



المندوبية السامية للتخطيط
HAUT-COMMISSARIAT AU PLAN

ROYAUME DU MAROC
._._*._*
HAUT COMMISSARIAT AU PLAN
._._*._*._*._*
INSTITUT NATIONAL
DE STATISTIQUE ET D'ECONOMIE APPLIQUEE



INSEA

Projet de Fin d'Etudes

Etude de rentabilité des tarifs de AXA Crédit

Préparé par : *M. Mehdi EL HAOUS*

Sous la direction de : *M. Driss EFFINA (INSEA)*
Mme Mariam BENZINEB (AXA Assurance Maroc)

Soutenu publiquement comme exigence partielle en vue de l'obtention du

Diplôme d'Ingénieur d'Etat

Filière : Actuariat et Finance

Devant le jury composé de :

- *M. Driss EFFINA (INSEA)*
- *M. khalil Mohammed SAID (INSEA)*
- *Mme Mariam BENZINEB (AXA Assurance Maroc)*

Résumé

L'activité d'inter-médiation bancaire repose sur deux composantes : la collecte des excédents de trésorerie et l'octroi de prêts. Les organisations opérationnelles ou financières distinguent généralement ces deux composantes. Ce travail se focalise sur l'octroi de prêts, en particulier sur le crédit à la consommation. Il porte sur la modélisation statistique de la probabilité de défaut des clients, et son intégration dans le calcul dans un nouveau tarif ,pour une marge d'intérêt fixée . Le modèle permet d'affecter des scores basés sur les facteurs socio-économiques des emprunteurs(Le type de client,l'état matrimonial , le capital emprunté ,la durée de remboursement, etc). L'analyse montre une incidence de chacune de ces caractéristiques sur la marge d'intérêt dégagée.

En raison de confidentialité, nous cachons certains chiffres.

Dédicace

*A ma chère mère qui ne m'ai jamais épargné un effort
pour m'aider et m'encourager.*

*Veillez trouver en ce travail la consolation et le témoin de
la patience et d'amour.*

*A mon cher père qui a été toujours près de moi, pour
m'écouter et me soutenir.*

Puisse ce travail exprimer le respect et l'amour que je vous porte.

A ma chère petite sœur, à qui je souhaite tout le bonheur du monde.

Vous avez toujours été d'une aide très précieuse.

A mes amis :Rafik, Mouad, Ayoub, Hanane, Kaoutar et Amal

*Je ne saurai terminer sans exprimer toute mon gratitude et mon respect le plus profond
à mon établissement et à mes chers professeurs de l'Institut National de Statistique et
d'Économie Appliquée*

ELHAOUS Mehdi

Remerciements

Qu'il m'ait permis, au terme de ce travail, d'exprimer ma gratitude et vifs remerciements à mon encadrante de stage de fin d'étude au sein de AXA Assurance Maroc, Mme. Mariam BENZINEB. Qu'elle trouve ici le témoignage de mon estime et de mon éternelle reconnaissance pour son disponibilité, ses conseils et son compétence qu'elle a su me prodiguer tout au long de mon stage malgré ses occupations extrêmes. Et son soutien qui m'a été précieux afin de mener mon travail à bon port.

Je tiens également à exprimer ma gratitude envers tout le personnel de AXA Assurance Maroc, notamment Mr. Abdelmounaim BJIJOU. Je tiens à remercier l'équipe de la Direction Risk Management pour leur hospitalité et pour l'esprit de service qu'ils ont eu à mon égard.

Une gratitude toute particulière revient à mon professeur à l'INSEA, Mr. Driss EFFINA, je tiens à le remercier sincèrement pour son aide, ses encouragements, ses conseils ainsi que ses précieuses remarques qui m'a grandement contribué à améliorer la qualité de ce mémoire.

Mes gratitude et estime vont finalement au corps professoral de l'Institut National de Statistique et d'Économie Appliquée qui veille à m'assurer une formation de valeur. Ma reconnaissance va à tous ceux qui, de près ou de loin, ont contribué à l'aboutissement et au bon déroulement de ce modeste travail.

Liste des abréviations

AAM : AXA Assurance Maroc

AUC : Area under curve

LOA : Location avec option d'achat

CMR : Caisse Marocaine des retraites

CRD : Capital restant dû

PD : Probabilité de défaut

ROC : Receiver operating characteristic

TMIC : Taux maximum des intérêts conventionnels

TVA : Taxe sur la valeur ajoutée

Liste des tableaux

I.1	Répartition du chiffre d'affaire du crédit personnel et crédit automobile	15
I.2	Répartition de la production brute selon l'affectation réseau et le type client	15
II.1	Tableau d'amortissement à la j^{eme} échéance	18
II.2	Répartition des Intérêts payées par AXA crédit	19
II.3	Tableau de marge d'intérêts favorable	23
II.4	Tableau de marge d'intérêts défavorable	24
III.1	Tableau des effectifs des bons et mauvais emprunteurs	25
III.2	Table de contingence entre type de client et défaut de remboursement	25
III.3	Table de contingence entre l'état matrimonial et défaut de remboursement	26
III.4	Table de contingence entre l'affectation réseau et le défaut de remboursement	26
III.5	Table de contingence entre le mode d'habitation et le défaut de remboursement	26
III.6	Matrice de confusion	32
III.7	Interprétation des valeurs du critère AUC	35
IV.1	Liste des Variables prédictives du modèle logit	42
IV.2	Resultats V cramer	50
IV.3	Matrice de confusion de l'échantillon d'apprentissage	54
IV.4	Matrice de confusion de l'échantillon de test	54
IV.5	Comparaison entre scénarios à PD nulle et ceux à PD estimée	59

Table des figures

I.1	Répartition du chiffre d'affaires d'AAM en fin 2016 et 2017	11
I.2	Résultat net de AAM en fin 2017	11
I.3	Répartition du Capital de AAM en fin 2017	12
I.4	Répartition de la production selon le type de client en 2019	15
I.5	Répartition de la production selon l'affectation réseau en 2019	16
III.1	La fonction Logistique	29
IV.1	les statistiques de base de jeu de données du crédit conso	43
IV.2	Histogramme de l'échantillon d'étude	44
IV.3	Histogramme de la durée de remboursement	45
IV.4	Histogramme de l'age	45
IV.5	Histogramme du montant de crédit	46
IV.6	Résultat du test Kruskal-wallis	47
IV.7	Taux d'impayés selon l'age	48
IV.8	Taux d'impayés selon la durée de remboursement	48
IV.9	Taux d'impayés selon la production brute	49
IV.10	Taux d'impayés selon le revenu	49
IV.11	V de Cramer des variables explicatives entre elles	51
IV.12	Processus de sélection de variables - stepAIC de R - Stepwise	52
IV.13	Modèle sélectionné par le module stepAIC de R - Stepwise	53
IV.14	Diagramme de fiabilité	55
IV.15	La courbe ROC de l'échantillon d'apprentissage	56
IV.16	La courbe ROC de l'échantillon test	56
IV.17	Comparaison du taux probabilisé et non probabilisé en fonction de la marge	59
IV.18	Comparaison entre Tarif fixe et Tarif par dossier pour une marge de 1 million	60
A.1	Extrait de la table de calcul du tarif par dossier pour une marge d'intérêt d'équilibre	68

Sommaire

Résumé	3
Dédicace	4
Remerciements	5
Liste des abréviations	6
Table des figures	8
Introduction	10
I Présentation générale	11
1 Organisme d'accueil	11
1.1 AXA Assurance Maroc	11
1.2 AXA Crédit	12
2 Marché du crédit à la consommation	12
2.1 Vue d'ensemble	12
2.2 Activités d'AXA Crédit	14
II Calcul du compte résultat d'AXA crédit	17
1 Les systèmes d'amortissement	17
1.1 Définition	17
1.2 Échéances constantes	17
1.3 Amortissement constant	18
2 Présentation du passif/Actif	18
2.1 Présentation du Passif	18
2.2 Présentation de L'actif	19
3 Méthodologie appliquée	22
3.1 Scénario favorable	22
3.2 Scénario défavorable	23

3.3	Résultat	23
III Approche théorique du Crédit Scoring		25
1	Analyse descriptive du portefeuille	25
2	Présentation du modèle	28
2.1	Principe et estimation	28
2.2	Évaluation de la régression	31
2.3	Tests de significativité des coefficients	35
2.4	Prédiction et intervalle de prédiction	36
2.5	Interprétation des coefficients	37
2.6	La sélection de variables	37
2.7	Diagnostic de la régression logistique	39
IV Résultats du modèle et refonte des tarifs		41
1	résultats du modèle	41
1.1	Notation	41
1.2	Données	41
1.3	Estimation des parametres	51
2	Refonte des tarifs	57
2.1	Approche théorique	57
2.2	Résultats	58
Conclusion		61
A Annexe		62
Bibliographie		69
Webographie		70

Introduction

AXA crédit est une société spécialisée dans le crédit à la consommation ; elle emprunte des capitaux à un taux bas, à partir desquels elle octroie des crédits aux particuliers.

D'un point de vue prêteur : AXA reçoit chaque mois une mensualité qu'elle découpe de manière à amortir une partie du capital restant dû et dégager des intérêts. D'un point de vue emprunteur, AXA crédit paye des intérêts sur les capitaux qu'elle a empruntés.

Ainsi, la société dégage une marge d'intérêt en retranchant les intérêts sur les crédits octroyés aux particuliers de ceux sur les capitaux qu'elle a empruntés. Cependant, la société fait souvent face à des impayés.

Un impayé survient lorsque le client n'arrive pas à payer une ou plusieurs mensualités pendant la durée de remboursement du crédit. Dans ce cas, AXA n'encaisse pas la mensualité. Elle doit payer la perte associée de son propre gain.

Dans un portefeuille de crédit aux particuliers, AXA crédit ne connaît pas sa marge d'intérêt future. De plus, pour des fins de commercialisation, la société ne fait pas de discrimination entre les bons et les mauvais clients.

Pour mutualiser le risque de défaut et diminuer le montant des impayés futures dans le portefeuille crédit à la consommation, nous procédons d'abord par un crédit scoring en estimant une probabilité de défaut à partir des caractéristiques du client (Age, Revenu, État matrimonial, Mode d'habitation ...) en utilisant une régression logistique binaire.

Nous créons un portefeuille fictif au dessein d'aboutir à une tarification qui :

- Prend en considération la probabilité de défaut de chaque client.
- Dégage une marge d'intérêts fixée du produit crédit à la consommation.
- Respecte la loi qui régleme le marché du crédit à la consommation.

Présentation générale

1 Organisme d'accueil

1.1 AXA Assurance Maroc

AXA Assurance Maroc est une société d'assurance et de réassurance appartenant au groupe AXA. Elle s'adresse aux particuliers et entreprises pour répondre à leurs besoins de services en matière d'assurance, épargne et retraite.

Chiffres clés

Chiffre d'affaires

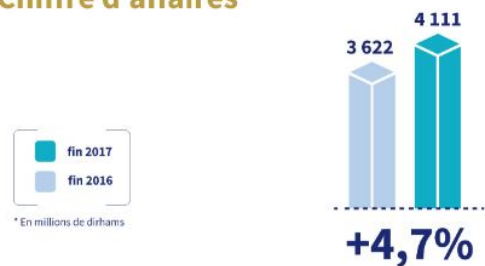


FIGURE I.1 – Répartition du chiffre d'affaires d'AAM en fin 2016 et 2017

Résultat net



FIGURE I.2 – Résultat net de AAM en fin 2017

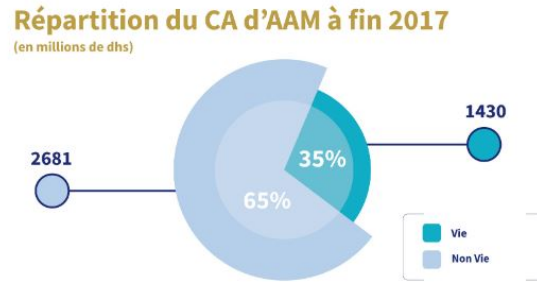


FIGURE I.3 – Répartition du Capital de AAM en fin 2017

1.2 AXA Crédit

Filiale d'AXA Assurance Maroc, AXA Crédit est un établissement financier spécialisé dans le crédit à la consommation. Elle emploie plus de 150 collaborateurs et a réalisé en 2015 un chiffre d'affaires de 214 millions de dirhams. L'offre commerciale d'AXA Crédit s'adresse aux particuliers. Elle octroie plusieurs types de crédits

- Crédit amortissable, affecté ou non affecté ;
- Crédit d'aide à l'immobilier, à taux fixe, réservé à une clientèle particulièrement sélectionnée ;
- Crédit permanent.

Ces offres concernent plusieurs profils de clients, pouvant faire, compte tenu des spécificités de chaque catégorie, l'objet d'approches commerciales adaptées et individualisées. Il s'agit de

- La clientèle directe, bancarisée ;
- Les fonctionnaires actifs relevant de la gestion du Centre National de Traitement qui assure le remboursement des prêts par retenue à la source ;
- Les fonctionnaires retraités du secteur public, relevant de la gestion de la Caisse Marocaine des Retraités ;
- Les salariés des entreprises, ou collectivités avec lesquelles AXA Crédit a conclu un accord de partenariat.

2 Marché du crédit à la consommation

2.1 Vue d'ensemble

Le crédit à la consommation a d'abord une portée économique ; il contribue à stimuler la demande globale dans le but de développer l'investissement et la croissance. Il a également une portée sociale ; il permet aux classes moyennes et aux catégories sociales économiquement faibles d'accéder à certains biens de consommation durable.

Les acteurs institutionnels du crédit à la consommation sont les banques et les sociétés de crédit à la consommation. D'autres opérateurs interviennent sur le marché en accordant des ventes à tempérament. Cela va de l'épicier de quartier au commerce moderne.

Un crédit à la consommation est un crédit accordé à un particulier par une banque (ou un organisme financier spécialisé) directement ou par l'intermédiaire d'un commerçant. Le particulier, en tant qu'emprunteur, s'engage à rembourser une somme d'argent mise à sa disposition majorée des intérêts. Il est destiné au financement de besoins privés, sans rapport avec l'activité professionnelle de l'emprunteur.

Nous distinguons deux catégories de crédits à la consommation

- Le crédit à la consommation affecté : lié à l'achat un bien précis ;
- Le crédit à la consommation non affecté : n'est pas lié à un achat précis. L'emprunteur dispose alors librement du montant emprunté.

Le crédit non affecté se décompose en

- Prêts personnels : un contrat de crédit sur une certaine période aux termes duquel est mise à la disposition d'un individu (particulier) une somme d'argent qui sera remboursée par des mensualités (versements périodiques et constants) pendant la durée du prêt.
- Crédits renouvelables : Un établissement de crédit met à disposition d'un emprunteur une somme d'argent sur un compte spécialement ouvert à cet effet. L'emprunteur, quant à lui, peut disposer de la somme librement, sans justificatifs d'utilisation, à la condition de rembourser son crédit renouvelable selon les échéances contractuelles.
- Locations avec option d'achat (LOA) : Est un mode d'acquisition différent du crédit automobile classique. C'est un contrat par lequel un organisme financier achète le bien et le loue à un individu selon certaines conditions. Au terme du contrat ce dernier a la possibilité de racheter le bien au montant de la valeur de rachat déterminée au début de contrat.
- Découverts en compte autorisés : Le découvert bancaire désigne la situation dans laquelle le compte d'un client d'une banque est débiteur, c'est-à-dire lorsque celui-ci affiche un solde négatif car le montant des débits est supérieurs au montant des crédits.

Le secteur du crédit à la consommation a fait ses débuts au Maroc à partir de la fin des années 30. Il se caractérise actuellement par une concurrence ardue ce qui a engendré un mouvement de concentration. En effet, le nombre des sociétés du secteur n'a cessé de baisser depuis quelques années, passant de 36 sociétés en 1996 à 19 sociétés en 2006. Cette évolution s'explique par :

- Le désencadrement du crédit à partir de 1991 et la libéralisation des taux qui ont engendré un grand intérêt des banques pour le secteur du crédit à la consommation et ont ainsi, recouru à la filialisation de cette activité. De ce fait, le secteur s'élargit par l'arrivée de nouveaux opérateurs.
- L'avènement de la loi du 6 juillet 1993 qui réforme le système bancaire et érige les sociétés de crédit à la consommation en établissements de crédit. Cependant, à partir de 1996, les sociétés qui

n'ont pas pu se conformer à la nouvelle loi, dont notamment les fonds propres minimums, ont dû cesser leur activité.

- La recrudescence de la concurrence au sein du secteur, conjuguée à une décreue du taux maximum des intérêts conventionnels (TMIC), amenant les sociétés de financement à resserrer leurs marges ce qui a enclenché un processus de concentration au sein du secteur à partir de 2001-2002 et s'est traduit par des opérations de fusion- absorption.

En effet, selon le rapport de Bank Al-Maghrib relatif aux sociétés de financement, trois sociétés de crédit à la consommation détenaient, à fin 2006, environ 65% du total-actif de l'ensemble du secteur. Cette part augmente à 78% pour les 5 premiers établissements.

Les sociétés de crédit à la consommation adossées à des institutions financières, au nombre de 10, détenaient une part de près de 93% du total- actif. Globalement, les sociétés de crédit à la consommation adossées à des banques ou à d'autres institutions financières réalisent de bonnes performances comparativement aux sociétés indépendantes. En effet, ces dernières, confrontées à la fois à la baisse du taux maximum des intérêts conventionnels (TMIC) et à la hausse du coût du risque de crédit, supportent un coût de refinancement plus élevé par rapport à la catégorie précédente.

A fin 2006, l'encours des crédits des sociétés de crédit à la consommation a enregistré un accroissement de 13% pour s'établir à 26,9 milliards de dirhams. Cet encours se répartit à hauteur de 9,2 milliards de dirhams pour les crédits affectés, en hausse de 28% et de 17,7 milliards de dirhams pour les crédits non affectés, en progression de 6,5%. Parmi les crédits affectés, l'encours du crédit automobile a atteint 6,9 milliards de dirhams, en hausse de 39,2%. Cette hausse a concerné aussi bien le crédit automobile classique (+5,7% à 1,8 milliard de dirhams), que l'encours LOA (location avec option d'achat) (+57,3% à 5,1 milliards de dirhams). S'agissant des crédits non affectés, l'encours des prêts personnels a atteint 14,8 milliards de dirhams, en hausse de 4%, tandis que le crédit revolving s'est établi à 469 millions de dirhams, en progression de 10,3%. En 2017 l'encours total des crédits des sociétés de crédits à la consommation s'est établi à 48,7 milliards de dirhams. Soit un taux de croissance annuel moyen de 5.54% au cours de la période 2006-2017. Cet encours se répartit comme suit :

- Crédit automobile : 26,3 milliards
- Crédit d'équipement domestique et autres crédits : 395 millions de dirhams
- Prêts personnels : 21,8 milliards
- Crédit revolving : 158 millions de dirhams

2.2 Activités d'AXA Crédit

Positionnement d'AXA Crédit

AXA crédit intervient sur deux segments distincts des crédits de financement. Il s'agit des crédits personnels et des crédits automobiles

Montants en MMAD	Auto			Montants en MMAD	Prêts personnels		
	2015	2014	évolution		2015	2014	évolution
WAFASALAF	1389	1455	-5%	WAFASALAF	1975	1985	-1%
SOFAC	1187	852	39%	SOFAC	623	581	4%
AXA CREDIT	96	36	167%	AXA CREDIT	383	378	1%

TABLE I.1 – Répartition du chiffre d’affaire du crédit personnel et crédit automobile

Au niveau du crédit automobile, AXA crédit a triplé son chiffre d’affaires en passant de 36 MMD en 2014 à 96 MMD en 2015. La société a connu une évolution de 1% dans la même période au niveau des prêts personnels.

Clientèle et gamme de produits d’AXA crédit

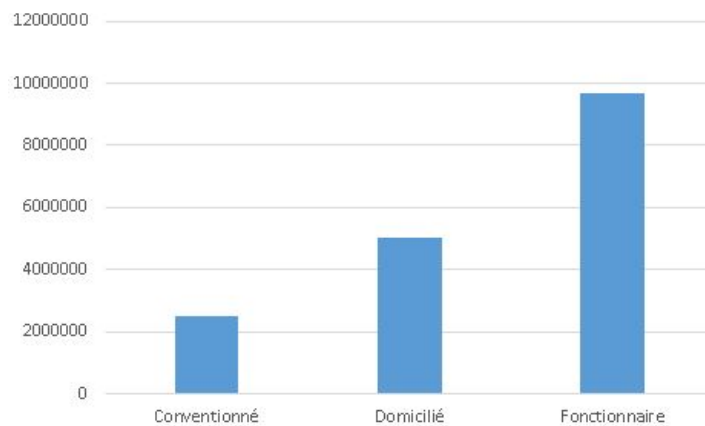


FIGURE I.4 – Répartition de la production selon le type de client en 2019

Les produits mis en place par AXA crédit visent principalement une clientèle issue des entreprises conventionnées ainsi que les fonctionnaires. Ces derniers détiennent 60% de la production brute en 2019. Les conventionnés détiennent 2,5 millions de dirhams.

Le réseau commercial

	AAM	Direct	total ligne
Fonctionnaire	0,07	0,93	1
Domicilié	0,03	0,97	1
Conventionné	0,17	0,83	1

TABLE I.2 – Répartition de la production brute selon l’affectation réseau et le type client

Parmi la clientèle conventionnée ; 17% est issue de AAM. 97% des domiciliés sont type direct d'affectation.

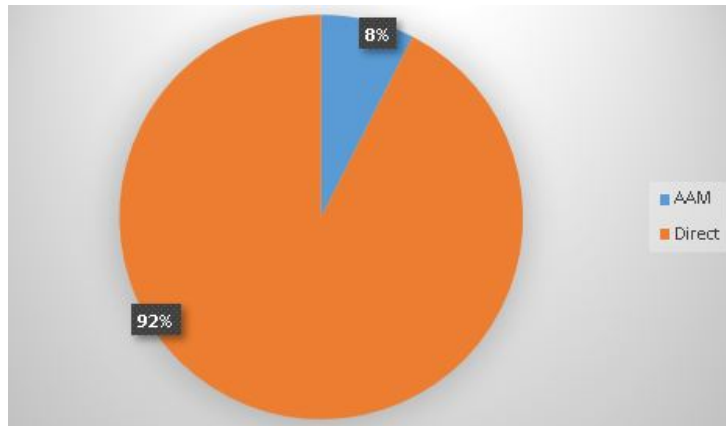


FIGURE I.5 – Répartition de la production selon l'affectation réseau en 2019

La part de production apportée par le réseau direct représente 92%(environ 16 millions de dirhams) en 2019.

Calcul du compte résultat d'AXA crédit

1 Les systèmes d'amortissement

1.1 Définition

L'amortissement d'un emprunt (bancaire ou obligataire) est la partie du capital qui est remboursée à chaque échéance périodique (par exemple chaque mois).

Ce paiement se fait en même temps que celui des intérêts dûs pour la même période. Le versement total (amortissement + intérêts) à chaque échéance est dénommé, selon sa périodicité, la mensualité ou annuité. Il y a deux principales formules possibles d'amortissement : amortissement constant ou échéances constante.

1.2 Échéances constantes

le montant de l'amortissement s'accroît au fur et à mesure que les intérêts diminuent, pour des échéances identiques tout au long de l'emprunt.

Si on note :

M_j : Mensualité à la j^{eme} échéance

CRD_j : Capital restant dû à j^{eme} échéance

A_j : Amortissement à la j^{eme} échéance

I_j : Intérêts à la j^{eme} échéance

i : Taux d'intérêt

n : Durée d'emprunt

C : Montant emprunté

t_j : Montant de la TVA à j^{eme} échéance

Le montant de la mensualité constante s'écrit : $M_j = M = C * \frac{i}{1 - (1 + i)^{-n}}$

Le tableau d'amortissement à la j^{eme} échéance est :

Mois	Intérêts	TVA	Mensualité	CRD	Amortissement
j	$I_j = CRD_j * i$	$t_j = 10\% * I_j$	$M_j = M$	$CRD_j = CRD_{j-1} - A_{j-1}$	$A_j = M - I_j - t_j$

TABLE II.1 – Tableau d'amortissement à la j^{eme} échéance

1.3 Amortissement constant

le montant de l'amortissement reste identique tout au long de la durée de l'emprunt, tandis que les intérêts sont dégressifs de même que les échéances.

Dans le cas de ce type d'amortissement, aussi qualifié de linéaire, il ne s'agit plus de payer le même montant à intervalles réguliers pendant toute la durée de l'emprunt mais de rembourser la même part du capital emprunté à chaque échéance. Les intérêts étant calculés par rapport au capital restant dû, leur montant diminue à chaque échéance. La somme totale à régler chaque mois, composée d'une part fixe liée au remboursement du capital et de la part due aux intérêts, décroît donc au fur et à mesure des échéances, d'où le nom de remboursement à échéances dégressives . Ce mode d'amortissement est très rarement proposé aux particuliers et s'adresse plutôt aux entreprises ou aux collectivités locales. Dans ce cas le tableau d'amortissement est obtenu en utilisant les formules suivantes :

$$\begin{aligned}
 A_j &= \frac{C}{n} \\
 I_j &= CRD_j * i \\
 t_j &= I_j \\
 M_j &= A_j + I_j + t_j \\
 CRD_j &= CRD_{j-1} - A_{j-1} \\
 CRD_1 &= C
 \end{aligned}$$

2 Présentation du passif/Actif

2.1 Présentation du Passif

AXA crédit est une société intermédiaire de financement, elle emprunte à taux bas des capitaux auprès des banques auxquelles elle doit payer en contrepartie des échéances périodiques pendant la durée du remboursement des montants empruntés.

La base de données retenue pour évaluer le passif contient les tableaux d'amortissements de AXA crédit auprès de chaque banque. Pour faciliter le calcul, on regroupe les données dans un tableau croisé dont chaque ligne correspond un contrat et chaque colonne correspond à une date de paiement.

Si on note :

P_i le montant total du passif à payer par AXA crédit à la date i .

$e_{i,j}$: l'intérêt à payer par AXA crédit à la banque j à la date i .

n : L'effectif des banques

	01/2019	02/2019	...
.
Banque j	$e_{j,1}$	$e_{j,2}$...
.
.
Total	$P_1 = \sum_{i=1}^n e_{j,1}$	$P_2 = \sum_{i=1}^n e_{j,2}$...

TABLE II.2 – Répartition des Intérêts payées par AXA crédit

on a alors $P_i = \sum_{i=1}^n e_{i,j}$

Ensuite on actualise les P_i en utilisant la courbe des taux zero coupons au 31 décembre 2018.

En notant :

P^* : le montant total du passif actualisé au 31 décembre 2018.

r_i : le taux d'actualisation au 31 décembre 2018 d'une unité de monnaie payée à une date future i .

on a $P^* = \sum_{i=1}^L \frac{P_i}{(1 + r_i)}$

2.2 Présentation de L'actif

Pour évaluer l'actif, nous utilisons un fichier qui comporte le CRD, les impayées des clients au 31 décembre 2018 ainsi que toutes les informations concernant ces clients à savoir :

Le numéro de dossier : Il désigne le numéro de dossier relatif au client souhaitant bénéficier des offres d'AXA crédit. Un seul client peut avoir plusieurs numéro de dossiers, ceci dépend du nombre de crédits souhaités.

Production nette : Le montant financé.

Production brute : Le montant financé sans prise en compte du montant racheté.

Capital restant dû : Il désigne la partie de la somme empruntée qui reste à rembourser par l'emprunteur. Cette dernière nous sert de base pour le calcul des intérêts à venir, constituant ainsi un élément essentiel du crédit.

Impayé : Il s'agit du montant manquant au paiement total ou partiel d'une ou plusieurs mensualités du crédit. Notons qu'il s'agit du montant cumulé des impayés évalué au 31 décembre de chaque exercice.

Affaires en cours : Il s'agit du statut des affaires en cours. Il comporte deux modalités :

- Encours sain : faisant références au client avec un bon profil et comportement.
- Créances en souffrance (les dossiers provisionnés).

Type clients : les offres de crédit concernent plusieurs profils de clients, pouvant faire, compte tenu des spécificités de chaque catégorie, l'objet d'approches commerciales adaptées et individualisées. Il s'agit de :

- Domicilié : La clientèle directe, bancarisée
- Fonctionnaire : Les fonctionnaires actifs relevant de la gestion du Centre National de Traitement (CNT) qui assure le remboursement des prêts par retenue à la source, et Les retraités du secteur public, relevant de la gestion de la Caisse Marocaine des Retraités (CMR).
- Conventionné : Les salariés des entreprises, ou collectivités avec lesquelles AXA Crédit a conclu un accord de partenariat.

Type de réseau dont on cite trois types :

- Directe
- Partenaire
- Agent générale d'AXA Assurance

Taux : (également appelé le taux nominal) est un taux d'intérêt qui, appliqué à la somme prêtée dans le cadre d'un crédit, permet de calculer le montant des intérêts qui seront payés par l'emprunteur. Avec la durée du prêt accordé par la banque, il permet de définir le nombre et le montant des mensualités. Les intérêts calculés avec le taux de crédit permettent à la banque de compenser son propre coût de refinancement et le risque qu'elle prend en prêtant de l'argent.

Mensualité : Désignant la somme d'argent payé chaque mois par l'emprunteur à AXA crédit

Type de produit : AXA offre plusieurs types de crédits à savoir :

- Le crédit automobile : C'est un crédit à la consommation souscrit auprès d'AXA Crédit, mettant à la disposition des clients une somme d'argent définie en vue d'acheter une voiture neuve ou d'occasion.
- Le prêt personnel : C'est un crédit à la consommation. Il peut permettre au client de financer leurs projets, même sans apport. Le taux d'intérêt, le montant des mensualités et le coût total du crédit sont fixes. Ils dépendent du montant emprunté et de la durée de remboursement choisie par l'emprunteur.
- Le crédit renouvelable : Il est également appelé «crédit revolving» ou «crédit permanent», est un crédit à la consommation par lequel une personne emprunte une somme d'argent dont il reconstitue le montant disponible au fur et à mesure de ses remboursements. Contrairement au crédit affecté, le crédit renouvelable n'est pas lié à l'achat d'un bien en particulier. L'emprunteur peut donc disposer de la somme comme il le souhaite pour financer différents achats.

Libellé apporteur : Ce terme désigne les différents émetteurs des demandes de prêt : une agence directe mandataires ou autres partenaires.

Libellé barème : Un barème de crédit est un document financier utilisé essentiellement dans le crédit bancaire et qui regroupe des données mentionnant les modalités et les conditions de remboursement d'un

crédit.

Date début : Faisant référence à la date de souscription du client à un crédit.

Date fin : Indiquant la date où le remboursement de l'emprunt arrive à terme.

Le numéro de client : il s'agit d'un numéro unique attribué à chaque client souhaitant bénéficier des offres de AXA crédit.

Date de 1 ère échéance : Il s'agit de la date à partir de laquelle le client cesse de rembourser une part ou la totalité de ses mensualités. Date de naissance : La date de naissance du client concerné.

Revenu : Revenu du client concerné

Age : Age de client concerné à la date d'octroi du crédit

Frais de dossier HT : Sont des frais hors taxe pris afin de rémunérer le temps consacré à l'étude d'un dossier

Pour évaluer le montant total de l'actif de AXA crédit à une date j , on construit les tableaux d'amortissement dossier par dossier. Puis, on projette les intérêts à encaisser sur la durée de remboursement du prêt. Ensuite, on fixe une date j et on agrège les intérêts à encaisser à cette date.

on note :

Ac_j : Le montant de l'actif à une date j .

I_j^i : l'intérêt à encaisser à une date j par un dossier i .

$$\text{on a : } Ac_j = \sum_{i=1}^N I_j^i$$

cette formule suppose qu'on a aucun impayé à la date j . En réalité, Lorsque nous faisons face à un impayé, nous ne recevons aucun intérêt et nous perdons la partie amortissable du capital restant dû.

On définit ainsi Y comme étant une variable binaire qui prend la valeur 1 si le client fait défaut ; c'est à dire qu'il n'arrive pas à payer une ou plusieurs mensualités pendant la durée de remboursement du prêt.

Soit Y une variable binaire telle que :

$Y=1$ si le client fait défaut

$Y=0$ sinon

le flux, qu'on note Ac_j^i , perçu à une date j par un dossier i s'écrit :

$$Ac_j^i = I_j^i * 1_{Y=0} - a_j^i * 1_{Y=1}$$

où :

I_j^i le montant d'intérêts que paye le dossier i à la date j .

a_j^i L'amortissement du capital restant dû du dossier i à la date j .

Le flux total, noté Ac_j , perçu à une date j est :

$$Ac_j = \sum_{i=1}^N Ac_j^i$$

3 Méthodologie appliquée

Dans cette section nous expliquons la méthodologie de calcul de la marge d'intérêts de AXA crédit. En effet, cette dernière peut être composée en deux parties

une partie déterministe : La somme du montant total du passif P_j à une date j .

une partie aléatoire qui correspond au montant de l'actif Ac_j à une date j .

La formule de la marge d'intérêts dégagée par AXA Crédit à une date j est :

$$MG_j = Ac_j - P_j \quad (\text{II.1})$$

Le montant des impayés à percevoir à la date j n'est pas connu à l'avance. Puisque il existe deux types d'emprunteurs dans le portefeuille encours :

Les bons emprunteurs sont ceux qui n'ont pas d'impayés depuis le début du contrat jusqu'au 31/12/2018.

Les mauvais emprunteurs sont ceux qui ont des impayés depuis le début du contrat jusqu'au 31/12/2018.

Une première étude considère deux scénarios extrêmes pour calculer la marge d'intérêts.

3.1 Scénario favorable

Le scénario est dit favorable car il suppose que tous les clients sont sains à partir du 31/12/2018.

Nous supposons que :

Pas d'impayés futurs pour les bons clients.

Pas d'impayés futurs pour les mauvais clients.

nous avons alors :

$$\begin{aligned} MG_j &= Ac_j - P_j \\ &= \sum_{i=1}^N Ac_j^i - P_j \\ &= \sum_{i=1}^N (I_j^i) - P_j \\ &= \sum_{i=1}^N (I_j^i) - P_j \end{aligned} \quad (\text{II.2})$$

Pour Calculer la richesse de l'institution au 31/12/2018, nous actualisons les MG_j en utilisant les taux issus de la courbe des taux zero coupon de ce même date. on note r_{ZC}^T le taux annuel zero coupon de maturité T

Le taux d'actualisation d'une date(un mois) j se situant entre les années T et $T-1$ est tel que :

maturité(par année)	1	2	...	T
Taux _{ZC}	r_{ZC}^1	r_{ZC}^2	...	r_{ZC}^T

$$1 + r_j = (1 + r_{ZC}^T)^{T-1} + \frac{j}{12}$$

Nous déduisons que la marge d'intérêts actualisée au 31/12/2018 est :

$$MG^{favorable} = \left(\sum_{j=1}^J \frac{MG_j}{1 + r_j} \right) - C \quad (\text{II.3})$$

C est le montant total des impayés du portefeuille évalué au 31/12/2018.

3.2 Scénario défavorable

Le scénario est dit défavorable car il ne prend pas en considération la probabilité de défaut des clients. on suppose que :

Pas d'impayés futurs pour les bons clients.

Pas de remboursements futurs pour les mauvais clients.

Dans ce cas, le flux financier à une date j d'un dossier i s'écrit :

$$Ac_j^i = \begin{cases} I_j^i & \text{si } Y^i = 0 \\ -(I_j^i + a_j^i) & \text{si } Y^i = 1 \end{cases}$$

si on note N_0 le nombre des bons dossiers et N_1 le nombre les mauvais dossiers, on a :

$$\begin{aligned} MG_j &= Ac_j - P_j - C \\ &= \sum_{i=1}^N Ac_j^i - P_j - C \\ &= \left(\sum_{i=1}^{N_0} I_j^i \right) - \left(\sum_{i=1}^{N_1} (I_j^i + a_j^i) \right) - P_j - C \end{aligned} \quad (\text{II.4})$$

Enfin, La marge d'intérêts au 31/12/2018 s'écrit :

$$MG^{defavorable} = \left(\sum_{j=1}^J \frac{MG_j}{1 + r_j} \right) - C$$

3.3 Résultat

Les montants sont calculés en millions

Date d'actualisation	Actif actualisé (Ac^*)	Passif actualisé(P^*)	Impayés(C)	Marge d'intérêts(MG^*)
31/12/2018	73,8	8,2	46,1	19,5

TABLE II.3 – Tableau de marge d'intérêts favorable

Date d'actualisation	Actif actualisé (Ac^*)	Passif actualisé(P^*)	Impayés(C)	Marge d'intérêts(MG^*)
31/12/2018	32,9	8,2	46,1	-21,4

TABLE II.4 – Tableau de marge d'intérêts défavorable

Approche théorique du Crédit Scoring

1 Analyse descriptive du portefeuille

L'analyse des relations entre le défaut de remboursement et certaines caractéristiques qualitatives des emprunteurs a été menée en utilisant les tableaux croisés et les statistiques y afférents (Khi-deux).

Bons emprunteurs	Mauvais emprunteurs
24994	5573

TABLE III.1 – Tableau des effectifs des bons et mauvais emprunteurs

Les mauvais emprunteurs représentent 18.2% de la population.

Type client	Bons emprunteurs	Mauvais emprunteurs
Fonctionnaire	17917	1030
Domicilié	1605	1631
Conventionné	5472	2912

TABLE III.2 – Table de contingence entre type de client et défaut de remboursement

Les Fonctionnaires représentent 61.9% de l'échantillon ; 94.5% sont considérés comme bons emprunteurs contre 5.5% de mauvais. Environ 27.4% de l'échantillon étaient des Conventionnés. Parmi eux, environ 35% ont eu un défaut de remboursement contre 65% qui ont toujours bien remboursé leurs prêts. Dans l'échantillon, il y a 3236 Domicilié ; 50% d'entre eux ont fait défaut. Le statut de Conventionné être lié à un haut risque de crédit alors celui du Fonctionnaire peut être associé à un faible risque de crédit. Le Chi-deux est significatif, ce qui implique une forte relation entre le défaut et le type de client.

Etat matrimonial	Bons emprunteurs	Mauvais emprunteurs
Marié	19944	3656
Célibataire	4178	1733
Divorcé	711	149
Veuf	161	35

TABLE III.3 – Table de contingence entre l'état matrimonial et défaut de remboursement

Les Mariés représentent 77.2% de l'échantillon ; 84.5% d'entre eux sont considérés comme bons emprunteurs contre 15.5% de mauvais. Parmi les Célibataires , environ 29.3% ont eu un défaut de remboursement contre 70.7% qui ont toujours bien remboursé leurs prêts. Dans l'échantillon, il y a 1056 entre Divorcés et Veufs ; 17.4% d'entre eux ont fait défaut. Le Chi-deux est significatif, ce qui implique une forte relation entre le défaut et l'état matrimonial.

Affectation réseau	Bons emprunteurs	Mauvais emprunteurs
AAM	1879	456
Direct	23115	5117

TABLE III.4 – Table de contingence entre l'affectation réseau et le défaut de remboursement

Les dossiers bancarisés représentent 92.4% de l'échantillon ; 81.8% d'entre eux sont considérés comme bons emprunteurs contre 18.2% de mauvais. Dans l'échantillon, il y a 7.6% des AAM ; 19.5% d'entre eux ont fait défaut. Toutefois, le Chi-deux n'est pas significatif, ce qui implique une faible relation entre le défaut et l'affectation réseau.

Mode d'habitation	Bons emprunteurs	Mauvais emprunteurs
Propriétaire	22124	4609
Habite chez ses parents	2208	777
Locataire	337	130
Logement de fonction	185	24
Autres	140	33

TABLE III.5 – Table de contingence entre le mode d'habitation et le défaut de remboursement

Les Propriétaires représentent 87.5% de l'échantillon ; 82.8% d'entre eux sont considérés comme bons emprunteurs contre 17.2% de mauvais. Parmi les emprunteurs qui habitent chez leurs parents, environ 26% ont eu un défaut de remboursement contre 74% qui ont toujours bien remboursé leurs prêts. Le Chi-deux est significatif, ce qui implique une forte relation entre le défaut et l'état matrimonial.

la régression logistique est utilisée pour développer le modèle . Nous avons eu recours au logiciel R pour mettre en œuvre la technique statistique ci-dessus mentionnée.

L'objectif principal de ce chapitre est de développer un modèle statistique qui puisse permettre de distinguer les bons emprunteurs des mauvais. Une des premières étapes est donc de définir ce que nous entendons par bons et mauvais emprunteurs.

- Un emprunteur est considéré comme bon s'il rembourse (ou a toujours remboursé) correctement son prêt et n'a jamais été en retard de paiement pour trente (30) jours ou plus.
- Un mauvais emprunteur est celui qui a connu au moins une fois un retard dans le remboursement de son prêt pour 30 jours ou plus.

Nous avons principalement utilisé le dossier de crédit pour collecter les données sur les performances de crédit des emprunteurs. Nous avons également eu recours aux agents de crédit pour avoir plus d'information sur certains clients, sur les causes du défaut de remboursement et aussi dans le but d'identifier les variables liées au défaut.

A partir du cadre théorique développé et de la disponibilité des données, nous avons identifié sept variables dont quatre sont qualitatives. Dans le tableau ci- dessous, les variables clés sont définies.

2 Présentation du modèle

Dans cette section, Nous modélisons la probabilité de défaut d'un client selon ses caractéristiques. Nous utilisons la régression logistique car elle accepte une large gamme de distributions. Elle recourt à l'approche du Maximum de Vraisemblance pour estimer les paramètres du modèle. Le terme d'erreur est supposé suivre une distribution logistique.

2.1 Principe et estimation

Pour la régression logistique binaire, Y prend uniquement deux modalités (1, 0) pour simplifier. Nous disposons d'un échantillon Ω de taille n . La valeur prise par Y pour un individu ω est notée $Y(\omega)$. Le fichier de crédit comporte J descripteurs (X_1, X_2, \dots, X_J) . Le vecteur de valeurs pour un individu ω s'écrit $(X_1(\omega), X_2(\omega), \dots, X_J(\omega))$. Dans le cadre binaire, pour un individu donné, sa probabilité a priori d'être positif s'écrit $P[Y(\omega) = 1] = p(\omega)$. Lorsqu'il ne peut y avoir d'ambiguïtés, nous la noterons simplement p .

La probabilité a posteriori d'un individu Ω d'être positif - sachant les valeurs prises par les descripteurs - est notée $P[Y(\omega) = 1/X(\omega)] = \pi(\omega)$. Ici également, lorsqu'il ne peut y avoir de confusions, nous écrirons π . Ce dernier terme est très important. En effet, c'est la probabilité que l'on cherche à modéliser.

Pour un individu ω , on appelle transformation logit de $\pi(\omega)$ l'expression

$$\ln\left(\frac{\pi(\omega)}{1 - \pi(\omega)}\right) = a_0 + a_1 X_1(\omega) + \dots + a_J X_J(\omega) \quad (\text{III.1})$$

a_0, a_1, \dots, a_J sont les paramètres que l'on souhaite estimer à partir des données. Enfin, pour alléger l'écriture, nous omettrons le terme ω lorsque cela est possible.

La quantité $\frac{\pi(\omega)}{1 - \pi(\omega)}$ exprime un rapport de chances. Par exemple, si un individu présente un rapport de chances de 2, cela veut dire qu'il a 2 fois plus de chances d'être positif que d'être négatif.

Posons $C(X) = a_0 + a_1 X_1 + \dots + a_J X_J$, nous pouvons revenir sur π avec la fonction logistique

$$\pi = \frac{1}{1 + e^{-C}} \quad (\text{III.2})$$

Quelques commentaires et remarques

- Le LOGIT = $C(X)$ est théoriquement défini entre $-\infty$ et $+\infty$.
- En revanche, $0 \leq \pi \leq 1$ issue de la transformation de $C(X)$ représente une probabilité.

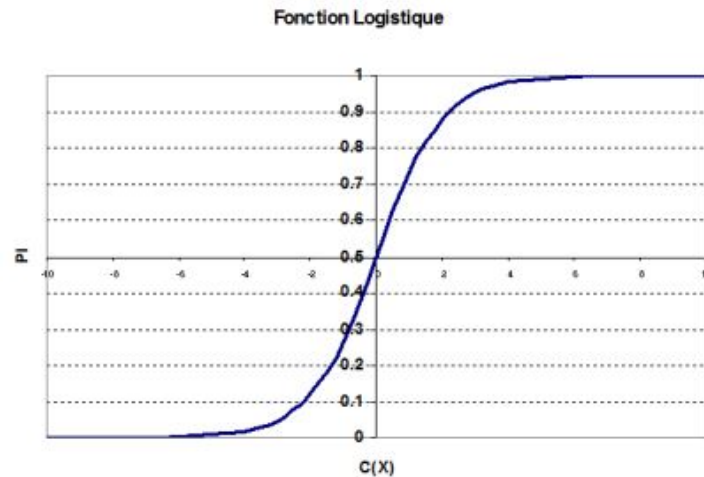


FIGURE III.1 – La fonction Logistique

- $C(X)$ et π permettent tous deux de "scorer" les individus, et par là de les classer selon leur propension à être "positif". Cette fonctionnalité est très utilisée dans le ciblage marketing. On parle de "scoring".
- π représente une probabilité, avec les propriétés inhérentes à une probabilité, entre autres $P(Y = 1/X) + P(Y = 0/X) = 1$.

Les applications de la quantification de la propension d'un individu à être positif (ou négatif) sont nombreuses, certains touchent directement à notre vie quotidienne :

- Déterminer la viabilité d'un client sollicitant un crédit à partir de ses caractéristiques (age, type d'emploi, niveau de revenu, état matrimonial, etc.)
- Quantifier le risque de survenue d'un sinistre pour une personne sollicitant un contrat d'assurance
- Discerner les facteurs de risque de survenue d'une maladie cardio-vasculaire chez des patients (ex.l'âge, le sexe, le tabac, l'alcool, etc.)
- Pour une enseigne de grande distribution, cibler les clients qui peuvent être intéressés par tel ou tel type de produit.

Estimation des paramètres par la maximisation de la vraisemblance

Pour estimer les paramètres de la régression logistique par la méthode du maximum de vraisemblance, nous devons tout d'abord déterminer la loi de distribution de $P(Y / X)$.

Y est une variable binaire définie dans $\{ 1,0 \}$. Pour un individu ω , on modélise la probabilité à l'aide de la loi binomiale $B(1, \pi)$

$$P[Y(\omega)/X(\omega)] = \pi(\omega)^{y(\omega)} * (1 - \pi(\omega))^{(1-y(\omega))} \quad (\text{III.3})$$

Cette modélisation est cohérente avec ce qui a été dit précédemment, en effet

- Si $y(\omega) = 1$, alors $P[Y(\omega) = 1/X(\omega)] = \pi$
- Si $y(\omega) = 0$, alors $P[Y(\omega) = 0/X(\omega)] = 1 - \pi$

La vraisemblance d'un échantillon Ω s'écrit

$$L = \prod_{\omega} \pi^y * (1 - \pi)^{1-y} \tag{III.4}$$

la vraisemblance correspond à la probabilité d'obtenir l'échantillon Ω à partir d'un tirage dans la population. Elle varie donc entre 0 et 1. La méthode du maximum de vraisemblance consiste à produire les paramètres $a = (a_0, a_1, \dots, a_J)$ de la régression logistique qui rendent maximum la probabilité d'observer cet échantillon.

Pour faciliter les manipulations, nous préférons souvent travailler sur la log-vraisemblance

$$LL = \sum_{\omega} y \ln(\pi) + (1 - y) \ln((1 - \pi)) \tag{III.5}$$

Le logarithme étant une fonction monotone, le vecteur a qui maximise la vraisemblance est le même que celui qui maximise la log-vraisemblance. Cette dernière en revanche varie entre $-\infty$ et 0. Puisque \hat{a} est un estimateur du maximum de vraisemblance, il en possède toutes les propriétés

- Il est asymptotiquement sans biais ;
- Il est de variance minimale ;
- Il est asymptotiquement gaussien.

Ces éléments, notamment le dernier, seront très importants pour l'inférence statistique (intervalle de confiance, test de significativité, etc.).

Comme dans toute démarche de modélisation, plusieurs questions se posent immédiatement :

- Choisir la forme de la fonction.
- Estimer les paramètres du modèle à partir d'un échantillon Ω .
- Évaluer la précision des estimations.
- Mesurer le pouvoir explicatif du modèle.
- Vérifier s'il existe une liaison significative entre l'ensemble des descripteurs et la variable dépendante.
- Identifier les descripteurs pertinents dans la prédiction de Y , évacuer celles qui ne sont pas significatives et/ou celles qui sont redondantes.
- Pour un nouvel individu à classer, déterminer la valeur de π à partir des valeurs prises par les X .

La régression logistique permet de répondre précisément à chacune de ces questions. Elle le fait surtout de manière complètement cohérente avec sa démarche de maximisation de la vraisemblance.

Maintenant que nous avons construit un modèle de prédiction, il faut en évaluer l'efficacité. Nous confrontons les valeurs observées de la variable dépendante $Y(\omega)$ avec les prédictions $\hat{Y}(\omega)$. Dans cette section, nous nous consacrons aux méthodes d'évaluation basées sur les prédictions $\hat{y}(\omega)$ fournies par le modèle.

De fait, les techniques et ratios présentés ci-dessous peuvent s'appliquer à tout classifieur issu d'un processus d'apprentissage supervisé, pourvu qu'il sache fournir $\hat{y}(\omega)$ et $\hat{\pi}$ (ex. analyse discriminante, arbres de décision, réseaux de neurones, etc.).

2.2 Évaluation de la régression

Pour évaluer la capacité à bien classer du modèle, nous produisons la démarche suivante :

- Construire la colonne prédiction ;
- Construire la colonne erreur $\Delta(Y, \hat{Y})$;
- Comptabiliser le nombre de mauvais classement ;
- Déduire le taux d'erreur.

Il est plus judicieux de construire une matrice de confusion . Elle confronte toujours les valeurs observées de la variable dépendante avec celles qui sont prédites, puis comptabilise les bonnes et les mauvaises prédictions. Son intérêt est qu'elle permet à la fois d'appréhender la quantité de l'erreur (le taux d'erreur) et de rendre compte de la structure de l'erreur (la manière de se tromper du modèle).

$\Delta(Y, \hat{Y})$	0	1	Total ligne
0	a	b	a+b
1	c	d	c+d
Total colonne	a+c	b+d	a+b+c+d

TABLE III.6 – Matrice de confusion

Dans un problème à 2 classes (1 vs. 0), à partir de la forme générique de la matrice de confusion, plusieurs indicateurs peuvent être déduits pour rendre compte de la concordance entre les valeurs observées et les valeurs prédites. Nous nous concentrons sur les ratios suivants :

- a sont les vrais positifs ; les observations qui ont été classées positives et qui le sont réellement.
- c sont les faux positifs ; les individus classés positifs et qui sont réalité des négatifs.
- b sont les faux négatifs
- d sont les vrais négatifs.

Mais ces termes sont peu utilisés en pratique car les positifs et les négatifs n'ont pas le même statut dans la majorité des études (dans notre cas, nous supposons que les positifs sont les bons clients les négatifs sont les mauvais clients).

Le taux d'erreur est égal au nombre de mauvais classement rapporté à l'effectif total

$$\epsilon = \frac{b+c}{n} = 1 - \frac{a+d}{n}$$

Il estime la probabilité de mauvais classement du modèle.

Le taux de succès correspond à la probabilité de bon classement du modèle, c'est le complémentaire à 1 du taux d'erreur

$$\theta = 1 - \epsilon \tag{III.6}$$

La sensibilité (ou le rappel, ou encore le taux de vrais positifs [TVP]) indique la capacité du modèle à retrouver les positifs

$$Se = TVP = rappel = \frac{a}{a+b} \tag{III.7}$$

La précision indique la proportion de vrais positifs parmi les individus qui ont été classés positifs

$$precision = \frac{a}{a+c} \tag{III.8}$$

Elle estime la probabilité d'un individu d'être réellement positif lorsque le modèle le classe comme tel. Dans certains domaines, on parle de valeur prédictive positive (VPP).

La spécificité, à l'inverse de la sensibilité, indique la proportion de négatifs détectés

$$Sp = \frac{d}{c+d} \tag{III.9}$$

Parfois, on utilise le taux de faux positifs (TFP), il correspond à la proportion de négatifs qui ont été classés positifs

$$TFP = \frac{c}{c+d} = 1 - Sp \quad (\text{III.10})$$

Quelques remarques sur le comportement de ces indicateurs

Un "bon" modèle doit présenter des valeurs faibles de taux d'erreur et de taux de faux positifs (proche de 0); des valeurs élevées de sensibilité, précision et spécificité (proche de 1). Le taux d'erreur est un indicateur symétrique, il donne la même importance aux faux positifs (c) et aux faux négatifs (b).

La sensibilité et la précision sont asymétriques, ils accordent un rôle particulier aux positifs. Enfin, en règle générale, lorsqu'on oriente l'apprentissage de manière à améliorer la sensibilité, on dégrade souvent la précision et la spécificité. Un modèle qui serait meilleur que les autres sur ces deux groupes de critères antinomiques est celui qu'il faut absolument retenir.

Autres indicateurs

La sensibilité et la spécificité jouent un rôle particulier dans l'évaluation des classifieurs. En effet : Un bon modèle doit présenter des valeurs élevées sur ces deux critères d'évaluation. Comme nous le disions plus haut, lorsqu'on oriente l'apprentissage pour améliorer la sensibilité, on dégrade (souvent) la spécificité.

Tous deux partagent une propriété importante : ils ne dépendent pas du schéma d'échantillonnage. Même si l'échantillon n'est pas représentatif; la proportion des positifs (resp. des négatifs) ne reflète pas la probabilité d'être positif (resp. négatif), la sensibilité et la spécificité n'en sont pas affecté.

Diagramme de fiabilité

La régression logistique produit une bonne approximation de la quantité $\pi(\omega)$. La première idée qui vient à l'esprit est de confronter les probabilités estimées par le modèle et celles observées dans le fichier de données. On construit pour cela le diagramme de fiabilité.

Ici également, si nous en avons la possibilité, nous avons tout intérêt à construire le diagramme à partir des données tests n'ayant pas participé à l'élaboration du classifieur. Les indications obtenues n'en seront que plus crédibles.

Voici les principales étapes de la construction du diagramme de fiabilité :

- Appliquer le classifieur sur les données pour obtenir le score $\hat{\pi}(\omega)$.
- Trier le fichier selon le score croissant.
- Sur la base du score, subdiviser les données en intervalles (ex. 0.0-0.2, 0.2-0.4, etc.).
- Dans chaque intervalle, calculer la proportion de positifs.
- Dans le même temps, toujours dans chaque intervalle, calculer la moyenne des scores.

- Si les chiffres concordent dans chaque intervalle, les scores sont bien calibrés, le classifieur est de bonne qualité.
- Nous pouvons résumer l'information dans un graphique nuage de points appelé diagramme de fiabilité, avec en abscisse la moyenne des scores, en ordonnée la proportion de "positifs".
- Si les scores sont bien calibrés, les points devraient être alignés sur une droite, la première bissectrice.
- Les points s'écartant sensiblement de la première bissectrice doivent attirer notre attention.

La courbe ROC

La courbe ROC est un outil très riche. Son champ d'application dépasse largement le cadre de l'apprentissage supervisé. Elle est par exemple très utilisée en épidémiologie. Pour nous, elle présente surtout des caractéristiques très intéressantes pour l'évaluation et la comparaison des performances des classifieurs :

- Elle propose un outil graphique qui permet d'évaluer et de comparer globalement le comportement des classifieurs.
- Elle est indépendante des coûts de mauvaise affectation. Elle permet par exemple de déterminer si un classifieur surpasse un autre, quelle que soit la combinaison de coûts utilisée.
- Elle est opérationnelle même dans le cas des distributions très déséquilibrées. Mieux, même si les proportions des classes ne sont pas représentatives des probabilités à priori dans le fichier, c'est le cas lorsque l'on procède à un tirage rétrospectif. on fixe le nombre de positifs et négatifs à obtenir, et on tire au hasard dans chaque sous-population, la courbe ROC reste valable.
- Enfin, on peut lui associer un indicateur synthétique, le critère AUC (aire sous la courbe, en anglais area under curve), que l'on sait interpréter.

La courbe ROC met en relation le taux de vrais positifs TVP (la sensibilité, le rappel) et le taux de faux positifs TFP ($TFP = 1 - \text{Spécificité}$) dans un graphique nuage de points.

Habituellement, nous comparons $\pi^*(\omega)$ à un seuil $s = 0.5$ pour effectuer une prédiction $Y^*(\omega)$.

Nous pouvons ainsi construire la matrice de confusion et en extraire les 2 indicateurs précités. La courbe ROC généralise cette idée en faisant varier s sur tout le continuum des valeurs possibles entre 0 et 1. Pour chaque configuration, nous construisons la matrice de confusion et nous calculons TVP et TFP.

C'est l'idée directrice. Elle est un peu lourde à mettre en place. Dans la pratique, il n'est pas nécessaire de construire explicitement la matrice de confusion, nous procédons de la manière suivante :

- Calculer le score $\pi^*(\omega)$ de chaque individu à l'aide du modèle de prédiction.
- Trier le fichier selon un score décroissant.
- Considérons qu'il n'y a pas d'ex-aequo. Chaque valeur du score peut être potentiellement un seuil s . Pour toutes les observations dont le score est supérieur ou égal à s , les individus dans la partie haute du tableau, nous pouvons comptabiliser le nombre de positifs $n^+(s)$ et le nombre de négatifs $n^-(s)$. Nous en déduisons :

$$TVP = \frac{n^+(s)}{n^+}$$

$$TFP = \frac{n^-(s)}{n^-}$$

- La courbe ROC correspond au graphique nuage de points qui relie les couples (TVP, TFP). Le

premier point est forcément (0, 0), le dernier est (1, 1).

Deux situations extrêmes peuvent survenir. La discrimination est parfaite. Tous les positifs sont situés devant les négatifs, la courbe ROC est collée aux extrémités Ouest et Nord du repère. Les scores sont totalement inopérants, le classifieur attribue des valeurs au hasard, dans ce cas les positifs et les négatifs sont mélangés. La courbe ROC se confond avec la première bissectrice.

Le critère AUC

Il est possible de caractériser numériquement la courbe ROC en calculant la surface située sous la courbe. C'est le critère AUC. Elle exprime la probabilité de placer un individu positif devant un négatif. Ainsi, dans le cas d'une discrimination parfaite, les positifs sont sûrs d'être placés devant les négatifs, nous avons $AUC = 1$. A contrario, si le classifieur attribue des scores au hasard, il y a autant de chances de placer un positif devant un négatif que l'inverse, la courbe ROC se confond avec la première bissectrice, nous avons $AUC = 0.5$. C'est la situation de référence, notre classifieur doit faire mieux. Certains auteurs proposent généralement différents paliers pour donner un ordre d'idées sur la qualité de la discrimination.

Valeur de l'AUC	Commentaire
$AUC = 0.5$	Pas de discrimination.
$0.7 \leq AUC < 0.8$	Discrimination acceptable
$0.8 \leq AUC < 0.9$	Discrimination excellente
$AUC \geq 0.9$	Discrimination exceptionnelle

TABLE III.7 – Interprétation des valeurs du critère AUC

Pour calculer l'AUC, la méthode des trapèzes par exemple.

Au final, il apparaît que le critère AUC est un résumé très commode. Il permet, entre autres, les comparaisons rapides entre les classifieurs. Mais il est évident que si l'on souhaite analyser finement leur comportement, rien ne vaut la courbe ROC.

2.3 Tests de significativité des coefficients

Hypothèses à tester

L'objectif des tests de significativité est d'éprouver le rôle d'une, de plusieurs, de l'ensemble, des variables explicatives. Formellement, les hypothèses nulles peuvent se décliner comme suit :

- Évaluer la contribution individuelle d'une variable

$$H_0 : a_j = 0$$

Ce test de significativité est systématiquement donné par les logiciels. Nous verrons plus loin que seule une de ses formes (test de Wald) est en réalité proposée. L'autre (test du rapport de vraisemblance) est passée sous silence. Or ces approches ne se comportent pas de la même manière. Il faut le savoir pour interpréter les résultats en connaissance de cause.

- Évaluer la contribution d'un bloc de "q" variables. Sans restreindre la généralité du propos (les coefficients à tester ne sont pas forcément consécutifs dans la régression), nous écrirons H_0 de la

manière suivante

$$H_0 : a_j = a_{j+1} = \dots = a_{j+q} = 0$$

Nous ne pouvons pas le transformer en une succession de tests individuels. En effet, les coefficients ne sont pas indépendants (en tous les cas, ils ont une covariance non-nulle). Il faut bien tester la nullité simultanée des q coefficients.

- Évaluer l'apport de l'ensemble des variables explicatives.

$$H_0 : a_1 = a_2 = \dots = a_j = 0$$

Il s'agit d'une évaluation globale de la régression. En effet, si l'hypothèse nulle est compatible avec les données, cela signifierait qu'aucun des descripteurs ne contribue à l'explication de la variable dépendante. Le modèle peut être jeté aux orties.

Dans tous les cas, l'hypothèse alternative correspond à : "un des coefficients au moins est non-nul". Notons que ces tests s'inscrivent dans le cadre d'une formulation générale de la forme

$$H_0 : M * a = 0$$

où M est une matrice de contrastes indépendants à m lignes et $J+1$ colonnes, de rang m . La procédure et les formules sont un peu complexes, mais nous pouvons évaluer tout type de configuration.

Deux approches pour les tests

Nous disposons de cette stratégies pour implémenter ces tests :

- S'appuyer sur la normalité asymptotique des estimateurs (du maximum de vraisemblance). On parle de test de Wald. Le principal avantage est que les informations que l'on souhaite exploiter sont toutes disponibles à l'issue de l'estimation du modèle complet, incluant l'ensemble des variables. L'obtention des résultats est donc immédiate. L'inconvénient est que le test de Wald est conservateur. Il a tendance à favoriser l'hypothèse nulle.

2.4 Prédiction et intervalle de prédiction

Un des principaux objectifs de l'apprentissage supervisé est de fournir un système de classement qui, pour un nouvel individu quelconque ω issu de la population (ex. un nouveau client pour une banque, un malade qui arrive au service des urgences, etc.), fournit une prédiction $\hat{y}(\omega')$. Avec exactitude si possible. La régression logistique sait faire cela. Mais, à la différence d'autres méthodes, elle peut fournir en plus un indicateur de fiabilité de la prédiction avec une estimation de la probabilité $\hat{\pi}(\omega')$. Ainsi, lorsque $\hat{\pi}$ est proche de 1 ou de 0, la prédiction est plutôt sûre ; lorsqu'elle prend une valeur intermédiaire, proche du seuil d'affectation s ($s = 0.5$ habituellement), la prédiction est moins assurée. Dans les domaines où les conséquences des mauvaises affectations peuvent être dramatiques (dans le domaine de la santé par exemple), on pourrait même imaginer un système qui ne classe qu'à coup (presque) sûr du type :

Si $\hat{\pi} \leq s_1$ Alors $\hat{Y} = -$

Si $\hat{\pi} \geq s_2$ Alors $\hat{Y} = +$, avec $s_2 \gg s_1$ bien entendu.

Sinon, indétermination. On demande des analyses complémentaires ou on présente le sujet à un expert.

Obtenir une estimation $\hat{\pi}$ et une indication sur sa précision nous est donc fort utile. Dans ce chapitre, nous montrons comment calculer $\hat{\pi}$ pour un nouvel individu à classer, puis nous étudierons la construction d'un intervalle (fourchette) de prédiction. Ce dernier point constitue aussi une avancée considérable par rapport aux d'autres méthodes supervisées. Nous disposons d'une indication sur la plage de valeurs crédibles de π .

Prédiction ponctuelle

Pour obtenir une prédiction du LOGIT pour un nouvel individu ω' à classer, il nous suffit d'appliquer les coefficients estimés de la régression logistique, soit

$$\hat{c}(X(\omega')) = \hat{a}_0 + \hat{a}_1 * X_1(\omega') + \dots + \hat{a}_J * X_J(\omega') \quad (\text{III.11})$$

Si nous adoptons une écriture matricielle, avec $X(\omega') = (1, X_1(\omega'), \dots, X_J(\omega'))$ la description de l'individu à classer et $\hat{a}' = (\hat{a}_0, \hat{a}_1, \dots, \hat{a}_J)$ le vecteur des paramètres estimés, nous écrivons

$$\hat{c}(X(\omega')) = X(\omega') * \hat{a}$$

Pour alléger l'écriture, nous écrivons simplement \hat{c} dans ce qui suit. A partir du LOGIT, nous pouvons déduire une estimation de la probabilité a posteriori d'être positif de l'individu, soit

$$\hat{\pi}(\omega') = \frac{1}{1 + e^{-\hat{c}}} \quad (\text{III.12})$$

Et en appliquant la règle d'affectation standard, nous obtenons \hat{Y}

2.5 Interprétation des coefficients

Dans certains domaines, l'explication est bien plus importante que la prédiction 1. On souhaite comprendre les phénomènes de causalité, mettre à jour les relations de cause à effet. Bien entendu, les techniques statistiques n'ont pas vocation à répondre mécaniquement à des problèmes complexes. En revanche, elles ont pour rôle de donner aux experts les indications adéquates pour qu'ils puissent se concentrer sur les informations importantes. La régression logistique propose des outils qui permettent d'interpréter les résultats sous forme de risques, de chances, de rapports de chances. C'est certainement une des raisons pour laquelle elle a gagné les faveurs d'un large public d'utilisateurs. Un signe qui ne trompe pas, une large documentation est dédiée à l'interprétation des sorties de la régression logistique dans les ouvrages qui font référence.

Risque relatif : On appelle risque relatif le surcroît de chances d'être positif du groupe exposé par rapport au groupe témoin.

2.6 La sélection de variables

La sélection de variables est une étape clé de la modélisation. Dans les études réelles, nous sommes confrontés à des bases de données avec un nombre considérable de descripteurs. Ce sont autant de variables explicatives potentielles. Certaines d'entre elles sont redondantes, d'autres n'ont aucun rapport

avec la variable dépendante.

La méthode statistique doit nous donner des indications sur le sous- ensemble des bonnes variables à inclure dans le modèle. Dans l'idéal, elles devraient être orthogonales entre elles et toutes fortement liées avec la variable dépendante.

Même si l'expert du domaine a une certaine idée des explicatives à retenir, une sélection automatique peut l'aiguiller sur les pistes à étudier.

Plusieurs raisons nous poussent à réduire le nombre de variables explicatives :

Moins il y aura de variables, plus facile sera l'interprétation. En évacuant les descripteurs qui ne sont pas nécessaires à l'explication de la variable dépendante, nous pouvons plus facilement cerner le rôle de celles qui sont retenues.

La régression logistique nous propose des outils merveilleux pour lire les coefficients en termes de surcroît de risque. Réduire le nombre de variables permet d'en profiter pleinement.

Le déploiement sera facilité. Lorsque le modèle sera mis en production, on a toujours intérêt à poser peu de questions pour identifier la classe d'appartenance d'un individu.

Idem, vous sollicitez un crédit auprès d'une banque, elle commence à vous demander la date de naissance de votre arrière grand-père, la question d'après vous êtes déjà dans l'établissement d'à-côté. Au fil du temps, je me suis rendu compte qu'un système aussi efficace soit-il n'est vraiment adopté par les utilisateurs que s'il est peu contraignant, simple d'utilisation.

Dernier argument en faveur de la sélection, pour un même nombre d'observations, un modèle avec peu de variables a de meilleures chances d'être plus robuste en généralisation.

En effet, lorsque le nombre de paramètres du modèle est trop élevé, le sur-apprentissage nous guette. Le classifieur "colle" trop aux données et, au lieu d'intégrer les informations essentielles qui se rapportent à la population, il ingère les particularités de l'échantillon d'apprentissage.

Introduire des variables explicatives non pertinentes augmente artificiellement les variances des coefficients, les estimations sont numériquement instables.

Dans cette section, nous étudierons deux approches : la sélection par optimisation implémentée dans R,. Tous deux se rejoignent sur le mode d'exploration de l'espace des solutions, il s'agit de procédures pas-à-pas qui évaluent une succession de modèles emboîtés : R, de plus, dispose de la méthode STEP-WISE , elle consiste à vérifier si chaque ajout de variable ne provoque pas le retrait d'une explicative qui aurait été intégrée précédemment.

Ces techniques numériques nous proposent des scénarios de solutions. Il ne faut surtout pas prendre pour argent comptant les sous- ensembles de variables explicatives proposées. D'autant qu'ils peuvent varier d'une stratégie à une autre, et même d'un échantillon d'apprentissage à un autre. Il faut plutôt les considérer comme des alternatives que l'on peut soumettre et faire valider par un expert du domaine.

La sélection de variables est un maillon de la démarche exploratoire. Nous pouvons nous appuyer sur ses

résultats pour essayer des combinaisons de variables, des transformations, réfléchir sur la pertinence de ce que l'on est en train de faire, etc.

Sélection par optimisation

La sélection par optimisation consiste à trouver le sous-ensemble de variables prédictives qui minimise l'un des critères suivants

Le critère AIC d'Akaike

$$AIC = -2LL + 2 * (J + 1) \tag{III.13}$$

et le critère BIC de Schwartz

$$BIC = -2LL + \ln(n) * (J + 1) \tag{III.14}$$

où $-2LL$ est la déviance, $(J + 1)$ est le nombre de paramètres à estimer, avec J le nombre de variables explicatives.

Ces deux critères sont assez similaires finalement. BIC pénalise plus la complexité du modèle dès que l'effectif n augmente (dès que $\ln(n) > 2$). Ça ne veut pas dire qu'il est meilleur ou moins bon. Il privilégie simplement les solutions avec moins de variables explicatives par rapport à AIC.

Ce n'est pas parce que la variable a été sélectionnée via cette procédure d'optimisation qu'elle sera significative au sens du test du rapport de vraisemblance ou du test de Wald dans la régression.

2.7 Diagnostic de la régression logistique

L'analyse des résidus permet de diagnostiquer la qualité de la régression. Plusieurs questions se posent à l'issue du processus de modélisation, nous devons y apporter des réponses :

- Déterminer les points qui "clochent" dans les données, qui s'écartent fortement des autres dans l'espace de représentation. On parle de données "atypiques".
- Déterminer les points qui sont mal modélisés (mal expliqués) par la régression logistique. On parle de résidus.
- Déterminer les points qui pèsent fortement dans la régression. On parle de points "leviers".
- Déterminer les points qui pèsent exagérément sur les résultats. Si on les retirait de l'ensemble d'apprentissage, le modèle obtenu serait très différent. On parle de points "influent".

La modélisation de la variable $Y \in \{1,0\}$ peut s'écrire sous la forme suivante

$$Y(\omega) = \pi(\omega) + \epsilon(\omega) \tag{III.15}$$

$\epsilon(\omega)$ est l'erreur de modélisation, avec $\epsilon(\omega) = Y(\omega) - \pi(\omega)$, elle peut prendre deux valeurs possibles :

$\epsilon(\omega) = 1 - \pi(\omega)$ avec la probabilité $\pi(\omega)$

$\epsilon(\omega) = -\pi(\omega)$ avec la probabilité $1-\pi(\omega)$

Nous calculons aisément :

$$E(\epsilon) = \pi(1 - \pi) + (1 - \pi)(-\pi) = 0$$

$$V(\epsilon) = \pi(1 - \pi)$$

La variance de l'erreur n'est pas constante, elle dépend des individus. Il y a hétéroscédasticité. Pour un individu ω , le résidu de Pearson permet d'identifier les points mal modélisés

$$r(\omega) = \frac{y(\omega) - \hat{\pi}(\omega)}{\sqrt{\hat{\pi}(\omega)(1 - \hat{\pi}(\omega))}} \quad (\text{III.16})$$

Le résidu de Pearson prend une valeur d'autant plus élevée que $\hat{\pi}$ est proche de 0 ou de 1. La distribution de r est approximativement gaussienne $N(0,1)$. Ainsi, tout point en dehors de l'intervalle ± 2 (au niveau de confiance 95%) sont suspects .

Il est aussi important de détecter les éventuels décrochements, les observations qui prennent des valeurs inhabituelles par rapport aux autres. Un graphique est très précieux pour cela.

Chapitre IV

Résultats du modèle et refonte des tarifs

1 résultats du modèle

1.1 Notation

L'objectif est de prédire les valeurs prises par la variable aléatoire Y définie dans (y_1, y_2, \dots, y_K) . Pour la régression logistique binaire, Y prend uniquement deux modalités $\{1, 0\}$. Nous disposons d'un échantillon Ω de taille n . La valeur prise par Y pour un individu ω est notée $Y(\omega)$. Le fichier comporte J descripteurs (X_1, X_2, \dots, X_J) . Le vecteur de valeurs pour un individu ω s'écrit $(X_1(\omega), X_2(\omega), \dots, X_J(\omega))$. Dans le cadre binaire, pour un individu donné, sa probabilité a priori d'être positif s'écrit $P[Y(\omega) = 1] = p(\omega)$.

La probabilité a posteriori d'un individu ω d'être positif c.-à-d. sachant les valeurs prises par les descripteurs est notée $P[Y(\omega) = 1/X(\omega)] = \pi(\omega)$. Ici également, lorsqu'il ne peut y avoir de confusions, nous écrirons π . Ce dernier terme est très important. En effet, c'est la probabilité que l'on cherche à modéliser. Le logit d'un individu ω s'écrit

$$\ln\left(\frac{\pi(\omega)}{1 - \pi(\omega)}\right) = a_0 + a_1 X_1(\omega) + \dots + a_J X_J(\omega) \quad (\text{IV.1})$$

(a_0, a_1, \dots, a_J) sont les paramètres que l'on souhaite estimer à partir des données. Enfin, toujours pour alléger l'écriture, nous omettrons le terme ω lorsque cela est possible.

1.2 Données

Nous utiliserons une base de données comportant 30567 observations et 7 variables prédictives pour illustrer la régression logistique binaire. L'objectif est de prédire la présence ou l'absence d'un défaut de paiement à la date d'octroi du crédit ($Y=1$ si il y a défaut et $Y=0$ sinon) à partir des variables citées ci-dessous.

Nous obtenons une série d'indicateurs lorsque nous le traitons avec le logiciel R. Certaines permettent d'évaluer la qualité globale de la régression, d'autres permettent de juger la contribution individuelle de chaque variable. Expliciter les principes qui régissent la méthode et décrire les formules associées pour

Notation	Nom des Variables	Type des Variables	Modalités
X_1	Type client	Qualitative	Fonctionnaire, Domicilié ou Conventionné
X_2	Affectation réseau	Qualitative	Direct ou AAM
X_3	État matrimonial	Qualitative	Marié, Célibataire, Divorcé ou Veuf
X_4	Mode d'habitation	Qualitative	Propriétaire ou non
X_5	Production brute	Quantitative	Continue
X_6	Age	Quantitative	Continue
X_7	Durée de remboursement	Quantitative	Discrète

TABLE IV.1 – Liste des Variables prédictives du modèle logit

que nous sachions lire en connaissance de cause les résultats constituent les objectifs de ce support. La taille du fichier est suffisante pour que l'on puisse détailler tous les calculs. Même si L'effectif induit une certaine stabilité des résultats. Dans certains cas ils ne concordent pas avec nos connaissances usuelles. Il ne faudra pas s'en formaliser. L'intérêt d'avoir recours à un expert du domaine justement est qu'il a la possibilité de valider ou d'invalider le fruit de calculs purement mécaniques.

Parmi les solutions envisageables, la plus simple consiste à évaluer le classifieur sur des données à part qui n'ont pas participé au processus d'apprentissage. Nous procédons de la manière suivante lorsque l'on dispose d'un échantillon Ω de taille n :

1. Nous tirons au hasard n_a individus parmi n , il s'agit de l'échantillon d'apprentissage, nous les utilisons pour construire le modèle de prédiction M_a . On dédie généralement 70% des données à l'apprentissage. Mais ce n'est pas aussi simple, nous en discuterons plus loin.
2. Sur les n_t observations restantes, l'échantillon test, nous appliquons le modèle M_a , et nous élaborons la matrice de confusion en confrontant les valeurs observées et les valeurs prédites.

Habituellement, $\frac{n_t}{n} = 1 - \frac{n_a}{n} = 30\%$.

Principal atout de cette approche, les indicateurs ainsi obtenus sont non-biaisés. Ils permettent de comparer les mérites respectifs de plusieurs modèles, même s'ils sont de complexité différente, même s'ils ne reposent pas sur des systèmes de représentation identiques. C'est la démarche à privilégier si l'on dispose de suffisamment d'observations.

Bref, les proportions habituellement mises en avant (70% vs. 30%) ne doivent pas être prises au pied de la lettre. Tout est affaire de compromis : il en faut suffisamment pour l'apprentissage afin de produire un modèle consistant ; il en faut suffisamment pour le test afin d'obtenir une évaluation fiable des performances. Les "bonnes" proportions dépendent souvent des caractéristiques du classifieur et des données analysées (rapport entre le nombre d'observations et le nombre de variables, degré de difficulté du concept à apprendre, etc.).

Exploration des données

LA modélisation des données doit être précédée de leur exploitation, afin de se familiariser avec ces données, de s'assurer de l'absence d'anomalie, de mesurer la liaison des variables explicatives entre elles et avec la variable à expliquer ; d'examiner tous les points susceptibles de devoir être pris en compte au moment de la modélisation. R est bien adapté à cette tâche.

Voici les statistiques de base de jeu de données étudié. On constate l'absence de valeurs manquantes. Les variables quantitatives sont décrites dans R par des quartiles. Les variables qualitatives sont décrites par effectifs de chaque modalité.

```
> summary(b)
  Type.client      aff.réseau      PROD.BRUTE      Durée.de.remboursement
Conventionné : 8384      AAM      : 2335      Min.      :      1      Min.      :  8.00
Domicilié   : 3236      Direct:28232  1st Qu.: 4884      1st Qu.: 60.00
Fonctionnaire:18947      Mean     : 9860      Mean     : 73.02
                                     3rd Qu.:14700     3rd Qu.: 84.00
                                     Max.     :21127     Max.     :240.00

  age      revenu      em      mh      Y
Min.   :19.00      Min.   :  666.7      autres: 6967      Autre: 3834      0:24994
1st Qu.:37.00      1st Qu.: 3355.0      Marié :23600      Prop :26733      1: 5573
Median :48.00      Median : 4633.3
Mean   :45.35      Mean   : 6319.9
3rd Qu.:54.00      3rd Qu.: 7450.9
Max.   :77.00      Max.   :432064.0
```

FIGURE IV.1 – les statistiques de base de jeu de données du crédit conso

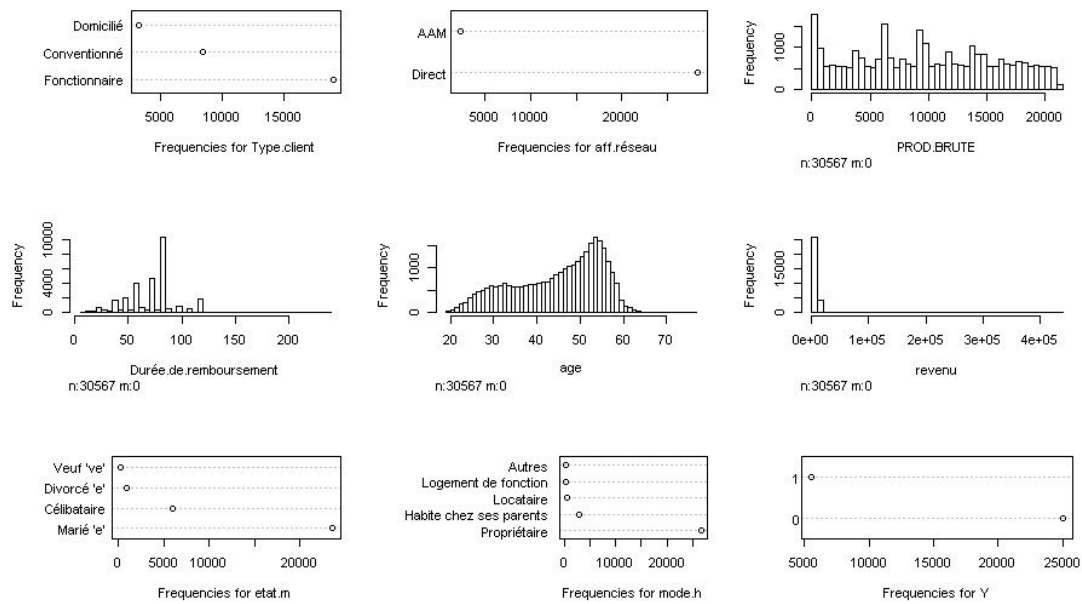


FIGURE IV.2 – Histogramme de l'échantillon d'étude

Analyse des variables continues

Avant de commencer la modélisation, nous allons examiner rapidement la base de données, en commençant par les quatre variables continues : la durée de remboursement, le montant du crédit (PROD.BRUTE), l'âge à l'octroi du crédit et le revenu. Nous le faisons en juxtaposant la distribution de ces variables sur les bons et mauvais dossiers.

Ces graphiques permettent de détecter aussitôt des liaisons entre la variable continue et la variable à expliquer, ainsi que d'éventuelles anomalies, par exemple des valeurs anormalement grandes ou des pics inattendus dans la distribution. Ces derniers ne se détectent pas nécessairement aussi facilement en lisant de longs tableaux de chiffres qu'en regardant un graphique, surtout si un pic est proche de la moyenne. Et épilucher des pages de chiffres est peut-être plus précis mais devient vite plus fastidieux que de consulter des graphiques.

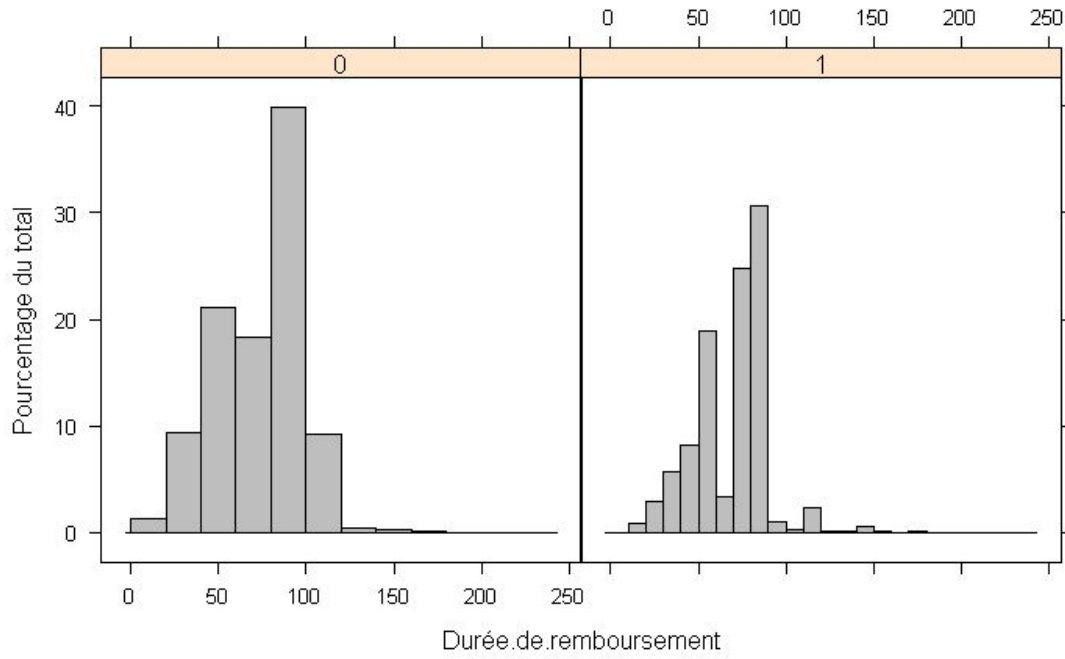


FIGURE IV.3 – Histogramme de la durée de remboursement

La durée de remboursement présente des pics prévisibles entre 48 et 100 mois..

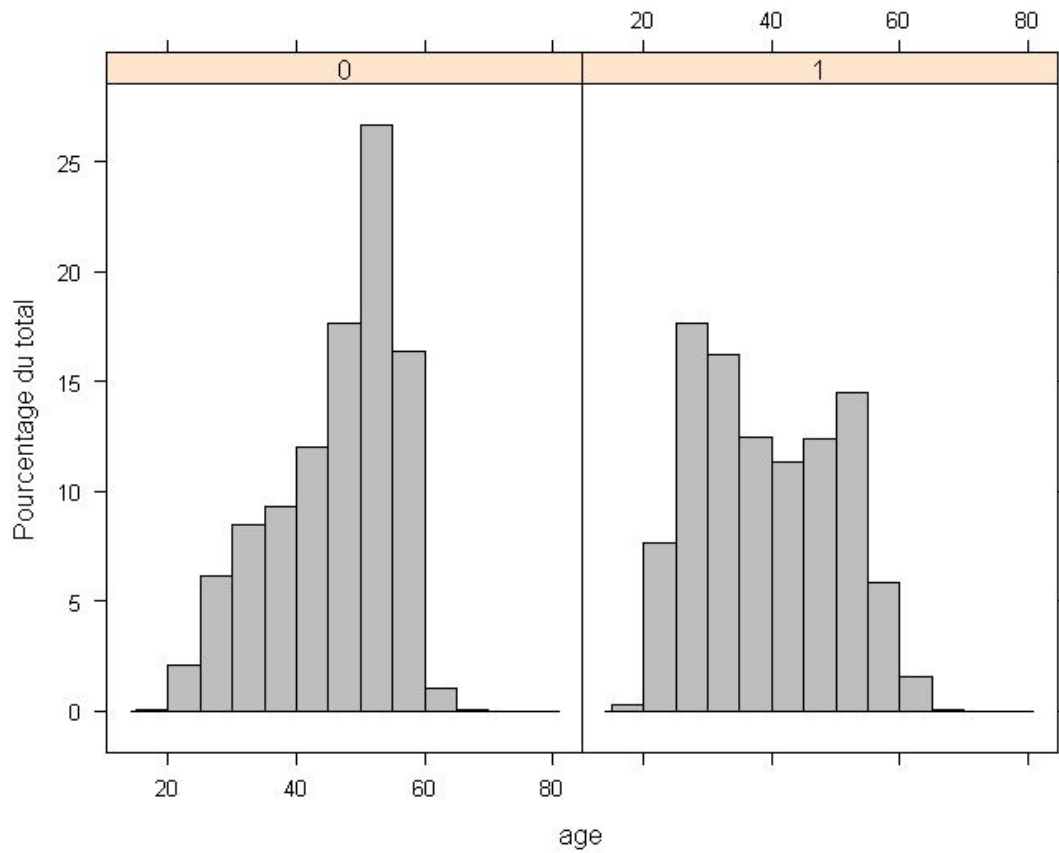


FIGURE IV.4 – Histogramme de l'âge

L'âge des contractants est compris entre 19 et 77 ans. Il a une distribution plus homogène pour les crédits sans impayé que pour les autres. Les impayés concernent majoritairement les contractants de moins de 40 ans lors de l'octroi du crédit.

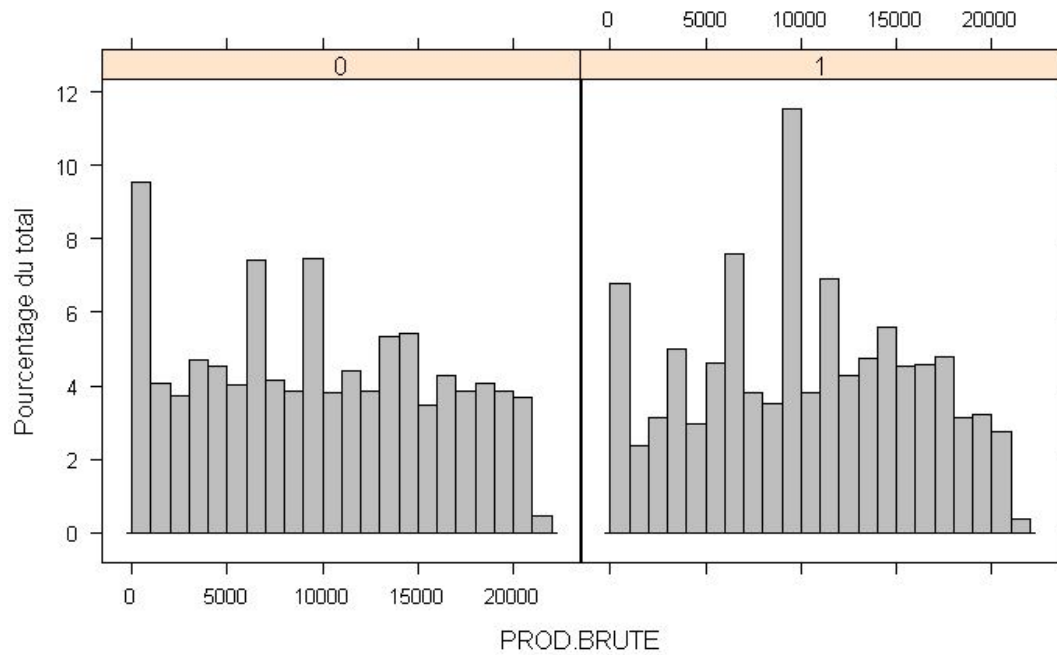


FIGURE IV.5 – Histogramme du montant de crédit

On atteint une fréquence maximale autour de 10000 dh pour les crédits qui ont des impayés. On constate la plus forte proportion de montants plus élevés parmi les crédits qui ont des impayés. La distribution présente une régularité acceptable et pas d'anomalie apparente.

```

> kruskal.test(b$age~b$Y)

Kruskal-Wallis rank sum test

data: b$age by b$Y
Kruskal-Wallis chi-squared = 1796.6, df = 1, p-value < 2.2e-16

> kruskal.test(b$PROD.BRUTE~b$Y)

Kruskal-Wallis rank sum test

data: b$PROD.BRUTE by b$Y
Kruskal-Wallis chi-squared = 22.178, df = 1, p-value = 2.484e-06

> kruskal.test(b$Durée.de.remboursement~b$Y)

Kruskal-Wallis rank sum test

data: b$Durée.de.remboursement by b$Y
Kruskal-Wallis chi-squared = 217.55, df = 1, p-value < 2.2e-16

> kruskal.test(b$revenu~b$Y)

Kruskal-Wallis rank sum test

data: b$revenu by b$Y
Kruskal-Wallis chi-squared = 670.83, df = 1, p-value < 2.2e-16

```

FIGURE IV.6 – Résultat du test Kruskal-wallis

Nous voyons que les trois variables continues présentent chacune une liaison significative avec la variable à expliquer, que nous pouvons mesurer à l'aide d'un test non-paramétrique, ici celui de Kruskal-Wallis. Ce type de test permet en effet de s'affranchir des hypothèses de normalité et d'homoscédasticité habituelles dans les tests paramétriques.

Discrétisation des variables continues

Les graphiques et tableaux précédents ne nous disent pas si les variables continues pourront être utilisées telles quelles dans la modélisation, ou s'il vaudra mieux les discrétiser. Pour chaque variable continue X , la discrétisation est superflue si la liaison de cette variable avec la variable à expliquer est linéaire, c'est-à-dire ici si le taux d'impayés est une fonction linéaire de X ou d'une transformation simple de X , par exemple X_2 , mais pourra sinon s'imposer, surtout si le taux d'impayés n'est pas une fonction monotone de X .

Pour représenter le taux d'impayés en fonction de l'âge, de la durée et du montant de crédit, voici comme nous procédons.

Prenons d'abord l'âge. On commence par calculer les déciles de l'âge, puis on calcule les taux d'impayés par déciles. qui montrent une nette inflexion entre le deuxième et le troisième décile. La fonction 1 : découpe la variable du premier argument en un certain nombre d'intervalles ou selon, comme ici, des seuils prescrits.

On représente le taux d'impayés par intervalle, et signaler par un pointillé horizontal le taux d'impayés moyen (17%).

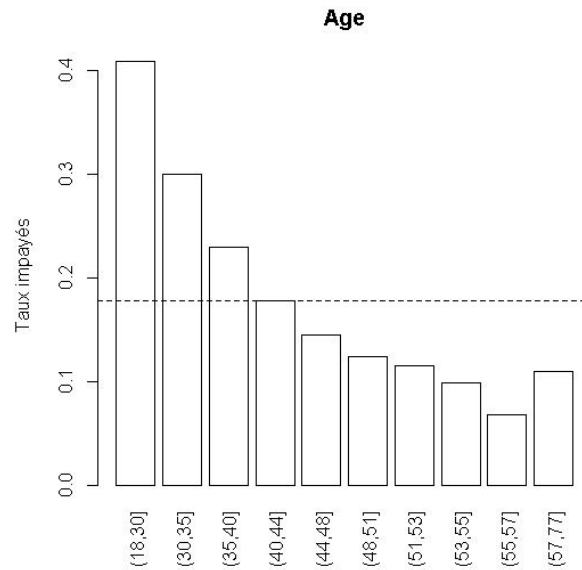


FIGURE IV.7 – Taux d'impayés selon l'âge

Sur la Figure se distinguent très nettement les trois premiers déciles, dont le taux d'impayés est sensiblement supérieur à la moyenne. En revanche, aucune autre tendance ne se dessine aussi fortement dans les autres déciles. Le taux d'impayés baisse à 40 ans puis remonte à 57 ans. Nous découperons donc l'âge en deux tranches : moins de 40 ans et plus de 40 ans.

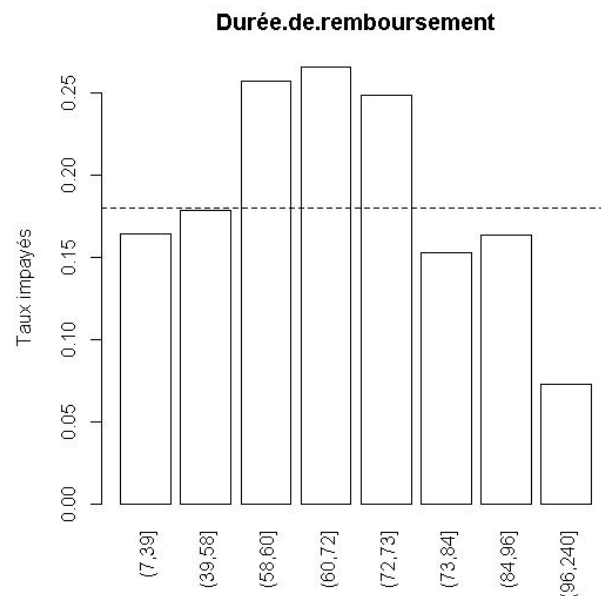


FIGURE IV.8 – Taux d'impayés selon la durée de remboursement

Le taux d'impayés dépasse largement la moyenne entre les 58 et 73 mois . En revanche, aucune autre tendance. il baisse pour les autres modalités. Nous découperons donc la durée de remboursement en 3 tranches : moins de 58 mois , plus de 73 mois et entre 58 et 73 mois.

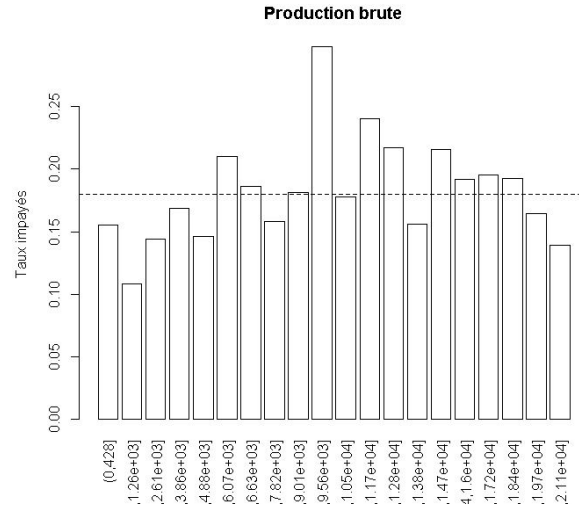


FIGURE IV.9 – Taux d'impayés selon la production brute

Nous découperons la production brute en deux tranches : moins de 9000 dh et plus de 9000dh.

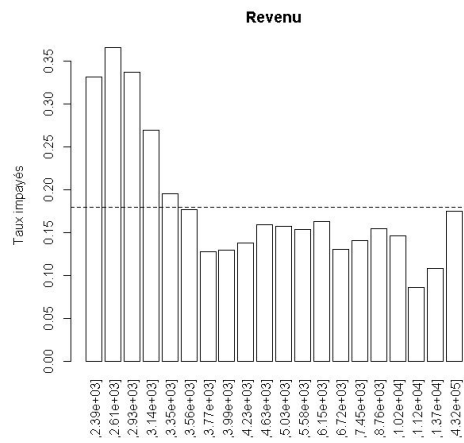


FIGURE IV.10 – Taux d'impayés selon le revenu

Nous découperons le revenu en deux tranches : moins de 4000 dh et plus de 4000 dh.

Liaison des variables explicatives avec la variable à expliquer

Nous allons commencer par inspecter le pouvoir discriminant de chaque variable explicative, en la croisant avec la variable à expliquer, et en quantifiant la liaison à l'aide du V de Cramer.

Au vu des taux d'impayés et des effectifs des modalités des variables explicatives qualitatives ou discrètes, nous allons aussi procéder à certains regroupements de modalités, lorsqu'elles ont des taux impayés proches et des effectifs très petits.

Nous allons tout d'abord quantifier la liaison entre chacun des prédicteurs et la variable à expliquer, en calculant le V de Cramer de chaque paire, puis en affichant la liste des variables explicatives par valeur décroissante du V de Cramer.

Pour le calcul du V de Cramer, nous avons choisi d'appliquer la définition du V de Cramer, qui ici est simplement la racine carrée du ratio ($\chi / \text{effectif}$) puisque la variable à expliquer n'a que deux classes.

Nous créons un data frame intermédiaire, en excluant les 3 variables continues, et nous lui ajoutons les formes précédemment discrétisées des variables continues, afin d'avoir une idée du V de Cramer de l'âge, et de la durée et du montant du crédit, et de leur importance par rapport à celui des autres variables.

Variable	V de Cramer	$pvalue(\chi_2)$
Typede client	0.43	0
age discrétisé	0.21	0
revenu discrétisé	0.17	0
durée discrétisée	0.13	0
état matrimonial	0.13	0
mode d'habitation	0.06	0
production discrétisée	0.04	0
affectation réseau	0.009	0.1

TABLE IV.2 – Resultats V cramer

Nous constatons que les variables explicatives sont statistiquement significatives au seuil de 5%. Seule la variable affectation réseau est statistiquement non significative.

Liaison des variables explicatives entre elles

La modélisation logit exige de s'assurer de l'absence de liaisons trop fortes entre les variables explicatives elles-mêmes. Avec des variables ici toutes qualitatives ou discrètes, le V de Cramer est une mesure appropriée de liaison.



FIGURE IV.11 – V de Cramer des variables explicatives entre elles

Les liaisons les plus fortes sont aisément repérables dans la figure par leur couleur plus foncée. La plus forte est entre l'age et le type de client (V de Cramer = 0.48). Ensuite vient la liaison entre l'age et le état matrimonial (V de Cramer = 0.34), puis entre le type de client et le revenu (0.29), entre l'etat matrimonial et le mode d'habitation(0,28), et entre le type de client et l'état matrimonial (0.27). Entre des variables figurant dans un modèle de régression. on peut généralement considérer comme gênants les V de Cramer dépassant 0,40 en valeur absolue.

1.3 Estimation des parametres

Nous utilisons le logiciel R qui, avec la commande stepAIC du package MASS, implémente la sélection de variables par optimisation.

Selection STEPWISE

Pour initier une sélection stepwise, nous utilisons la commande stepAIC. Elle utilise par défaut le critère AIC. Le modèle constitué uniquement de la constante (modele) sert de point de départ. stepAIC lance la procédure de recherche, et modele.stepwise réceptionne la régression finale intégrant les variables sélectionnées.

```

Start:  AIC=14638.73
Y ~ 1

      Df Deviance  AIC
+ Type.client  2    11762 11768
+ age.d        1    13917 13921
+ r.d          1    14238 14242
+ em           1    14372 14376
+ duree.d      2    14394 14400
+ mh           1    14555 14559
+ prod.d       1    14608 14612
<none>                14637 14639

Step:  AIC=11768.49
Y ~ Type.client

      Df Deviance  AIC
+ r.d        1    11695 11703
+ duree.d    2    11710 11720
+ em         1    11751 11759
+ age.d      1    11753 11761
+ mh         1    11755 11763
<none>                11762 11768
+ prod.d     1    11761 11769
- Type.client 2    14637 14639

Step:  AIC=11702.83
Y ~ Type.client + r.d

```

FIGURE IV.12 – Processus de sélection de variables - stepAIC de R - Stepwise

Le processus se poursuit tant que l'on réduit le critère AIC. Dès que le critère stagne ou reparte à la hausse, le processus de recherche est stoppé.

```
> summary(modele.stepwise)

Call:
glm(formula = Y ~ Type.client + r.d + duree.d + mh + em, family = binomial,
    data = appren)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4548 -0.4229 -0.3288 -0.2668  2.6857

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.82065   0.07565 -10.847 < 2e-16 ***
Type.clientDomicilié  0.69357   0.06057  11.451 < 2e-16 ***
Type.clientFonctionnaire -2.14469   0.06045 -35.477 < 2e-16 ***
r.d(3.35e+03,4.32e+05] -0.45049   0.05104  -8.826 < 2e-16 ***
duree.d(58,73]      0.49752   0.06415   7.755 8.82e-15 ***
duree.d(73,240]    0.42719   0.06571   6.501 7.97e-11 ***
mhProp             0.26147   0.06503   4.021 5.80e-05 ***
emMarié           -0.16321   0.05364  -3.043 0.00234 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14637  on 15282  degrees of freedom
Residual deviance: 11607  on 15275  degrees of freedom
AIC: 11623

Number of Fisher Scoring iterations: 5
```

FIGURE IV.13 – Modèle sélectionné par le module stepAIC de R - Stepwise

Au final, cinq variables explicatives sont sélectionnées. Dans le modèle qui en découle, nous constatons qu'elles sont toutes statistiquement significatives au sens du test de Wald à 5% . une variable peut être intégrée au sens du critère AIC, sans pour autant être significative au sens du test de Wald. Nous remarquons que la variable age est écartée du modèle, ceci est du sa liaison forte avec les autres variables.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.09925	0.10217	-10.759	< 0.05
Type.clientDomicilié	0.69380	0.06066	11.437	< 0.05
Type.clientFonctionnaire	-2.15785	0.06054	-35.643	< 0.05
aff.réseauDirect	0.34878	0.08472	4.117	< 0.05
duree.d(58,73]	0.49423	0.06421	7.697	< 0.05
duree.d(73,240]	0.42283	0.06581	6.425	< 0.05
r.d(3500,43200]	-0.46672	0.05118	-9.118	< 0.05
mhProp	0.23600	0.06532	3.613	0.000302
emMarié	-0.16421	0.05366	-3.060	0.002212

Le modèle estimé s'écrit :

$$\hat{\pi} = \frac{1}{1 + e^{-\hat{C}}}$$

telle que :

$$\begin{aligned} \hat{C} = & -1.09925 + 0.69380 * 1_{Domicil} - 2.15785 * 1_{Fonctionnaire} + 0.34878 * 1_{Direct} \\ & + 0.49423 * 1_{d(58,73]} + 0.42283 * 1_{d(73,240]} - 0.46672 * 1_{R(3500,43200]} \\ & + 0.23600 * 1_{Proprietaire} - 0.16421 * 1_{marie} \end{aligned}$$

matrice de confusion

La règle de classement utilisée est la suivante : Si $\hat{\pi} > 0.5$ alors $\hat{Y} = 1$

Nous pouvons former la matrice de confusion en confrontant les colonnes Y et \hat{Y} .

$Y * \hat{Y}$	<i>bons</i>	<i>mauvais</i>
bons	12057	2341
mauvais	399	486

TABLE IV.3 – Matrice de confusion de l'échantillon d'apprentissage

Nous en déduisons les principaux indicateurs d'évaluation des classifieurs :

- Taux d'erreur=18%
- Taux de succès=82%
- Sensibilité =83.7%
- Précision=96%
- Spécificité=54%

Nous pouvons aussi former la matrice de confusion de l'échantillon de test

$Y * \hat{Y}$	<i>bons</i>	<i>mauvais</i>
bons	12139	2239
mauvais	399	507

TABLE IV.4 – Matrice de confusion de l'échantillon de test

Nous en déduisons les principaux indicateurs d'évaluation des classifieurs :

- Taux d'erreur=17%
- Taux de succès=83%
- Sensibilité =84%
- Précision=96.8%
- Spécificité=55.9%

En termes de performances, nous constatons que le modèle issu de la régression logistique est bon puisque le taux d'erreur pour les deux échantillons est faible et est le même.

Diagramme de fiabilité

De nouveau, nous reproduisons les étapes permettant d'obtenir le diagramme de fiabilité :

- Nous avons estimé les paramètres du modèle à l'aide de R
- Nous calculons alors le LOGIT pour chaque individu.
- Nous en déduisons le score
- Une fois calculé tous les scores, et le tableau trié, nous décidons de procéder à un découpage en 6 intervalles, définies par (0.00-0.1, 0.1-0.2, etc.)
- Dans chaque intervalle nous comptabilisons la proportion de positifs et, dans le même temps, nous calculons la moyenne des scores (nous avons utilisé les tableaux croisés dynamiques pour cela).
- Il ne reste plus qu'à produire le diagramme de fiabilité. Concernant le fichier CREDIT, nous constatons que le modèle produit une bonne estimation des quantités $\pi(\omega)$, les points sont quasiment alignés sur une droite.

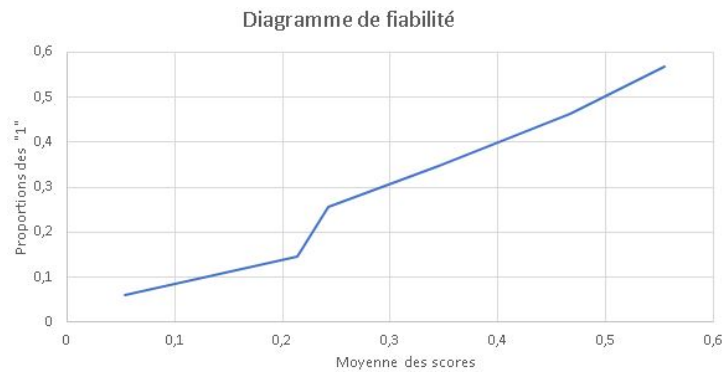


FIGURE IV.14 – Diagramme de fiabilité

Courbe ROC et Critère AUC

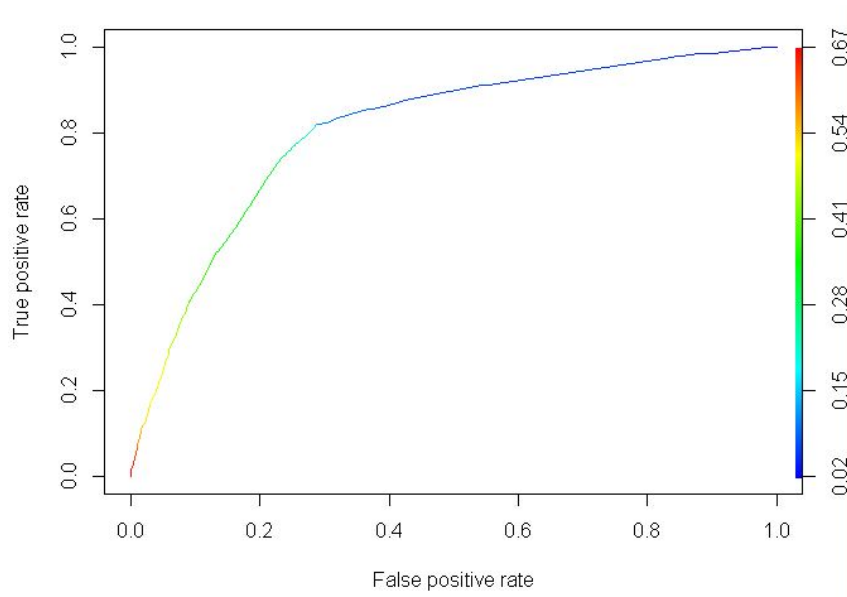


FIGURE IV.15 – La courbe ROC de l'échantillon d'apprentissage

Nous avons 80.5% de chances de placer un bon client devant un mauvais client en "scorant" avec notre classifieur, à comparer avec les 50%(première bissectrice) de la situation de référence. Ce résultat est plutôt encourageant. En observant le graphique, la courbe s'écarte sensiblement de la première bissectrice. Elle indique que notre modèle est plutôt acceptable avec des estimations $\hat{\pi}$ discriminatoires.

Nous calculerons aussi La valeur de AUC de l'échantillon de test

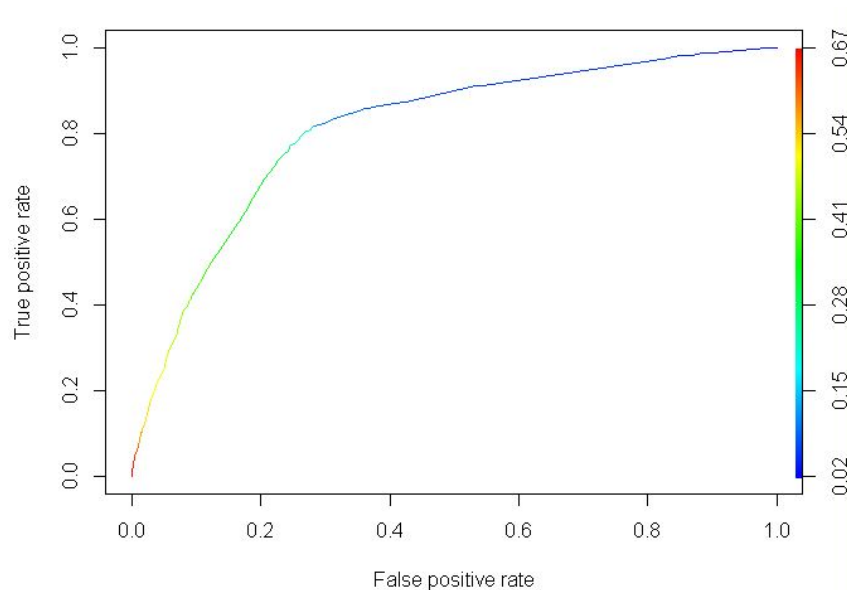


FIGURE IV.16 – La courbe ROC de l'échantillon test

Nous avons 80.8% de chances de placer un bon client devant un mauvais client en "scorant" avec notre classifieur dans l'échantillon test. En termes de performances, le modèle est acceptable

2 Refonte des tarifs

Dans cette, nous intégrons la probabilité de défaut $\hat{\pi}$ d'un individu ω dans le calcul de la marge d'intérêts. Le modèle de scoring précédemment estimé sera acceptable si la marge d'intérêts calculée - en tenant compte des $\hat{\pi}$ - est positive.

2.1 Approche théorique

Les caractéristiques d'un individu ω à une échéance h sont :

- Durée de remboursement n
- Différé d
- Montant du crédit C
- Taux nominal t
- Taux de TVA tva

Nous posons $X = 1+t \Leftrightarrow X-1 = t$ et travaillons dans le cas d'un remboursement à échéances constantes. Soit h le nombre de mensualités payées. Nous avons $h \in [d; n-d]$ La mensualité à l'échéance h s'écrit :

$$\begin{aligned} M_h = M &= C * \frac{t}{(1+t)^{-d} - (1+t)^{-n}} \\ &= C * \frac{X-1}{X^{-d} - X^{-n}} \end{aligned} \quad (IV.2)$$

Le capital restant dû après l'échéance h s'écrit :

$$\begin{aligned} CRD_h &= C * (1+t)^{d+h} - M * \left(\sum_{i=0}^{h-1} (1+t)^i \right) \\ &= C * (X)^{d+h} - (X^h - 1) * \frac{M}{X-1} \end{aligned} \quad (IV.3)$$

L'intérêt après l'échéance h est :

$$\begin{aligned} I_h &= CRD_h * t \\ &= CRD_h * (X-1) \\ &= (C * (X)^{d+h} - (X^h - 1) * \frac{M}{X-1}) * (X-1) \end{aligned} \quad (IV.4)$$

L'amortissement après l'échéance h est :

$$\begin{aligned} a_h &= M - I_h - tva * I_h \\ &= M - (1+tva) * I_h \end{aligned} \quad (IV.5)$$

Le flux reçu par AXA crédit après la h^{eme} échéance d'un individu ω , dont la probabilité de défaut est $p(\omega)$, s'écrit :

$$\begin{aligned} A_h &= I_h * (1 - p(\omega)) - a_h * p(\omega) \\ &= (1 + p * tva) I_h - M * p \end{aligned} \quad (IV.6)$$

Comme la valeur p est inconnu, nous l'estimons avec les calculs faits dans le chapitre précédent en la remplaçant par $\hat{\pi}$. La formule devient :

$$\begin{aligned}\hat{A}_h(\omega) &= I_h * (1 - \hat{\pi}(\omega)) - a_h * \hat{\pi}(\omega) \\ &= (1 + \hat{\pi} * tva)I_h - M * \hat{\pi}\end{aligned}\tag{IV.7}$$

La valeur estimée de l'actif que détient AXA crédit après la h^{eme} échéance est :

$$\hat{R}_h = \sum_{\omega \in \Omega} \hat{A}_h(\omega)\tag{IV.8}$$

La valeur estimée de la marge d'intérêt actualisée de AXA crédit est :

$$\hat{M}^* = \left(\sum_{h \in [d; n-d]} \hat{R}_h \right) - P - C\tag{IV.9}$$

Où P est la valeur actualisée du Passif et C est la valeur des impayés évaluée à la date d'actualisation.

Dans ce qui suit, nous allons définir un tarif qui prend en considération la probabilité de défaut de l'individu,. Nous procédons de la manière suivante :

- D'abord, nous définissons un tarif fixe pour tous les dossiers qui réalise une marge d'intérêt Mg , que nous notons X' et qui vérifie :

$$\begin{aligned}\hat{M}^*(X') &= Mg \\ X(\omega) &= X' \quad \forall \omega \in \Omega\end{aligned}\tag{IV.10}$$

- Puis, nous recalculons les valeurs estimés des flux reçus par AXA crédit pour chaque individu ω avec le tarif X' :

$$\begin{aligned}\hat{A}_h(X', \omega) &= I'_h * (1 - \hat{\pi}(\omega)) - a'_h * \hat{\pi}(\omega) \\ &= (1 + \hat{\pi} * tva)I'_h - M' * \hat{\pi}\end{aligned}\tag{IV.11}$$

- Ensuite, nous calculons la valeur actualisée des flux reçus d'un individu ω

$$\hat{A}^*(X', \omega) = \sum_{h \in [d; n-d]} \frac{\hat{A}_h(X', \omega)}{1 + r_h}$$

- Enfin, nous définissons un tarif individuel que nous notons \tilde{X} qui vérifie l'équation suivante :

$$\hat{A}^*(X', \omega) = \sum_{h \in [d; n-d]} \frac{I_{h, \omega}(\tilde{X})}{1 + r_h}\tag{IV.12}$$

Comme il n'existe pas de formule mathématique explicite des solutions de ces équations dont le degré supérieur à 2, nous avons recours aux solutions numériques en utilisant EXCEL.

2.2 Résultats

Par souci de confidentialité, nous avons construit un portefeuille fictif de 572 observations sur lequel nous appliquons la démarche de calcul citée ci-dessus.

Nous supposons que :

- P+C=1771380

Les calculs de rentabilité aboutissent aux résultats suivants :

Scénario	Marge d'intérêt \hat{M}^*	Taux fixe(PD estimée)	Taux fixe avec PD nulle
0	0	6.88%	3.88%
1	500000	7.84%	4.83%
2	1000000	8.8%	5.77%
3	1500000	9.74%	6.7%
4	2000000	10.68%	7.63%
5	2500000	11.6%	8.55%
6	3000000	12.53%	9.46%
7	3500000	13.45%	10.36%

TABLE IV.5 – Comparaison entre scénarios à PD nulle et ceux à PD estimée

Nous constatons que le taux d'intérêts risqué (2eme colonne du tableau) dépasse le taux d'intérêt à probabilité de défaut nulle de 3% quelque soit la marge d'intérêt. Graphiquement, Ceci est traduit par une translation vers le haut de la courbe du taux probabilisé par rapport au taux non probabilisé. Le résultat obtenu est logique puisque le taux probabilisé doit couvrir la partie incertaine de la marge d'intérêt.

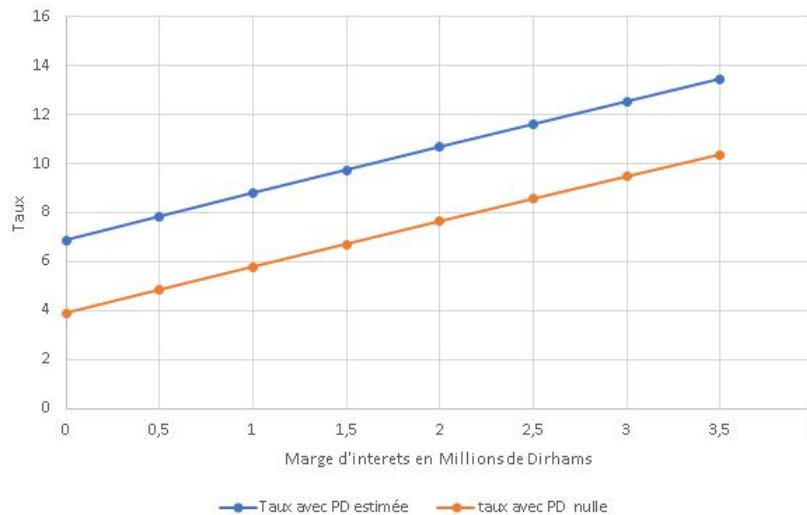


FIGURE IV.17 – Comparaison du taux probabilisé et non probabilisé en fonction de la marge

un taux probabilisé de 7.8% dégage une marge de 0.5 millions tandis q'un taux non probabilisé de même valeur dégage environ 2 millions.

Pour réaliser une marge d'intérêts de 1 million, le modèle suggère un taux fixe de 8,36% pour tous les dossiers du portefeuille.

L'inclusion de la probabilité de défaut d'un individu dans le caclul du taux dégage une faible rentabilité. Plusieurs éléments permettent d'expliquer ce constat :

				Recette	2 771 379,99
Bouton 9					
tarif fixe	tarif/Dossier	taux mensuel	X	VPM	
8,37	5,70	0,48	1,0052294	264,75	
8,37	4,18	0,35	1,00383321	282,80	
8,37	6,14	0,51	1,00563231	2984,98	
8,37	5,70	0,48	1,0052294	3706,28	
8,37	3,47	0,29	1,0031788	2181,00	
8,37	6,48	0,54	1,00593547	7102,91	
8,37	6,48	0,54	1,00593547	2623,05	
8,37	5,70	0,48	1,0052294	3446,88	
8,37	4,18	0,35	1,00383321	1307,97	
8,37	6,48	0,54	1,00593547	2776,35	
8,37	5,70	0,48	1,0052294	3613,81	
8,37	5,24	0,44	1,00480631	1112,09	

FIGURE IV.18 – Comparaison entre Tarif fixe et Tarif par dossier pour une marge de 1 million

Nous constatons que les dossiers à faibles probabilités de défaut reçoivent des taux plus bas que ceux à grande probabilité de défaut. Nous pouvons conclure que le modèle logit estimé est acceptable

- le capital emprunté est faible, et génère donc peu d'intérêt ;
- ce même capital est rapidement remboursé et génère donc peu d'intérêt sur leur durée de vie totale ;
- les frais d'ouverture et les commissions versées au réseau ne sont pas incluses dans le calcul de la marge.

Conclusion

Cette étude a permis de concevoir un modèle de crédit scoring des prêts aux particuliers mettant ainsi en évidence des tendances de comportement des crédits. Le modèle élaboré présente de nombreux avantages, en particulier :

- les paramètres correspondent aux caractéristiques initiales du crédit, elles sont facilement accessibles dans les systèmes et connues dès la mise en place du crédit.
- Il fonctionne bien pour prévoir la capacité d'un individu à rembourser un crédit.
- un modèle facilement intégrable dans le calcul de la marge d'intérêts et dans la tarification des crédits.

Le modèle apparaît satisfaisant, pour autant il ne peut être considéré comme un modèle prédictif statistiquement fiable. Pour cela il faudrait analyser l'influence d'autres facteurs conjoncturels (taux de croissance, chômage, niveau de taux, concurrence...)

D'autre part, des études relatives aux risques doivent être menées pour étudier le comportement dans le temps, en fonction du capital emprunté, des taux ou des objets et ainsi affiner la marge nette dégagée.

Les conclusions d'une telle analyse ne doivent pour autant pas conduire à arrêter cette activité, il est indispensable de conserver une offre de crédit complète pour capter l'ensemble des besoins des clients.

Annexe

Code R du modèle Logit

```
b=read.csv2("b.csv")
b=b[,c("Type.client","aff.réseau","PROD.BRUTE","Durée.de.remboursement","age","revenu",
"em","mh","Y")]
head(b)
str(b)
b$PROD.BRUTE=as.double(b$PROD.BRUTE)
b$Y=as.factor(b$Y)
str(b)
summary(b)
b=na.omit(b)
attach(b)
Breaksage = c(min(b$age),45, max(b$age))
age.d = cut(b$age, breaks = Breaksage, include.lowest = TRUE)

Breaksage = c(min(b$revenu),3350, max(b$revenu))
r.d = cut(b$revenu, breaks = Breaksage, include.lowest = TRUE)

Breaksage = c(0,58,73,max(b$Durée.de.remboursement))
duree.d = cut(b$Durée.de.remboursement, breaks = Breaksage, include.lowest = TRUE)

Breaksage = c(min(b$PROD.BRUTE),9000, max(b$PROD.BRUTE))
prod.d = cut(b$PROD.BRUTE, breaks = Breaksage, include.lowest = TRUE)

b2 =cbind(b,age.d,r.d,duree.d,prod.d)
detach(b)
```

```
attach(b2)

xtabs(~Y + age.d, data = b2)
b3=subset(b2, select = -c(age, revenu,Durée.de.remboursement,PROD.BRUTE))
summary(b3)

chisq.test(table(b2$r.d,b2$duree.d))
chisq.test(table(b3$Y,b3$r.d))
chisq.test(table(b3$Y,b3$duree.d))
chisq.test(table(b3$Y,b3$prod.d))
chisq.test(table(b3$Y,b3$em))
chisq.test(table(b3$Y,b3$mh))
chisq.test(table(b3$em,b3$Type.client))
chisq.test(table(b3$Y,b3$aff.réseau))

# Tirage aléatoire et sans remise des 50% des individus de l'échantillon On
# initialise le tirage aléatoire afin de retomber sur nos pieds à chaque
# fois
set.seed(1111)
d = sort(sample(nrow(b3), nrow(b3) * 0.50))
# Echantillon d'apprentissage
appren <- b3[d, ]
# Echantillon de test
test <- b3[-d, ]

summary(appren)

attach(appren)

str(appren)
# modèle trivial réduit à la constante
str_constant <- "~ 1"
# modèle complet incluant toutes les explicatives potentielles
str_all <- "~Type.client+em+mh+prod.d+duree.d+age.d+r.d"
require(MASS)
## Loading required package: MASS
modele <- glm(Y ~ 1, data = appren, family = binomial)
```

```
modele.forward <- stepAIC(modele, scope = list(lower = str_constant, upper = str_all), trace = TRUE,

# affichage du modèle final
summary(modele.forward)

modele <- glm(Y ~ 1, data = appren, family = binomial)
modele.stepwise <- stepAIC(modele, scope = list(lower = str_constant, upper = str_all),
  trace = TRUE, data = appren, direction = "both")
summary(modele.stepwise)

logit = function(formula, lien = "logit", data = NULL) {
  glm(formula, family = binomial(link = lien), data)
}

m.logit <- logit(Y ~Type.client+aff.réseau+ duree.d + r.d + mh + em , data = appren)
# résultats du modèle
summary(m.logit)

exp(cbind(OR = coef(m.logit), confint(m.logit)))

par(mfrow = c(1, 1))
plot(rstudent(m.logit), type = "p", cex = 0.5, ylab = "Résidus studentisés ",
  col = "springgreen2", ylim = c(-3, 3))
abline(h = c(-2, 2), col = "red")

appren.p <- cbind(appren, predict(m.logit, newdata = appren, type = "link",
  se = TRUE))
head(appren.p)
```

```
appren.p <- within(appren.p, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})
tail(appren.p)
summary(appren.p$PredictedProb)

appren.p <- cbind(appren.p, pred.chd = factor(iffelse(appren.p$PredictedProb >
  0.5, 1, 0)))
head(appren.p)

# Matrice de confusion
(m.confusion <- as.matrix(table(appren.p$pred.chd, appren.p$Y)))

m.confusion <- unclass(m.confusion)
# Taux d'erreur
Tx_err <- function(y, ypred) {
  mc <- table(y, ypred)
  error <- (mc[1, 2] + mc[2, 1])/sum(mc)
  print(error)
}
Tx_err(appren.p$pred.chd, appren.p$Y)

test.p <- cbind(test, predict(m.logit, newdata = test, type = "response", se = TRUE))
test.p <- cbind(test.p, pred.chd <- factor(iffelse(test.p$fit > 0.5, 1, 0)))
(m.confusiontest <- as.matrix(table(test.p$pred.chd, test.p$Y)))

m.confusiontest <- unclass(m.confusiontest)
# calcul du taux d'erreur sur l'échantillon test
Tx_err(test.p$pred.chd, test.p$Y)

require(ROCR)
```

```

Pred = prediction(appren.p$PredictedProb, appren.p$Y)
Perf = performance(Pred, "tpr", "fpr")
plot(Perf, colorize = TRUE, main = "ROC apprentissage")

perf <- performance(Pred, "auc")
perf@y.values[[1]]

Predtest = prediction(test.p$fit, test.p$Y)
Perftest = performance(Predtest, "tpr", "fpr")
perftest <- performance(Predtest, "auc")
perft <- performance(Predtest, "auc")
plot(perft, colorize = TRUE, main = "ROC test")
perftest@y.values[[1]]

par(mfrow = c(1, 2))
plot(Perf, colorize = TRUE, main = "ROC apprentissage - AUC= 0.8")
plot(Perftest, colorize = TRUE, main = "ROC Test")
require(ggplot2)
pl=function(a){
  ggplot(appren.p, aes(x = a, y = PredictedProb)) + geom_ribbon(aes(ymin = LL,
ymax = UL, fill = a), alpha = 0.15) + geom_line(aes(colour = a), size = 1)}

hist(table(bm))

# histogramme
library(lattice)
par(mfrow = c(1, 3))
histogram(~ Durée.de.remboursement | Y , data = b, type="percent", col="grey", breaks=15)
histogram(~PROD.BRUTE | Y , data = b, type="percent", col="grey", breaks=20)
histogram(~revenu | Y , data = b, type="percent", col="grey", breaks=100)
histogram(~age | Y , data = b, type="percent", col="grey", breaks=15)

kruskal.test(b$age~b$Y)
kruskal.test(b$PROD.BRUTE~b$Y)
kruskal.test(b$Durée.de.remboursement~b$Y)
kruskal.test(b$revenu~b$Y)

```

```
# 1e solution pour le V de Cramer :
# utilisation du package rgrs ou questionr
library(rgrs)
install.packages("questionr")
library(questionr)
cramer <- matrix(NA,ncol(b3),3)
for (i in (1:ncol(b3)))
{   cramer[i,1] <- names(b3[i])
    cramer[i,2] <- cramer.v(table(b3[,i],b3$Y))
    cramer[i,3] <- chisq.test(table(b3[,i],b3$Y))$p.value
}
colnames(cramer) <- c("variable","V de Cramer","p-value chi2")

vcramer <- cramer [order(cramer[,2], decreasing=T),]

b4=subset(b3, select = -c(Y))

cramer <- matrix(NA,ncol(b4),ncol(b4))
for (i in (1:ncol(b4)))
{   for (j in (1:ncol(b4)))
    {
    cramer[i,j] <- cramer.v(table(b4[,i],b4[,j]))
    }
}
colnames(cramer) <- colnames(b4)
rownames(cramer) <- colnames(b4)
library(corrplot)
corrplot(cramer)
corrplot(cramer, method="shade", shade.col=NA, tl.col="black", tl.srt=45)
par(omi=c(0.4,0.4,0.4,0.4))
corrplot(cramer,type="upper",tl.srt=45,tl.col="black",tl.cex=1,diag=F,addCoef.col="black",
addCoefasPercent=T)
```

Macros de calcul du tarif pour une marge fixée

```

Sub Macro1()

For i = 6 To 577

    Application.CutCopyMode = False
    Application.CutCopyMode = False
    Application.CutCopyMode = False
    Application.CutCopyMode = False

    Range("DV" & i).GoalSeek Goal:=0, ChangingCell:=Range("DX" & i)
Next i

End Sub

Sub fixe()

    b = InputBox("Entrer la marge d'interets à réaliser")
    Range("AA2").Value = b

    Application.CutCopyMode = False
    Application.CutCopyMode = False
    Application.CutCopyMode = False
    Application.CutCopyMode = False

    Range("AA3").GoalSeek Goal:=0, ChangingCell:=Range("AD3")

End Sub

```

Bouton 9				Recette	1 771 379,99	
tarif fixe	tarif/Dossier	taux mensuel	X	VPM	Mensualité	
6,45	3,80	0,32	1,00348732	257,87	257,868066	
6,45	2,29	0,19	1,00210065	275,44	275,437425	
6,45	4,24	0,35	1,00388743	2907,49	2907,49388	
6,45	3,80	0,32	1,00348732	3610,00	3609,99819	
6,45	1,58	0,13	1,00145062	2124,12	2124,1199	
6,45	4,57	0,38	1,00418847	6918,65	6918,6471	
6,45	4,57	0,38	1,00418847	2555,00	2555,00208	
6,45	3,80	0,32	1,00348732	3357,34	3357,34164	
6,45	2,29	0,19	1,00210066	1273,90	1273,89822	

FIGURE A.1 – Extrait de la table de calcul du tarif par dossier pour une marge d'intérêt d'équilibre

Bibliographie

Rapports annuels de Bank Al-Maghrib sur le contrôle, l'activité et les résultats des établissements de crédit.

Rapports annuels de l'association professionnelle des sociétés de financement.

Webographie

[http://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/0/573fe62d2e94b613c1257c1700454809/\\$FILE/Me%CC%81moire%20DEMICHELIS.002.pdf/Me%CC%81moire%20DEMICHELIS.pdf](http://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/0/573fe62d2e94b613c1257c1700454809/$FILE/Me%CC%81moire%20DEMICHELIS.002.pdf/Me%CC%81moire%20DEMICHELIS.pdf)

<https://books.google.co.ma/books?id=C8GfCgAAQBAJ&printsec=frontcover&dq=methode+d+apprentissage+statistique&hl=fr&sa=X&ved=0ahUKEwil3IyWlYDjAhWMlhQKHyoZDLQQ6AEIKTAA#v=onepage&q=methode%20d%20apprentissage%20statistique&f=false>

https://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf

<https://corporate.axa.ma/fr/>

http://www.casablanca-bourse.com/Documents/AXC/fr/NI_AXC.pdf_fr.pdf