalité des distributions des variables quantitatives (cf. tests de Kolmogorov-Smirnov, Cramér-von-Mises et Anderson-Darling). Il en résulte que de même ici, aucune variable ne suit une distribution normale. Le test non paramétrique de Kruskal-Wallis s'applique alors pour tester la significativité des variables quantitatives.

Les variables « mt_dmd » et « mt_charges » ne sont pas significatives, nous ne les conservons pas pour la suite de l'analyse. De plus, et selon le même raisonnement que pour le premier cas, nous supprimons également les variables « no_cnt_crypte » et « no_par_crypte » du modèle de façon arbitraire, puisque conserver le numéro de contrat du client crypté et/ou le numéro d'identifiant du client crypté pour la prédiction d'un évènement semble illogique.

4. Traitement des variables candidates

a. Discrétisation des variables continues de PRO

La discrétisation des variables à la suite du test de Kruskal-Wallis est une étape indispensable pour améliorer le modèle final. Nous ferons comme pour la première partie, c'est-à-dire que nous discrétisons les variables retenues selon les 2 méthodes évoquées.

La discrétisation[10] des variables sans contrainte sur les déciles, en constituant des groupes représentatifs comportant au moins 15% de l'échantillon total. Comme exemple, la discrétisation de la variable « duree » est représentée de la manière suivante :

- La durée du financement est inférieure ou égal à 12 mois
- La durée du financement est comprise entre 12 et 36 mois

[10] Pour simplifier la lecture, les variables discrétisées sans contrainte sur les déciles portent le même nom avec le suffixe _cl1.

- La durée du financement est comprise entre 36 et 48 mois
- La durée du financement est comprise entre 48 et 60 mois
- La durée du financement est supérieure à 60 mois

La procédure pour valider les regroupements effectués nous indique que les variables « mt_dmd_cl1 » et « part_loyer_cl1 » ne sont pas significatives pour un risque de première espèce de 5%. Elles sont donc supprimées du modèle.

La deuxième méthode de discrétisation[11] est de mettre des contraintes de déciles sur la variable de réponse. Le recodage des modalités reste sensiblement identiques, si l'on reprend la variable « duree », nous obtenons :

- La durée du financement est inférieure ou égal à 12 mois
- La durée du financement est comprise entre 12 et 37 mois
- La durée du financement est comprise entre 37 et 49 mois
- La durée du financement est comprise entre 49 et 60 mois
- La durée du financement est supérieure à 60 mois

Les mêmes résultats ressortent lors de la validation du recodage selon la deuxième méthode, ainsi, les variables non significatives comme « mt_dmd_cl2 » et « part_loyer_cl2 » sont définitivement retirées.

Certaines variables ayant subi un codage identique pour la deuxième méthode, par rapport à la première, sont également supprimées : « pc_appo_cl2 », « age_veh_cl2 », « ind_imp_regu_cl2 ».

b. Colinéarité entre les variables qualitatives

Cette sous-section aura pour rôle de vérifier les corrélations entre les variables qualitatives candidates restantes, y compris les variables tout juste discrétisées.

Tout d'abord, nous devons choisir quelles variables discrétisées conserver entre celles issues de la première méthode (_cl1) et la seconde (_cl2) : « anc_emp_cl2 », « mt_charges_cl1 », « mt_rev_cl2 », « age_cl1 » et « duree_cl1 » ne sont pas préférées pour le modèle, à cause de leurs corrélations trop importantes (relativement selon les recodées _cl1 ou _cl2).

Ensuite, nous relançons le code pour regarder en détail la colinéarité entre les variables. Il en ressort que « qual_veh » et « age_veh_cl1 » sont les plus corrélées (V de Cramer : 0,92472). Un véhicule neuf ne va pas avoir une modalité pour « age_veh_cl1 » supérieure à 0.

Les variables « copot_ » et « ind_imp_regu_cl1 » sont également très corrélées entre-elles (0,82536), il représente le comportement de paiement de l'individu (bon, moyen ou mauvais) selon le nombre d'impayés régularisés sur les 12 derniers mois.

Ensuite, pour essayer de garder l'information pouvant être utile, nous allons croiser certaines variables pour éviter la corrélation entre-elles. Nous croisons les variables « qual_veh » et « age_veh_cl1 » et aussi « mode_habi » et « mt_charges_cl2 ».

c. Sélection des variables pour le modèle final (PRO)

A présent, nous avons une idée générale de la structure du modèle, les variables corrélées et les variables importantes pour expliquer « we18 ». De même que pour les variables de la table PRI, nous effectuer une procédure LOGISTIC avec une sélection automatique des variables via la méthode *STEPWISE* : le critère d'arrêt de l'algorithme est fixé à 5%.

[11] Pour simplifier la lecture, les variables discrétisées avec contrainte sur les déciles portent le même nom avec le suffixe _cl2.

L'algorithme nous conseille de garder les variables suivantes pour la prédiction : « ind_cli_rnva », « etat_civil », « mode_habi », « copot_ », « secteur_ », « pc_appo_cl1 », « anc_emp_cl1 », « duree_cl2 » et « age_cl2 » (Annexe 13).

Nous lançons enfin une dernière vérification de la corrélation entre les variables sélectionnées : seules les variables « duree_cl2 » et « secteur_ » restent à un niveau de corrélation important (Annexe 14). L'idée est de supprimer une des deux variables, c'est pourquoi on relance le modèle deux fois, la première est sans « secteur_ » et nous obtenons une statistique C égale à 0,767. Le second modèle sans « duree_cl2 » donne une statistique C de 0,794. Étant donné qu'il s'agit d'une mesure d'analyse de la discrimination d'un test (correspond plus communément à l'aire sous la courbe AUC), plus celle-ci est proche de 1, meilleur le modèle sera. Le choix sera donc de supprimer la variable « duree_cl2 ».

Le modèle final pour la modélisation du score sera composé des variables suivantes :

Variables conservées pour la grille de score de la table PRO	
Ind_cli_rnva	Indicateur client renouvelant
Etat_civil	Code état-civil
Mode_habi	Code mode d'habitation
Copot_	Comportement de paiement
Secteur_	Secteur d'activité
Pc_appo_cl1	Pourcentage d'apport (%)
Anc_emp_cl1	Ancienneté de l'emploi (en mois)
Age_cl2	Age de l'entreprise (année)

IV. Modélisation du score

A présent, une fois que nous avons réussi à obtenir un modèle pour chacune des tables PRI et PRO, nous pouvons lancer la procédure de validation en mesurant les performances de celui-celui sur l'échantillon test que nous avons créée au début pour chacune des tables.

A. Validation du modèle et étude des performances de PRI

Tout d'abord, nous procédons par une régression logistique, permettant d'estimer les valeurs prises par la variable de réponse binaire « we18 » (« 1 » : le contrat tombe en défaut, « 0 » : le contrat est sain). Cette prédiction est réalisée en fonction des caractéristiques personnelles de chaque individu (c'est-à-dire à partir des variables explicatives que nous avons conservé à la suite du tri de la base de données). La régression logistique est faite sur SAS à partir de la procédure LOGISTIC, nous allons préciser quelques points concernant la syntaxe à adopter pour obtenir correctement les estimations :

- « Outest = est ; » : la table de sortie qui va contenir les estimations est appelée « est » et se trouve dans la bibliothèque temporaire WORK.
- « class … ; » : l'instruction permet de mentionner les variables qualitatives du modèles.
- « (ref=…) » : cette option permet de mentionner la modalité de référence de la variable lors de l'estimation du modèle, si on ne spécifie pas cette option, SAS prendra la dernière modalité (dans l'ordre alphabétique) pour en faire la modalité de référence.
- « param=ref ; » : remplace le codage des variables propre à SAS par la variable de référence spécifiée.
- « model … = … ; » : on mentionne ici la variable dépendante ainsi que ses variables explicatives.
- « (event='1') » : il s'agit de spécifier au logiciel que la modalité représentant l'évènement est noté '1'.

(Annexe 15)

Les résultats nous indiquent tout d'abord que le critère de convergence est respecté, ainsi, nous pouvons continuer l'analyse avec la constante et ses covariables (critère d'information minimale). Le modèle est globalement significatif en regardant le tableau du test de l'hypothèse nulle globale, pour un risque de première espèce de 5%. De même que les variables explicatives du modèle sont toutes significatives lorsque l'on se réfère au tableau des effets de type 3 (Annexe 16).

La courbe ROC (Receiver Operating Characteristic) est une mesure de la performance, elle va répertorier la proportion de vrais positifs en fonction de la proportion de faux positifs. Une courbe ROC idéale et utopique va passer par l'angle en haut à gauche du graphique, représentant une AUC (Area Under Curve) de 1. Ainsi, l'AUC mesurant l'aire sous la courbe ROC nous aide à la comparaison des modèles, plus la valeur de celle-ci est proche de 1, mieux le modèle est ajusté. Ici, l'AUC pour l'échantillon d'apprentissage est de 0,8031 et pour l'échantillon de validation, de 0,7848 (Annexe 17).

Nous devons à présent fixer le seuil pour lequel la spécificité ainsi que la sensibilité sera maximisée. Pour rappel, la spécificité correspond à la capacité du modèle à prédire l'évènement « 0 » (non-défaut) alors que le contrat est sain, tandis que la sensibilité est la capacité de prédire du modèle à prédire le défaut (« 1 ») alors que le contrat est bien en défaut. Nous souhaitons donc obtenir un seuil qui permet au modèle de bien prédire les défauts mais aussi de bien prédire les contrats sains. Ce seuil correspond à la valeur à partir duquel on décide que le contrat est considéré comme en défaut. Il est trouvé en calculant la spécificité et la sensibilité pour une multitude de valeurs de seuils différents et notre résultat pour la valeur optimale du seuil, dans le cas de la table PRI, est à 0,08 (Annexe 18).

Une fois ce seuil déterminé, nous pouvons déterminer la matrice de confusion. Cette matrice de confusion va regrouper les valeurs de spécificité, de sensibilité ainsi que les valeurs complémentaires. Elle permet de mesurer la qualité d'un système de classification. En appliquant le seuil de 0,08, nous obtenons une spécificité de 73,15% (« 0 » bien classés) et une sensibilité de 68,21% (« 1 » bien classés). Rappelons que lorsque ces valeurs se rapprochent de 1, la prédiction du modèle est d'autant meilleure (Annexe 19).

Une autre mesure de la performance d'un modèle peut se faire à l'aide de l'indice de Gini. De manière synthétique, celui-ci est utilisé pour évaluer la bonne (ou mauvaise) qualité d'une grille de résultats. Plus il est proche de 1, mieux c'est : dans notre cas, il correspond à 0,56098.

Enfin, la dernière mesure utilisée pour évaluer la régression logistique est l'utilisation de l'indice 10/X. Il représente le pourcentage des défauts identifiés par le modèle, en prenant les 10% de la population les moins bien notés par le score. De même que toutes les autres techniques de mesures évoquées, plus il est proche de 1, mieux c'est. Pour la table PRI, l'indice 10/X est de 0,41059.

B. Construction de la grille de score de PRI

Une fois que nous avons terminé toutes les étapes précédentes, nous pouvons à présent construit le score d'octroi de crédit. La construction de la grille de score consiste à répartir pour chaque modalité de chaque variable un poids (compris entre 0 et 1000 dans notre cas). Ce poids indique l'impact de la modalité sur la probabilité de défaut. Le score d'un individu va consister donc à faire la somme des scores de chaque modalité caractérisée par l'individu. Plus le score est proche de 1000, plus la probabilité de faire défaut pour ce dernier est élevée, et inversement.

Nous rappelons que nous avons construit la grille de score suivante à la suite des estimations de la régression logistique sur la table PRI.

Nous avons utilisé les éléments suivant pour la construction de la grille (Annexe 20) :

- Les estimations des coefficients pour chaque modalité de chaque variable, l'estimation de la modalité mise en référence est égale à 0.

- Le calcul du delta pour chaque modalité, grâce aux estimations. Il s'agit de la différence entre chaque estimation et l'estimation minimale de chaque variable.
- Le calcul du score, grâce aux valeurs des deltas. Il s'agit du delta de chaque modalité divisée par la somme des deltas maximum de chaque variable.

Variables	Modalités	Paramètres estimés	Répartition	Taux de défaut	Delta	Score
Constante	Intercept	-3,47175	.	.	0	0
Anc_emp_cl1 (ancienneté de l'emploi en mois)	Inférieur à 33	0	18,19	2,38	1,11461	126
	Entre 33 et 69	-0,33090	18,08	2,16	0,78371	89
	Entre 69 et 133	-0,10575	21,32	1,73	1,00886	114
	Entre 133 et 253	-0,75714	21,46	1,13	0,35747	41
	Supérieur à 253	-1,11461	20,94	0,76	0	0
Csp (catégorie socio-professionnelle)	Commerçants	1,07408	13,81	2,45	1,07408	122
	Artisans	1,15345	22,93	3,22	1,15345	131
	Agriculteurs	0,83316	5,68	0,93	0,83316	94
	Professions libérales	0	57,57	0,81	0	0
Duree_cl1 (durée prévisionnelle du financement en mois)	Inférieur à 37	-0,63953	30,15	0,35	0	0
	Entre 37 et 49	0	45,20	1,73	0,63953	73
	Entre 49 et 60	0,02433	17,92	2,77	0,66386	75
	Supérieur à 60	0,59878	6,72	3,14	1,23831	140
Etat_civil (code état-civil)	Séparé	0,97488	5,40	1,37	0,97488	111
	Marié	0	52,21	1,17	0	0
	Divorcé	0,73542	9,5	2,34	0,73542	83
	Célibataire	0,69334	23,65	1,97	0,69334	79
	Veuf	0,95828	1,55	0,68	0,95828	109
	Union libre, concubinage ou pacsé	0,22912	7,7	2,74	0,22912	26
Mode_habi (code mode d'habitation)	Propriétaire	0	80,40	1,17	0	0
	Locataire	0,78744	15,06	3,79	0,78744	89
	Hébergé	0,26988	4,14	1,53	0,26988	31
	Logement de fonction	0,94073	0,4	5,26	0,94073	107
Nb_imp_an_0_cl1 (nombre d'impayés des 12 derniers mois)	Égal à 0	-0,45186	37,16	1,05	0	0
	Supérieur à 0	1,39735	3,75	6,48	1,84921	210
	N/A	0	59,09	1,63	0,45186	51
Nb_pers_chg_cl1 (nombre de personnes à charge)	Égal à 0	0	50,56	1,63	0,32921	37
	Égal à 1	-0,32921	19,44	1,52	0	0
	Egal à 2	-0,08887	20,24	1,57	0,24034	27
	Supérieur à 2	0,40399	9,76	1,62	0,73320	83
Part_loyer_cl1	Inférieur à 0,0170	0	54,50	1,12	0	0

(part de l'échéance en %)	Supérieur à 0,0170	0,81248	45,50	2,16	0,81248	92

Une fois que nous avons obtenu la grille de score, nous pouvons désormais calculer le score pour chaque individu souhaitant souscrire à un contrat. Le principe est simplement d'additionner les scores correspondant à chacun des modalités à laquelle l'individu appartient, selon ses caractéristiques personnelles.

En se référant à la grille de score, le profil type de l'individu le plus risqué serait ainsi une personne ayant les caractéristiques suivantes :

- L'ancienneté de l'emploi est inférieure à 33 mois
- Il s'agit d'un artisan
- Souhaitant une durée de financement supérieure à 60 mois
- Séparé
- Disposant d'un logement de fonction
- Ayant au moins un impayé durant les 12 derniers mois
- Ayant au moins deux personnes à charge
- Dont la part de l'échéance représente plus de 1,70% du prix du véhicule

Ces caractéristiques font de l'individu concerné le profil le plus risqué, ayant une probabilité très forte de faire défaut et son score est de 1000/1000.

C. Autres méthodes de Machine Learning pour la table PRI

1. Arbre de décision

Dans le but de concurrencer le modèle de régression logistique, nous allons effectuer deux méthodes d'apprentissage automatique supplémentaire. La première étant les arbres de décision, il s'agit d'une méthode utilisée dans l'exploration de données et en informatique décisionnelle. Ils emploient une représentation hiérarchique de la structure des données sous forme de séquences de décisions (tests) en vue de la prédiction d'un résultat ou d'une classe. On parle de nœuds racine au début, puisque l'accès à l'arbre se fait par ce nœud. Les nœuds internes correspondent aux nœuds ayant une descendance et enfin, les nœuds terminaux sont les nœuds n'ayant pas de descendance. Chaque individu qui doit être attribué à une classe (défaut ou non défaut) est décrit par l'ensemble des variables explicatives qui sont testées dans les nœuds de l'arbre.

Nous obtenons alors une matrice de confusion pour l'arbre de décision sur l'échantillon test, répertoriant 73,66% des « non-défauts » bien prédits par le modèle, et 64,9% des « défauts » bien prédits par le modèle. L'arbre de classification a été réalisé sur 66 nœuds terminaux, car la procédure *HPSPLIT* sur SAS nous indique le nombre de feuilles optimales pour les arbres de classification.

La courbe ROC nous indique une aire sous la courbe (AUC) de 0,82 pour la méthode des arbres de décision, pour la table PRI.

On obtient également un indice de Gini égal 0,6325, ainsi qu'un indice 10/X de 0,27814.

2. Random Forest (forêts aléatoires)

La deuxième méthode de Machine Learning que nous allons introduire pour concurrencer le modèle de régression logistique est la technique de Random Forest (ou forêt aléatoire). Comme son nom l'indique, il s'agit d'un grand nombre d'arbres de décision, agissant comme un ensemble. Chaque arbre

de la forêt aléatoire donne une prédiction de classe, et la classe ayant obtenu le plus de « votes » devient la prédiction du modèle.

Nous constatons pour la méthode du Random Forest une matrice de confusion indiquant un taux de 65,51% de « 0 » bien prédits, et 73,51% de « 1 » bien prédits.

L'AUC utilisé dans le calcul de l'indice de Gini est celui de train, on obtient alors un indice égal à 0,6056.

L'indice 10/X obtenu avec cette méthode s'élève à 0,37086.

D. Validation du modèle et étude des performances de PRO

De la même manière que pour la table PRI (section IV. A.), nous effectuons la régression logistique à partir des variables retenues pour la table PRO, pour la validation du modèle (Annexe 21).

On s'assure encore ici que le critère de convergence pour le modèle est respecté, le modèle est globalement significatif pour un risque de première espèce de 5% et toutes les variables présentes sont significatives pour le même seuil de risque I (Annexe 22).

Retenons à présent seulement les mesures de performances qui nous intéressent : la courbe ROC dispose d'une AUC de 0,7935 pour l'échantillon d'apprentissage et une AUC de 0,7926 pour l'échantillon de validation.

Enfin, le choix du seuil *cut-off* sera de 9%, déterminée d'après le graphique croisant la sensibilité ainsi que la spécificité. Ainsi, nous rappelons que pour la prédiction du modèle, lorsque ce sera au-delà de 9%, nous considérerons l'évènement « 1 » comme celle prédite par le modèle.

La matrice de confusion déterminée indique une spécificité de 0,7191 (« 0 » bien classés) et une sensibilité de 0,6958 (« 1 » bien classés) (Annexe 23).

L'indice de Gini calculée renvoie la valeur 0,58530 pour le modèle de régression logistique.

L'indice 10/X est de 0,40833 pour le modèle.

E. Construction de la grille de score de PRO

La construction de la grille de score consiste à répéter l'opération de la section IV. B. pour la table PRO cette fois-ci. Nous choisissons à nouveau de construire un score allant de 0 à 1000.

Variables	Modalités	Paramètres estimés	Répartition	Taux de défaut	Delta	Score
Constante	Intercept	-3,4181	.	.	0	0
Anc_emp_cl1 (ancienneté de l'emploi en mois)	Inférieur à 38	0	22,08	3,43	1,17633	111
	Entre 38 et 88	-0,45443	21,23	2,97	0,7219	68
	Entre 88 et 144	-0,77871	17	1,86	0,39762	38
	Entre 144 et 243	-1,17633	19,68	1,47	0	0
	Supérieur à 243	-1,03097	20,01	0,17	0,14536	14
Etat_civil (code état-civil)	Séparé	0,37058	7,65	1,69	0,65665	62
	Marié	0	40,38	1,6	0,28607	27
	Divorcé	0,31183	8,97	2,06	0,59790	57
	Célibataire	0,32297	39,02	2,72	0,60904	58

	Veuf	-0,28607	1,36	2,04	0	0
	Union libre, concubinage ou pacsé	-0,16010	2,62	1,76	0,12597	12
Mode_habi (code mode d'habitation)	Propriétaire	0	85,26	1,6	0	0
	Locataire	0,92380	12,06	4,66	0,92380	88
	Hébergé	0,32238	2,48	6,04	0,32238	31
	Logement de fonction	0,50241	0,21	0	0,50241	48
Copot_ (comportement de paiement)	Bon	0	92,63	1,72	0	0
	Moyen	0,88512	5,17	5,41	0,88512	84
	Mauvais	2,33682	2,2	13,11	2,33682	221
Pc_appo_cl1 (part de l'apport)	Égal à 0	0	67,02	2,3	1,03645	98
	Entre 0 et 0,2	-0,30535	17,44	2,79	0,73110	69
	Supérieur à 0,2	-1,03645	15,55	0,87	0	0
Age_cl2 (âge de l'entreprise)	Inférieur à 38	0	21,51	3,1	0,63266	60
	Entre 38 et 43	-0,13176	14,56	2,66	0,50090	47
	Entre 43 et 49	-0,27250	20,16	2,15	0,36016	34
	Entre 49 et 57	-0,42667	26,68	1,76	0,20599	20
	Supérieur à 57	-0,63266	17,09	1,21	0	0
Secteur_ (secteur d'activité)	Réparateur automobile	1,34567	1,22	2,22	1,34567	127
	Métallurgie	1,12465	0,17	0	1,12465	107
	Immobilier	1,43408	2,02	1,79	1,43408	136
	Hôtellerie, Restauration	2,05388	2,6	4,51	2,05388	195
	Hôpitaux	0	26,13	0,55	0	0
	Industrie	1,87155	0,78	2,3	1,87155	177
	Fabrication Chimie Pharmacie	2,86145	0,17	0	2,86145	271
	Biens de consommation	1,77951	2,29	5,51	1,77951	169
	Enseignement et Auto-école	1,38520	22,77	2,1	1,38520	131
	Commerce de groupe	2,22961	1,46	0,04	2,22961	211
	Commerce de détail	1,93576	5,98	1,81	1,93576	183
	Bâtiment et Travaux public	1,93509	15,51	3,84	1,93509	183
	Agriculture	2,09787	4,21	1,07	2,09787	199
	Transport	2,10839	3,08	5,85	2,10839	200
	Services	1,68833	6,21	2,32	1,68833	160
	Autres	1,92284	5,39	2,01	1,92284	182
Ind_cli_rnva (indicateur client renouvelant)	Oui	0,94333	0,95	4,76	0,94333	89
	Non	0	99,05	2,14	0	0

Cette grille de score nous permet de calculer le score de chacun des individus catégorisée « PRO », en additionnant le score de chacune des modalités concernées par le client.

Nous établissons à nouveau le profil type de l'individu le plus risqué parmi les PRO :

- L'ancienneté de l'emploi est inférieure à 38 mois
- Situation maritale : séparé
- Locataire
- Catégorisé comme ayant un mauvais comportement de paiement
- Ayant un contrat avec 0% d'apport
- L'âge de l'entreprise est inférieur à 38 ans
- Dans le secteur de la location
- Étant un client renouvelant

F. Autres méthodes de Machine Learning pour la table PRO

1. Arbre de décision

La procédure *HPSPLIT* sur SAS nous indique le nombre de feuilles optimales pour les arbres de classification : 62 nœuds terminaux.

La matrice de confusion indique pour l'échantillon test 74,3% de « 0 » bien prédits et 63,33% de « 1 » bien prédits.

L'AUC (Area Under Curve) est de 0,78 pour la technique des arbres de classification.

Les indices obtenus sont assez proches de ceux trouvés avec la régression logistique, notamment pour l'indice de Gini qui est égale à 0,5577.

L'indice 10/X est de 0,3083.

2. Random Forest

La forêt nous donne une prédiction pour chaque individu :

La matrice de confusion indiquant un taux de 67,5% de « 0 » bien prédits, et 75,83% de « 1 » bien prédits.

L'indice de Gini a été calculé via l'AUC de l'échantillon d'apprentissage est équivalent à 0,5656.

Cette méthode nous donne l'indice 10/X le plus important, avec un indice de 0,42916.

V. Conclusion

Tableau récapitulatif des résultats obtenus pour PRI :

Méthode & Mesure	Régression logistique	Arbre de décision	Random Forest
Spécificité	0,7315	0,7366	0,6551
Sensibilité	0,6821	0,649	0,7351
Indice de Gini	0,5698	0.6325	0,6056
Indice 10/X	0,41059	0,27814	0,37086

Tableau récapitulatif des résultats obtenus pour PRO :

Méthode & Mesure	Régression logistique	Arbre de décision	Random Forest
Spécificité	0,72	0,743	0,6750
Sensibilité	0,7	0,6333	0,7583
Indice de Gini	0,58443	0.5577	0,5656
Indice 10/X	0,40833	0.3083	0.42916

La conclusion qu'on en retire, c'est que peu importe la table, PRI ou PRO, l'Area Under Curve (AUC) est aux alentours de 0,80 pour la méthode de régression logistique.

Aussi, la spécificité semble meilleure à chaque fois pour la technique avec les arbres de décision tandis que l'avantage d'une bonne sensibilité revient aux forêts aléatoires.

Les autres indicateurs de comparaisons sont très proches et dans notre cas, ils varient selon la table utilisée. Par exemple, l'indice 10/X est meilleur avec la régression logistique pour la table PRI tandis que c'est plutôt la technique des forêts aléatoires concernant la table PRO.

On peut ainsi dire que les indicateurs de performances d'un modèle dépendent beaucoup du tri effectué sur la table, de la méthode employée pour la prédiction et également d'autres facteurs externes et aléatoires. D'autres méthodes non présentées ici peuvent s'avérer plus précises, mais on perdrait en interprétabilité.

Pour conclure, les grilles de scores obtenus pour PRI ainsi que PRO semblent assez réaliste pour chaque variable en jeu.

Nous pensons donc avoir établi une grille de score plutôt correcte et proche de la réalité, grâce aux données fournies par RCI Bank & Services. Un des axes d'améliorations de cette étude serait de se renseigner sur la préférence de la banque vis-à-vis de la spécificité ou bien la sensibilité puisque chaque modèle semble avoir ses points forts.

VI. Bibliographie

DUMITRESCU Elena, HUE Sullivan, HURLIN Christophe, TOKPAVI Sessi. Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects. Février 2018. [19 novembre].
https://www.researchgate.net/publication/318661593_Machine_Learning_for_Credit_Scoring_Improving_Logistic_Regression_with_Non_Linear_Decision_Tree_Effects

Baha-Eddine Aouni Carminda Cid Mael Jauniaux (2008) – Technique de scoring
http://carmindacid.free.fr/fichiers/projet_scoring.pdf

GONZALEZ, Pierre-Louis (2018) - CALCUL D'UN SCORE (SCORING) Application de techniques de discrimination. maths.cnam..
http://maths.cnam.fr/IMG/pdf/Scoring-2013_cle09f1ab.pdf

Carl Nord, Jacob Keeley, Grand Valley State University, Grand Rapids, MI (2016) - An Introduction to the HPFOREST Procedure and its Options
https://www.mwsug.org/proceedings/2016/AA/MWSUG-2016-AA20.pdf

C.RAULT (2019) – Cours de scoring Master ESA

SAS HELP - The HPSPLIT Procedure
http://support.sas.com/documentation/cdl/en/stathpug/68163/HTML/default/viewer.htm#stathpug_hpsplit_overview01.htm

VII. Annexes

Annexe 1 : Procédure SURVEYSELECT pour partitionner les données

```
PROC SURVEYSELECT data=pri
                  outall
                  samprate=.70                  /* Proportion d'observation sélectionnée dans l'échantillon d'apprentissage */
                  out = pri(drop=selectionProb SamplingWeight)
                  method = srs                  /* Sélection équiprobable et sans remise  */
                  seed = 435                    /* Observation de départ puis sélection aléatoire */
                  NOPRINT;
  strata we18;                                  /* Echantillon stratifié par rapport à la REPONSE */
RUN;
```

Annexe 2 : Recodage des variables secteur_ et periode

```
data train (drop=secteur_ periode);
 set train;

   length secteur $10.;
    if secteur_="AGR" then secteur="Primaire";
    if secteur_="BTP" or secteur_="FCP" or secteur_="FEM" then secteur="Secondaire";
    if secteur_="CDD" or secteur_="CDG" or secteur_="EAE"
     or secteur_="FBC" or secteur_="HOP" or secteur_="LOA" or secteur_="SCE" or secteur_="TRA" then secteur="Tertiaire";
    if secteur_="ATR" then secteur="Autres";

   length periodes $4.;
    if periode="0" then periodes="0";
    if periode="1" then periodes="1";
    if periode="2" or periode="3" or periode="4" or periode="5"
    or periode="6" or periode="7" or periode="8" or periode="25" then periodes=">=2";

run;
```

Annexe 3 : Exemple du test de khi-deux entre secteur et we18

Statistiques pour la table de secteur par WE18

Statistique	DDL	Valeur	Prob
Khi-2	3	71.7442	<.0001
Test du rapport de vraisemblance	3	58.4687	<.0001
Khi-2 de Mantel-Haenszel	1	4.4385	0.0351
Coefficient Phi		0.1359	
Coefficient de contingence		0.1347	
V de Cramer		0.1359	

Taille de l'échantillon = 3885

Annexe 4 : Tableau récapitulatif des corrélations entre les variables

Obs.	Table	abs_V_Cramer
1	Table QUAL_VEH * QUAL_VEH	1.00000
2	Table CSP * CSP	1.00000
3	Table copot_ * copot_	1.00000
4	Table secteur * secteur	1.00000
5	Table periodes * periodes	1.00000
6	Table ETAT_CIVIL * ETAT_CIVIL	1.00000
7	Table MODE_HABI * MODE_HABI	1.00000
8	Table secteur * CSP	0.46619
9	Table CSP * secteur	0.46619
10	Table periodes * QUAL_VEH	0.33999
11	Table QUAL_VEH * periodes	0.33999
12	Table ETAT_CIVIL * MODE_HABI	0.18304
13	Table MODE_HABI * ETAT_CIVIL	0.18304
14	Table CSP * QUAL_VEH	0.17905
15	Table QUAL_VEH * CSP	0.17905
16	Table secteur * QUAL_VEH	0.13776
17	Table QUAL_VEH * secteur	0.13776
18	Table ETAT_CIVIL * CSP	0.09881
19	Table CSP * ETAT_CIVIL	0.09881
20	Table periodes * CSP	0.09197
21	Table CSP * periodes	0.09197
22	Table QUAL_VEH * ETAT_CIVIL	0.08805
23	Table ETAT_CIVIL * QUAL_VEH	0.08805
24	Table CSP * MODE_HABI	0.07722
25	Table MODE_HABI * CSP	0.07722
26	Table ETAT_CIVIL * secteur	0.07090
27	Table secteur * ETAT_CIVIL	0.07090
28	Table periodes * secteur	0.06655
29	Table secteur * periodes	0.06655
30	Table ETAT_CIVIL * copot_	0.05506
31	Table copot_ * ETAT_CIVIL	0.05506
32	Table MODE_HABI * QUAL_VEH	0.05426
33	Table QUAL_VEH * MODE_HABI	0.05426
34	Table copot_ * CSP	0.05234
35	Table CSP * copot_	0.05234
36	Table MODE_HABI * secteur	0.04952
37	Table secteur * MODE_HABI	0.04952
38	Table secteur * copot_	0.04826
39	Table copot_ * secteur	0.04826
40	Table ETAT_CIVIL * periodes	0.04173
41	Table periodes * ETAT_CIVIL	0.04173
42	Table copot_ * MODE_HABI	0.03777

Annexe 5 : Fréquences csp par secteur

La procédure FREQ

Fréquence Pourcentage	Table de CSP par secteur				
CSP(Classe socio-professionnelle)	secteur				
	Autres	Primaire	Secondaire	Tertiaire	Total
Agriculteurs	54 1.39	113 2.91	2 0.05	40 1.03	209 5.38
Artisans	330 8.49	16 0.41	302 7.77	331 8.52	979 25.20
Commerçants	295 7.59	8 0.21	6 0.15	251 6.46	560 14.41
Professions libérales	546 14.05	6 0.15	52 1.34	1533 39.46	2137 55.01
Total	1225 31.53	143 3.68	362 9.32	2155 55.47	3885 100.00

Annexe 6 : V de Cramer entre les variables qualitatives et we18

Obs.	Table	abs_V_Cramer
1	Table copot_ * WE18	0.17906
2	Table MODE_HABI * WE18	0.16721
3	Table CSP * WE18	0.16477
4	Table secteur * WE18	0.13589
5	Table ETAT_CIVIL * WE18	0.10794
6	Table QUAL_VEH * WE18	0.06153
7	Table periodes * WE18	0.04139

Annexe 7 : Exemple de test de normalité pour mt_ttc_veh

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistique		p-value	
Kolmogorov-Smirnov	D	0.047419	Pr > D	<0.010
Cramer-von Mises	W-Sq	27.455255	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	179.047410	Pr > A-Sq	<0.005

Annexe 8 : Regroupement non significatif pour mt_men_eng_mena_cl2

Statistiques pour la table de mt_men_eng_mena_cl2 par WE18

Statistique	DDL	Valeur	Prob
Khi-2	1	0.1992	0.6553
Test du rapport de vraisemblance	1	0.1948	0.6590
Khi-2 continuité ajustée	1	0.1125	0.7373
Khi-2 de Mantel-Haenszel	1	0.1992	0.6554
Coefficient Phi		-0.0072	
Coefficient de contingence		0.0072	
V de Cramer		-0.0072	

Annexe 9 : Recodage de qual_veh et periodes en qvperiode

```
data train;
set train;

 length qvperiode $ 13.;
  if qual_veh="VN" and periodes="0" then qvperiode="VN periode 0";
  if qual_veh="VN" and periodes="1" then qvperiode="VN periode 1";
  if qual_veh="VN" and periodes=">=2" then qvperiode="VN periode>=2";
  if qual_veh="VO" and periodes="0" then qvperiode="VO periode 0";
  if qual_veh="VO" and periodes="1" then qvperiode="VO periode 1";
  if qual_veh="VO" and periodes=">=2" then qvperiode="VO periode>=2";

run;
```

Annexe 10 : Test des corrélations des variables restantes (PRI)

Obs.	Variable1	Variable2	abs_V_Cramer
1	ETAT_CIVIL	nb_pers_chg_cl1	0.19182
2	ETAT_CIVIL	MODE_HABI	0.18304
3	duree_cl1	part_loyer_cl1	0.18232
4	MODE_HABI	anc_emp_cl1	0.17290
5	CSP	anc_emp_cl1	0.13633
6	CSP	duree_cl1	0.12550
7	duree_cl1	nb_imp_an_0_cl1	0.10732
8	anc_emp_cl1	nb_pers_chg_cl1	0.10649
9	nb_imp_an_0_cl1	part_loyer_cl1	0.10323
10	ETAT_CIVIL	anc_emp_cl1	0.10310
11	CSP	ETAT_CIVIL	0.09881
12	anc_emp_cl1	nb_imp_an_0_cl1	0.09597
13	ETAT_CIVIL	nb_imp_an_0_cl1	0.08680
14	MODE_HABI	nb_imp_an_0_cl1	0.08641
15	MODE_HABI	nb_pers_chg_cl1	0.08225
16	anc_emp_cl1	duree_cl1	0.07968
17	CSP	MODE_HABI	0.07722
18	MODE_HABI	duree_cl1	0.07238
19	CSP	nb_imp_an_0_cl1	0.06354
20	CSP	part_loyer_cl1	0.05735
21	ETAT_CIVIL	part_loyer_cl1	0.05645
22	duree_cl1	nb_pers_chg_cl1	0.04888
23	CSP	nb_pers_chg_cl1	0.04486
24	ETAT_CIVIL	duree_cl1	0.04164
25	MODE_HABI	part_loyer_cl1	0.04137
26	anc_emp_cl1	part_loyer_cl1	0.03644
27	nb_imp_an_0_cl1	nb_pers_chg_cl1	0.02549
28	nb_pers_chg_cl1	part_loyer_cl1	0.00686

Annexe 11 : Tableau des fréquences de la variable cote_bdf

Cotation de la Banque de France				
COTE_BDF	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
G3	3	0.05	3	0.05
G4	1	0.02	4	0.06
G4+	2	0.03	6	0.10
G5	3	0.05	9	0.15
G5+	5	0.08	14	0.23
G6	1	0.02	15	0.24
H3	1	0.02	16	0.26
H4	5	0.08	21	0.34
H5+	8	0.13	29	0.47
H6	1	0.02	30	0.48
K0	1	0.02	31	0.50
K5+	2	0.03	33	0.53
X0	3882	62.62	3915	63.16
X5	5	0.08	3920	63.24
X6	7	0.11	3927	63.35
X7	17	0.27	3944	63.62
X8	1	0.02	3945	63.64
XP	2	0.03	3947	63.67
ZZZZZZ	2252	36.33	6199	100.00

Annexe 12 : Test du khi-deux incompatible pour la variable cote_bdf

Statistiques pour la table de COTE_BDF par WE18

Statistique	DDL	Valeur	Prob
Khi-2	18	27.4705	0.0706
Test du rapport de vraisemblance	18	21.6556	0.2476
Khi-2 de Mantel-Haenszel	1	5.1284	0.0235
Coefficient Phi		0.0666	
Coefficient de contingence		0.0664	
V de Cramer		0.0666	
WARNING: 82% des cellules ont un effectif théorique inférieur à 5. Le test du Khi-2 peut ne pas convenir.			

Taille de l'échantillon = 6199

Annexe 13 : Récapitulatif de la sélection STEPWISE pour PRO

Récapitulatif sur la sélection Stepwise								
	Effet							
Etape	Saisi	Supprimé	DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > khi-2	Libellé de la variable
1	copot_		2	1	254.3224		<.0001	Comportement de paiement
2	anc_emp_cl1		4	2	141.7123		<.0001	
3	secteur		15	3	122.6323		<.0001	
4	MODE_HABI		3	4	77.6472		<.0001	Code mode d'habitation
5	pc_appo_cl1		2	5	36.0535		<.0001	
6	IND_CLI_RNVA		1	6	9.6354		0.0019	Indicateur client renouvelant
7	age_cl2		4	7	14.9129		0.0049	
8	duree_cl2		4	8	13.4597		0.0092	
9	ETAT_CIVIL		5	9	12.2333		0.0317	Code état civil

Annexe 14 : Corrélation entre les variables restantes

Obs.	Variable1	Variable2	abs_V_Cramer
1	duree_cl2	secteur	0.49012
2	age_cl2	anc_emp_cl1	0.29475
3	pc_appo_cl1	secteur	0.29278
4	duree_cl2	pc_appo_cl1	0.27031
5	ETAT_CIVIL	age_cl2	0.16098
6	MODE_HABI	age_cl2	0.15150
7	MODE_HABI	anc_emp_cl1	0.15080
8	copot_	secteur	0.11745
9	IND_CLI_RNVA	secteur	0.11675
10	IND_CLI_RNVA	copot_	0.11338
11	age_cl2	secteur	0.10856
12	ETAT_CIVIL	MODE_HABI	0.10762
13	anc_emp_cl1	secteur	0.09958
14	ETAT_CIVIL	anc_emp_cl1	0.09631
15	MODE_HABI	secteur	0.08607
16	ETAT_CIVIL	secteur	0.08598
17	IND_CLI_RNVA	duree_cl2	0.08563
18	age_cl2	duree_cl2	0.08494
19	copot_	duree_cl2	0.07540
20	anc_emp_cl1	duree_cl2	0.07275
21	ETAT_CIVIL	duree_cl2	0.06608
22	IND_CLI_RNVA	age_cl2	0.06061
23	copot_	pc_appo_cl1	0.05906
24	age_cl2	pc_appo_cl1	0.05800
25	MODE_HABI	duree_cl2	0.05744
26	anc_emp_cl1	pc_appo_cl1	0.05643
27	IND_CLI_RNVA	anc_emp_cl1	0.05608
28	MODE_HABI	pc_appo_cl1	0.04889
29	anc_emp_cl1	copot_	0.04680
30	IND_CLI_RNVA	pc_appo_cl1	0.04535
31	age_cl2	copot_	0.04527
32	IND_CLI_RNVA	MODE_HABI	0.04454
33	ETAT_CIVIL	pc_appo_cl1	0.04420
34	ETAT_CIVIL	IND_CLI_RNVA	0.04133
35	ETAT_CIVIL	copot_	0.03735
36	MODE_HABI	copot_	0.01924

Annexe 15 : Code pour la régression logistique PRI

```
proc logistic data=train outest=est ;
    class Duree_cl1 (ref="entre 37 et 49") part_loyer_cl1 (ref="inférieur à 0.0170") csp (ref="Professions libérales") etat_civil (ref="Marié")
          mode_habi (ref="Propriétaire") nb_pers_chg_cl1 (ref="égal à 0") anc_emp_cl1 (ref="inférieur à 33") nb_imp_an_0_cl1 (ref="N/A") / param=ref;
    model wel8(event='1') = Duree_cl1 part_loyer_cl1 csp etat_civil mode_habi nb_pers_chg_cl1 anc_emp_cl1 nb_imp_an_0_cl1
    / rsquare lackfit ctable outroc=testroc rsquare;
    score data=test out=test0 outroc=roc0 fitstat;
run;
```

Annexe 16: Validation du modèle PRI

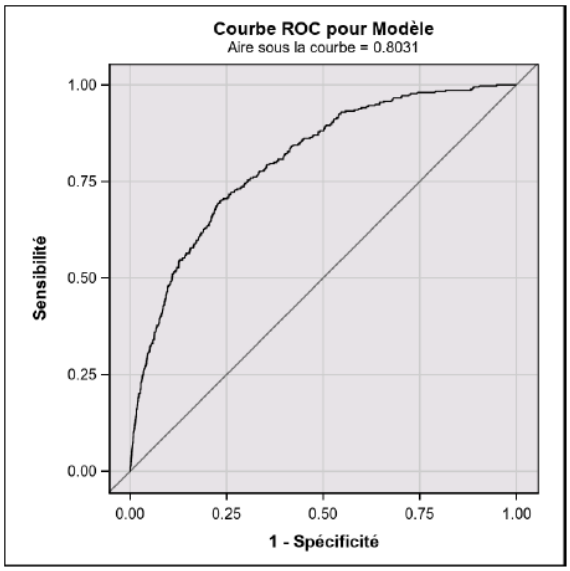
Etat de convergence du modèle
Critère de convergence (GCONV=1E-8) respecté.

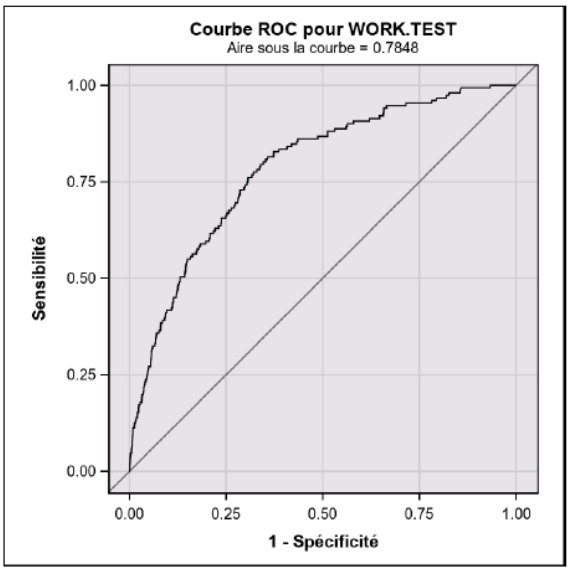
Statistique d'ajustement du modèle		
Critère	Constante uniquement	Constante et Covariables
AIC	2367.995	1997.504
SC	2374.259	2154.119
-2 Log L	2365.995	1947.504

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	418.4909	24	<.0001
Score	443.9275	24	<.0001
Wald	337.9894	24	<.0001

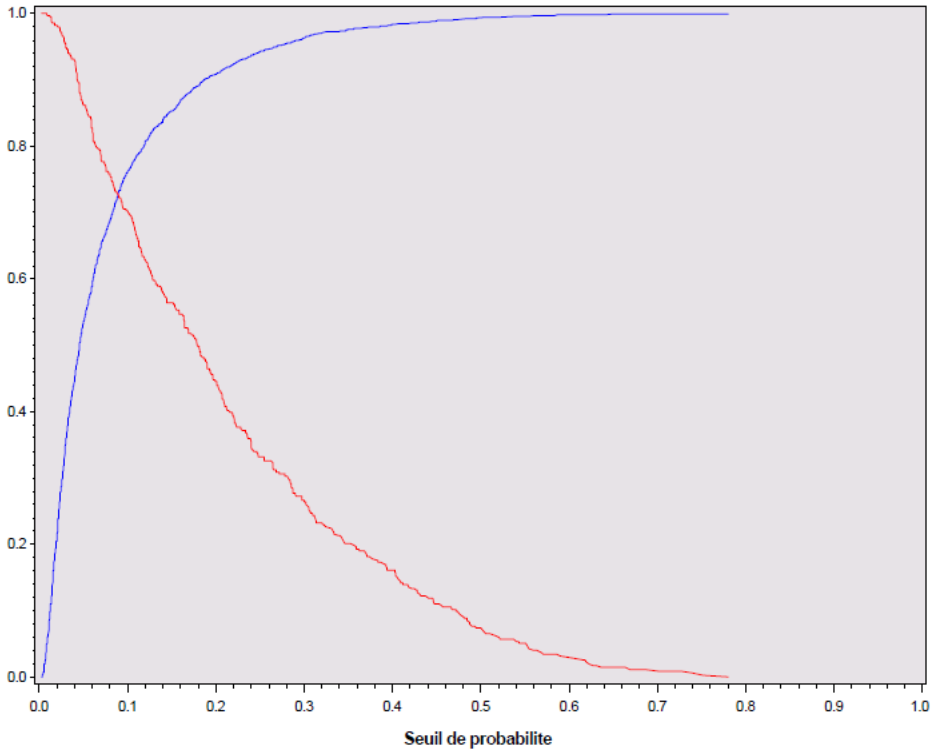
Analyse des effets Type 3			
Effet	DDL	Khi-2 de Wald	Pr > khi-2
duree_cl1	3	29.4548	<.0001
part_loyer_cl1	1	40.2037	<.0001
CSP	3	70.4388	<.0001
ETAT_CIVIL	5	30.0268	<.0001
MODE_HABI	3	32.1678	<.0001
nb_pers_chg_cl1	3	9.3985	0.0244
anc_emp_cl1	4	29.5742	<.0001
nb_imp_an_0_cl1	2	72.5317	<.0001

Annexe 17 : Courbes ROC apprentissage (gauche) et test (droite)





Annexe 18 : *Cut-off* optimal PRI



Annexe 19 : Matrice de confusion de la table PRI, pour la régression logistique

Fréquence Pct de ligne	Table de WE18 par I_cible		
WE18(Indicateur de défaut)	I_cible		
	0	1	Total
0	6444 69.16	2873 30.84	9317
1	36 23.84	115 76.16	151
Total	6480	2988	9468

Annexe 20 : Méthode pour construire les *deltas* et *deltas totaux* pour la grille de score PRI

```
/* Production des minimums des estimations */

proc means data=ref min noprint;
     class variables;
     var estimation;
     output out=mini min=minimum;
run;

proc sql;
create table ref2 as(
select Variables, Description, Moda, Repartition, Taux_defaut, Estimation, estimation-minimum as delta
from ref natural left join mini);
quit;


/* Production du delta total */

proc means data=ref2 max noprint;
     class variables;
     var delta;
     output out=maximum max=max;
run;

proc means data=maximum sum noprint;
     var max;
     where _Type_=1;
     output out= delta_tot sum=somme;
run;

data _null_;
     set delta_tot;
     call symputx ("delta_tot",somme);
run;

%put &delta_tot;

/* delta total = 8.81687 */

proc sql;
create table ref3 as(
select *, round(1000*(delta/&delta_tot.)) as score
from ref2);
quit;
```

Annexe 21 : Procédure LOGISTIC pour estimer le modèle issu de la table PRO

```
proc logistic data=train outest=est ;
     class ind_cli_rnva (ref="Non") etat_civil (ref="Marié") mode_habi (ref="Propriétaire")
           copot_ (ref="Bon") secteur_ (ref="Hôpitaux") pc_appo_cl1 (ref="égal à 0")
           anc_emp_cl1 (ref="inférieur à 38") age_cl2 (ref="inférieur à 38") / param=ref;
     model wel8(event='1') = ind_cli_rnva etat_civil mode_habi copot_ secteur_ pc_appo_cl1
           anc_emp_cl1 age_cl2 / rsquare lackfit ctable outroc=testroc rsquare;
     score data=test out=test0 outroc=roc0 fitstat;
run;
```

Annexe 22 : Validation du modèle pour PRO

Etat de convergence du modèle
Critère de convergence (GCONV=1E-8) respecté.

Statistique d'ajustement du modèle		
Critère	Constante uniquement	Constante et Covariables
AIC	3628.675	3105.422
SC	3635.371	3353.181
-2 Log L	3626.675	3031.422

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	595.2525	36	<.0001
Score	640.5681	36	<.0001
Wald	475.0500	36	<.0001

Analyse des effets Type 3			
Effet	DDL	Khi-2 de Wald	Pr > khi-2
IND_CLI_RNVA	1	10.1782	0.0014
ETAT_CIVIL	5	11.6087	0.0406
MODE_HABI	3	59.8201	<.0001
copot_	2	176.3003	<.0001
secteur	15	123.2181	<.0001
pc_appo_cl1	2	33.6434	<.0001
anc_emp_cl1	4	56.2253	<.0001
age_cl2	4	11.7496	0.0193

Annexe 23 : Matrice de confusion (reg. log.) PRO

Fréquence Pct de ligne	Table de WE18 par I_cible		
WE18(Indicateur de défaut)	I_cible		
	0	1	Total
0	7802 71.91	3048 28.09	10850
1	73 30.42	167 69.58	240
Total	7875	3215	11090

FIN