Université de Lorraine

OTODOO-IECL

Rapport de stage

Simulation des prêts immobiliers et étude du risque de crédit

---

Soutenu le 28 septembre 2017

Rédigé par : Ameni KANTASSI
Étudiante en Master 2 IMOI

Encadré par : Pierre VALLOIS
Membre de : Institut Élie Cartan

Alban CUNIN
Président de Créditéo

Table des matières

0.1 Introduction

Contexte :

• Aujourd'hui, lorsqu'un emprunteur veut obtenir un prêt, il peut contacter des banques directement, ou bien des courtiers. Le travail des courtiers est de négocier un meilleur prêt pour les clients en contactant de nombreuses banques et en obtenant les meilleurs taux.
Mais ces nombreuses demandes sont réglementées et nécessitent des procédures, et donc beaucoup de documents et de temps. C'est là que veut intervenir un simulateur des prêts immobiliers. En facilitant le travail des courtiers lors de la saisie de données, de la génération de documents et du transfert de ces documents directement aux banques, le logiciel ou le simulateur permettra aux courtiers d'économiser plus de 1000 d'heures de travail par an pour un cabinet de crédit immobilier moyen.

Un logiciel pour les simulation des différents crédits immobiliers constitue un outil très utilisé par les organismes bancaires ainsi par les cabinets de courtage, un autre outil bancaire très important dans nos jours est la gestion de risque de crédit.
Bon ou mauvais payeur, futur cancéreux ou non, client potentiel ou non,...Les banquiers, épidémiologistes, économistes et bien d'autres professions sont souvent menés à faire des prédictions pour prendre des décisions appropriées à leur objectifs.
Encore aujourd'hui, seules les banques et institutions financières de premier plan sont capable d'évaluer le risque de crédit avec un certain degré de confiance ou disposent d'une base de données fiable pour le «scoring».
L'analyse statistique leur permet de réaliser ces prédictions. Parmi toutes les méthodes qu'elle contient, la régression logistique s'est largement imposée par sa simplicité de mise en place d'utilisation et d'interprétation.

Objectif :

• Dans le cadre de ce mémoire, notre travail consiste, d'une part à développer les formules mathématiques des prêts immobiliers pour le logiciel "créditéo", qu'on le présentera dans le chapitre 1, et d'autre part à la mesure de risque de crédit par une notation statistique des emprunteurs. On y développe une méthode paramétrique, à savoir la régression logistique.

chapitre 1

Présentation de l'entreprise et concept de scoring

Ce premier chapitre de mon travail est consacré d'une part à présenter l'entreprise **Otodoo** dans laquelle j'ai effectué mon stage académique.
D'autre part, à présenter le concept de risque de crédit et particulièrement le risque de crédit, la raison d'être de notre travail.

1.1 Présentation de l'entreprise

1.1.1 L'entreprise

Otodoo, Société par Actions Simplifiée (SAS) est en activité depuis 11 ans. Installée à Vandœuvre-lès-Nancy (54500), elle est spécialisée dans le secteur d'activité de la programmation informatique. Sur l'année 2015 elle réalise un chiffre d'affaires de 169.100,00 €. Le total du bilan a augmenté de $15,57$ % entre 2014 et 2015.

Otodoo est une agence d'architecture digitale. Elle crée des outils informatiques permettant d'améliorer l'image, la productivité et le chiffre d'affaires de ses clients. Sa politique est qu'un bon projet doit être composé de $\frac{1}{3}$ de design, $\frac{1}{3}$ d'ingénierie,$\frac{1}{3}$ de marketing. Elle propose de nombreuses prestations dans la création de sites web comme la création de sites web 'wordpress', la création de cahier des charges, la création de chartes graphiques ou encore la création et la gestion de contenus

Aussi, **Otodoo** propose des services de communication à ses clients comme par exemple

du ' community ' management, des campagnes de communications avec photos et vidéos et du référencement de sites web. Par ailleurs, l'entreprise peut être amenée à développer de plus gros projets logiciels tels que des logiciels de gestion d'entreprises, des applications mobiles ou des applications web. En plus de toutes ces prestations, **Otodoo** propose aussi des formations aux outils numériques et de la maintenance informatique à ses clients

Les locaux d'**Otodoo** sont situés au 9 square de liège à Vandœuvre-lès-Nancy lès Nancy-lès-Nancy-lès-Nancy-lès-Nancy (54500). Il y a deux bureaux de type 'open space', l'un au 4ème étage et l'autre au 13ème. Au 4ème étage se trouve l'équipe de création de sites web et l'équipe commerciale. Et au 13ème l'équipe technique, responsable de la création d'applications. Mon stage a pris place avec cette équipe technique.

1.1.2 L'équipe

Otodoo emploie 10 salariés. Pendant mon stage, il y a également eu 3 stagiaires. L'équipe d'**Otodoo** est répartie en 4 pôles : commercial, administratif, web et technique.

Le pôle commercial est constitué de M. Raymond Gilles, le directeur d'**Otodoo** et co-fondateur de **Créditéo** et de M. Perrin Benjamin . Leur rôle est de démarcher des entreprises lors de salons par exemple, dans l'optique de signer des devis.

L'administration est gérée par Mlle. Toussaint Lolita . C'était avec elle que je gérais les sujets administratifs.
La pôle web comporte 3 personnes :
– Fred Jaillet, webmaster wordpress.
– Laurent Paris, web designer.
– Mylène Eisenberg, community manager.
Enfin, le pôle technique est constitué de 4 personnes :
– Romain Houpin, Chef de projet/d'équipe.
– Joffrey Dalencon, développeur web.
– Cédric Enclos, Ingénieur informatique.
– Stéphane Helbling, Développeur informatique.
Pendant toute la durée de mon stage, j'ai également été amené à travailler avec M. Cunin Alban.
M. Cunin Alban : Intermédiaire en opérations de banque et services de paiement (IOBSP) = courtier en crédit immobilier.

Pendant mon stage, M. Cunin a consacré son temps et sa expérience à m'expliquer plein de notions de base en courtage, particulièrement le fameux lissage des prêts et son intérêt dans le secteur des banques, le fonctionnement des prêts réglementés, les simulateurs de crédits immobiliers les plus utilisées en France (Altofice, Courtisia, Calcamo ..). Il m'a aidé à élaborer le projet d'évaluation du risque de crédit en mettant en disposition sa base de données de ses clients, des informations supplémentaires et ses conseils constructifs.

L'équipe **Otodoo** est très agréable, jeunes, serviable et dynamique. J'ai senti très à l'aise en travaillant avec eux.

1.1.3 Les différents projets en cours

Lors de mon stage, l'entreprise Otodoo ne développait pas que Créditéo. Le pôle technique travaillait principalement sur :

- Gestion : Application permettant la gestion des contacts, articles, factures, devis, commande, comptes bancaires, etc. (ERP d'Otodoo)
- TFN : Application Web permettant la gestion du planning des travaux de nettoyage à effectuer de la société TFN Propreté.
- Eclatec : Application multiplateforme (Android, iOS, Windows App) permettant de visualiser les produits de l'entreprise à l'aide de la réalité augmentée.
- Helio service : Plateforme d'achat et de gestion des commandes pour un client de l'imprimeur Helio Service.

Aussi, Otodoo a encore de nombreux projets pour l'avenir. Le pôle web travaille sur un gros site d'e-commerce nommé Plaisance et aussi des sites pour des salons tel que le Salon entreprise Lorraine et Portail des Réseaux. Non seulement ces activités montrent que la société se porte bien, mais permettront une visibilité accrue d'Otodoo (plus de 50 entreprises sur Salon entreprise Lorraine par exemple).
Dans le futur, l'équipe technique d'Otodoo va achever les projets en cours pour se concentrer sur Créditéo, que nous le présenterons dans la section suivante.

1.2 Le projet Créditéo

L'idée du projet Créditéo a émergé après la rencontre de M. Cunin Alban avec M. Raymond Gilles en novembre 2015 dans le club d'affaires BNI (Business Network International). Dans

ce club de recommandations il est préconisé de rencontrer chaque semaine l'un de ses comembres afin de réaliser un « tête à tête » pour mieux apprendre son activité et pouvoir le recommander. En cherchant à recommander Otodoo, M. Cunin Alban a eu l'idée de travailler avec M. Raymond Gilles sur un projet de création de logiciel destiné aux courtiers. Cette idée lui est venue car il a utilisé au cours de ses 5 années d'expériences dans le métier les deux seuls logiciels existants (Altoffice et Courtisia). Sa conclusion était sans équivoque : l'offre proposée ne le satisfaisait pas et il était convaincu qu'on pouvait faire mieux. Il a donc cherché comment innover dans son secteur.

Après 6 mois de réflexions,M. Gilles Raymond et M. Cunin Alban ont décidé de s'associer et de créer l'entreprise Créditeo en mai 2017. Cette entreprise est une SAS dont le président est M. Alban Cunin. C'est une startup qui développe et commercialise un service en mode SAAS (software as a service). Ce service est le logiciel Créditéo, qui sera donc loué par les entreprises lorsqu'il sera mis en vente. Le projet est porté par M. Alban Cunin, courtier en crédit immobilier, M. Gilles Raymond, le directeur d' Otodoo et M. Gilles Caumont, le président d'Adista.

Après une phase de rédaction complète du cahier des charges, il a été décidé de confier la sous-traitance du développement à Otodoo. C'est dans ce contexte que je me suis retrouvé à travailler sur la partie simulation des algorithmes pour l'application Créditéo dans l'équipe d'Otodoo.

Pourquoi ce projet ?

Aujourd'hui, lorsqu'un emprunteur veut obtenir un prêt, il peut contacter des banques directement, ou bien des courtiers. Le travail des courtiers est de négocier un meilleur prêt pour les clients en contactant de nombreuses banques et en obtenant les meilleurs taux.
Mais ces nombreuses demandes sont réglementées et nécessitent des procédures, et donc beaucoup de documents et de temps. C'est là que veut intervenir Créditéo. En facilitant le travail des courtiers lors de la saisie de données, de la génération de documents et du transfert de ces documents directement aux banques, le logiciel permettra aux courtiers d'économiser plus de 1000 d'heures de travail par an pour un cabinet de crédit immobilier moyen.
Et le marché est important. Plus de 12000 courtiers et 60000 agents immobiliers pourraient être intéressés. En revanche, il existe des concurrents sur ce marché. Il existe aujourd'hui deux acteurs majeurs : Altoffice (1991) le leader, et Courtisia (2013) le challenger. Altoffice possède beaucoup des parts du marché, ainsi que de nombreuses fonctionnalités, en plus d'être précis sur ses calculs. En revanche, ses graphismes deviennent obsolètes, et il ne propose ni une version en ligne, ni de système de transmission de dossiers automatique. Courtisia quant

a lui a l'avantage d'être en ligne, mais il n'a pas aucun automatisme et est moins précis au niveau des calculs.

Créditéo a pour ambition de les détrôner en fournissant une version en ligne avec de meilleures fonctionnalités. La saisie des données sera simplifiée avec un système d'OCR (optical character recognition), il signifie reconnaissance optique de caractères ou reconnaissance de texte et il fera gagner du temps à ses utilisateurs en automatisant certains transfert de fichiers aux banques.

Les fonctionnalités :
Les objectifs principaux du projet sont de réduire l'ensemble du travail du courtier et automatiser ses tâches. Le courtier ne doit pas avoir besoin d'utiliser autre chose que son logiciel pour tout faire et peu importe l'endroit où il se trouve (A la banque, chez le client, dans son bureau). Le logiciel se doit d'être beau et rapide, pour être utilisable devant les emprunteurs et asseoir la crédibilité du professionnel devant eux.
Les fonctionnalités principales seront :
- Un système d'agenda pour le courtier
- Un carnet de contacts regroupant clients, courtiers, apporteur d'affaires, agent en banque ...
- La gestion de dossier prospects.
- La gestion de dossiers clients.
- La gestion du système de facturation.
- La gestion des relations avec les banques, les apporteurs d'affaires et les interlocuteurs en agence.
- La création de simulations dans des dossiers.
- L'édition des documents nécessaires aux banques (ex : demande de prêt).
- L'automatisation de la transmission de ces documents.
- Un système simplifié de saisie de dossier.

La partie sur laquelle j'ai travaillé est la simulations des formules mathématiques et barèmes réglementés en utilisant le logiciel **Matlab**.

Pourquoi une partie simulation ?
Pour un courtier, utiliser un outil de simulation est un énorme gain de temps. En effet, les calculs impliqués sont lourds et parfois impossible à faire à la calculatrice (lissage de prêts par exemple). L'utilisation d'un tel outil permet de présenter des résultats rapidement et plus visuellement à ses clients. Aussi, utiliser un logiciel interactif permet de modifier certaines

données selon les désirs du client, et d'obtenir les résultats instantanément. En plus, il est possible de générer une présentation en pdf d'une simulation, qu'il sera facile à montrer à une banque.

Enfin, il est possible de sauvegarder ces simulations dans un dossier client ce qui facilite la gestion en général (on ne perd pas son travail).

FIGURE 1.1 – Diagramme de cas d'utilisation de la page de simulation

1.3 Concept de risque bancaire

Dans le domaine de la banque, une problématique très importante est celle de la gestion du risque. En effet, à partir de leurs activités, les banques sont confrontées à de nombreux risques, et un des défis les plus importants de leur secteur est de pouvoir établir un contrôle entre le risque à prendre et le gain espéré.

Elle gère toute une topologie des risques parmi lesquels :

- Le risque de crédit ou de contrepartie : Lorsqu'une banque octroie un prêt à un client, elle doit s'assurer que celui-ci soit capable de rembourser la somme empruntée. Sinon, le client se trouve en défaut et la banque doit prendre à sa charge les pertes liées à ce contrat.
- Les risques de marché : générés par les activités de marché (taux, change, perte de valeur d'instruments financiers).
- Les risques opérationnels : qui désignent les risques de pertes ou de sanctions notamment du fait de défaillances des procédures, d'erreurs humaines, d'événements extérieurs.
- Le risque de liquidité : Les banques reçoivent majoritairement des dépôts à court terme de leurs clients et font des prêts à moyen et long terme. Il peut donc se créer un décalage entre les sommes prêtées et les sommes disponibles (dépôts), ces dernières peuvent être insuffisantes. Dans ce cas on parle de manque de liquidités et le risque de liquidité représente l'éventualité de ne pas pouvoir faire face, à un moment donné, à ses engagements ou à ses échéances.

Un outil très important pour la gestion de risque de crédit est le **Scoring**, qu'on détaillera dans la section suivante.

1.4 Le crédit scoring

La crise financière qui secoue le monde, notamment les défaillances successives des grandes banques internationales ont remis sur le devant de la scène la problématique des risques bancaires dont le risque de crédit, on peut prendre à titre d'exemple la crise des **subprimes** en 2007 liée au problème du non remboursement des crédits immobiliers aux États-Unis.

Plusieurs travaux de recherches ont été réalisés pour qualifier à l'avance les emprunteurs qui seront des mauvais payeurs de ceux qui ne le seront pas. les établissements bancaires font ainsi appel aux méthodes statistiques afin de modéliser le risque qu'elles encourent, méthodes parmi lesquelles figure le scoring.

Les premiers travaux sur le scoring ont été entrepris aux États-Unis d'Amérique dans les années 1960, notamment par Altman (1968), Haldeman et Nerayanan (1977), etc.

1.4.1 Définition du crédit scoring

Le crédit scoring est un ensemble d'outils d'aide à la décision utilisés par les organismes financiers pour évaluer le risque de non-remboursement des prêts. C'est une note de risque ou une probabilité de défaut, en passant par la mise sur pied d'un modèle statistique, ceci passe par un travail de synthèse d'une grande masse de données collectée dans le passé.

Plusieurs explications peuvent être fournies sur la définition et le rôle du Credit Scoring, elles peuvent être résumées comme sur la figure ci-dessous :

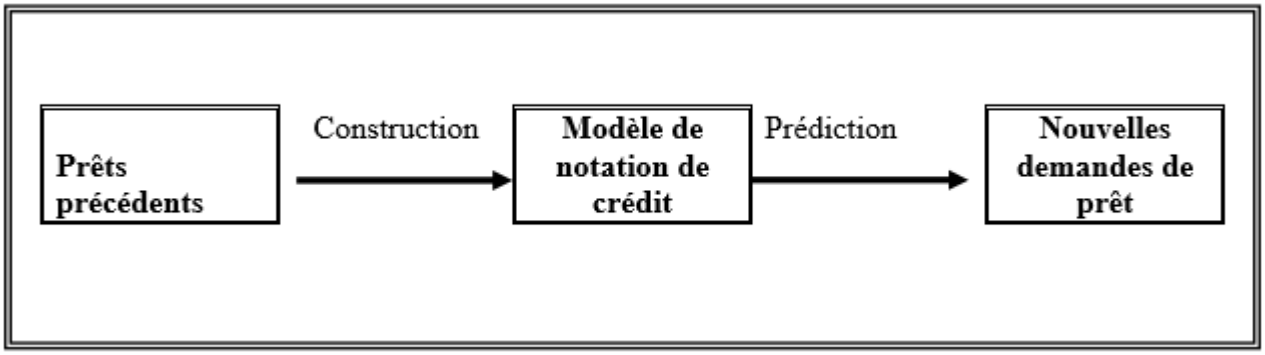


FIGURE 1.2 – processus du crédit scoring

Ce processus du crédit scoring sera plus détaillé dans le chapitre 3.

chapitre 2

Présentation et simulation des prêts immobiliers

Après avoir présenté dans le chapitre 1 le projet **Créditio**, le concept de risque de crédit et le scoring, le but de ce chapitre sera consacré à présenter une partie de travail élaborée pour Créditio

Le crédit immobilier, ou prêt immobilier est un emprunt destiné à financer tout ou une partie de l'acquisition d'un bien immobilier, de l'opération de construction, ou des travaux sur un tel bien.
Notre but est l'étude et la simulation (sous le logiciel Matlab) des différents types de prêts immobiliers , ses options et l'étude de lissages des prêts que nous le détaillerons dans la partie 2.2.

2.1 Prêts immobiliers bancaires

Lors de la réalisation d'un projet immobilier, un ménage peut faire appel à de nombreux financements. S'offrent à lui différents types de prêts immobiliers en fonction de sa situation, de son niveau de vie et de son projet. Nous les présenterons dans la partie qui suit.

2.1.1 Prêt immobilier amortissable à taux fixe

Le crédit à taux fixe ou constant propose un taux débiteur qui ne varie pas durant sa durée de remboursement, évidement des échéances de remboursement constantes ou modulables (revues à la hausse ou à la baisse à la demande de l'emprunteur et en fonction de ses possibilités).

2.1.1.1 Intérêt à taux fixe

Pour un emprunt bancaire, noté **C** de nominal **N** et pour un taux d'intérêt mensuel, fixé par les organismes bancaires, qu'on le notera **t/100**, l'intérêt à taux fixe **I** est la rémunération d'un prêt, sous forme généralement d'un versement mensuel de l'emprunteur au prêteur. Pour l'emprunteur, c'est un coût correspondant à une utilisation anticipée de l'emprunt (exprimée en pourcentage du montant prêté)

$$\boxed{I = \frac{C \times t \times N}{100}.} \tag{2.1}$$

Alors,

$$\boxed{t = \frac{I \times 100}{C \times N}.} \tag{2.2}$$

2.1.1.2 Mensualité ou échéance :

Une mensualité notée **M**, est la somme d'argent payée chaque mois par l'emprunteur à la banque lors d'un prêt contracté auprès d'une banque.
Dans un prêt amortissable, l'emprunteur rembourse chaque mois une partie du capital emprunté et paie des intérêts à la banque sur le capital emprunté, le tout constituant la mensualité calculée avec la formule suivante :

$$\boxed{M = \frac{C \times \frac{t}{100}}{1 - (1 + \frac{t}{100})^{-N}}.} \tag{2.3}$$

Démonstration A l'origine, la banque prête le client un capital C.
· Un mois plus tard, il verse M. Cette somme sert avant tout à rétribuer la banque pour lui avoir prêté C sur un mois, il lui versez donc un intérêt $I_1 = C \times \frac{t}{100}$.
Le reste de la mensualité, soit $M - I_1$, sert à rembourser la banque. Après quoi il ne lui serez donc plus redevable que de :

$$C_1 = C - (M - I_1) = C - (M - (C \times \frac{t}{100})) = C \times (1 + \frac{t}{100}) - M.$$

(Vous remarquerez que par principe on veut rembourser un peu de capital, la mensualité doit donc être supérieure à l'intérêt, c'est-à-dire que $M > C \times \frac{t}{100}$).

· Le deuxième mois, il verse encore M, dont une part rétribue la banque par un intérêt $I - 2 = C_1 \times \frac{t}{100}$, il rembourse $M - I_2$ de capital et il lui reste à rembourser que :

$$\begin{aligned} C_2 &= C_1 - (M - I_2) \\ &= C_1 \times (1 + \frac{t}{100}) - M \\ &= (C \times (1 + \frac{t}{100}) - M)(1 + \frac{t}{100}) - M \\ &= C(1 + \frac{t}{100})^2 - M(1 + \frac{t}{100}) - M. \end{aligned}$$

· Et ainsi de suite. Le i-ème mois, il paye un intérêt $I_i = C_{i-1} \times \frac{t}{100}$ et il rembourse $M - I_i$, il reste une dette de $C_i = C_{i-1}(1 + \frac{t}{100}) - M$.
En continuant le processus (par procédure de récurrence), on trouve :

$$\begin{aligned} C_i &= C(1 + \frac{t}{100})^i - M(1 + \frac{t}{100})^{i-1} - M(1 + \frac{t}{100})^{i-2} - ... - M \\ &= C(1 + \frac{t}{100})^i - M[(1 + \frac{t}{100})^{i-1} + (1 + \frac{t}{100})^{i-2} + ... + (1 + \frac{t}{100}) + 1] \\ &= C(1 + \frac{t}{100})^i - M \times \frac{(1 + \frac{t}{100})^i - 1}{\frac{t}{100}}. \end{aligned}$$

Revenons au principe de notre prêt, l'emprunteur doit le rembourser en N mois, on veut donc :

$$C_N = 0 = C(1 + \frac{t}{100})^N - M \times \frac{(1 + t)^N - 1}{\frac{t}{100}}.$$

Finalement, si on connaît C, t, N , on trouve la mensualité :

$$M = \frac{C \times \frac{t}{100}}{1 - (1 + \frac{t}{100})^{-N}}.$$

□

La mensualité peut tenir en compte des autres frais, à savoir les frais de l'assurance et de la garantie.

2.1.2 Prise en compte de l'assurance et de la garantie

2.1.2.1 L'assurance

L'assurance de prêt immobilier est définie dans le code des assurances sous l'appellation « assurance emprunteur ». C'est un contrat obligatoire en France dès lors que vous contractez

un crédit immobilier.

L'assurance emprunteur fait partie de la famille des assurances de Prévoyance. C'est à dire qu'elle vous couvre en cas de décès, d'accident ou encore de maladie.

L'assurance de prêt immobilier garantit, selon les garanties souscrites, le remboursement des échéances du prêt en cas de :

– Décès ou perte totale et irréversible d'autonomie (PTIA).
– Invalidité permanente et totale (IPT) ou partielle (IPP).
– Incapacité temporaire et totale de travail (ITT).
– Perte de votre emploi (en option).

En 2017, les taux d'une assurance de crédit immobilier bancaire ont subi une très légère augmentation par rapport à 2016, augmentation qui reste modérée.

Voici le tarif moyen des différentes banques en 2017 :

TAEA (prêt au taux nominal de 2,5% d'une durée de 20 ans) selon l'âge		**-30**	**31-35**	**36-40**	**41-45**	**46-50**	**51-55**	**56-60**	**61-65**
Crédit Agricole (1)	Min	0,36%	0,36%	0,60%	0,60%	0,74%	0,74%	0,88%	0,88%
	Max	0,51%	0,51%	0,71%	0,71%	0,87%	0,87%	1,03%	1,03%
Caisse d'Epargne (2)	Min	0,36%	0,36%	0,36%	0,47%	0,52%	0,59%	0,64%	0,71%
	Max	0,47%	0,59%	0,71%	0,71%	0,71%	0,87%	0,87%	1,03%
Banque Populaire		0,44%	0,44%	0,61%	0,61%	0,81%	0,81%	1,00%	1,00%
BNP (3)	Min	0,34%	0,37%	0,47%	0,54%	0,66%	0,87%	1,13%	2,41%
	Max	0,37%	0,42%	0,51%	0,59%	0,69%	0,97%	1,26%	2,57%
LCL (3)	Min	0,41%	0,41%	0,59%	0,61%	0,67%	0,97%	1,00%	2,03%
	Max	0,46%	0,46%	0,67%	0,77%	0,77%	1,05%	1,00%	2,03%
Société Générale		0,42%	0,42%	0,59%	0,59%	0,76%	0,76%	0,84%	0,84%
Crédit du Nord		0,42%	0,42%	0,57%	0,57%	0,77%	0,77%	0,84%	0,97%
La Banque Postale		0,42%	0,42%	0,64%	0,76%	0,92%	0,92%	0,92%	1,29%
Crédit Foncier France		0,65%	0,65%	0,71%	0,71%	0,71%	0,81%	0,81%	1,49%
C.Mutuel 11 - CIC (4)	Min	0,33%	0,39%	0,59%	0,59%	0,79%	0,79%	0,96%	0,96%
	Max	0,43%	0,57%	0,75%	0,75%	0,79%	0,79%	1,45%	1,45%
HSBC		0,59%	0,59%	1,00%	1,00%	1,00%	1,16%	1,16%	1,16%
Tarif moyen		0,43%	0,46%	0,62%	0,65%	0,76%	0,82%	0,98%	1,23%

(1) : tarifs liés au niveau de couverture retenue par les co-emprunteurs (-20% en cas de double assurance à 100%) et aussi aux primo-accédants favorisés jusqu'à 35 ans sur les non-primo-accédants
(2) : tarif "public" et tarif avec concession commerciale maximale accordable.
(3) : prêts >15 ans pour la ligne la plus chère et prêts de 15 ans pour la ligne la moins chère.
(4) : la ligne chère correspond aux prêts de 25 ans et l'autre aux prêts de 15 ans.

FIGURE 2.1

Remarque 2.1. ***(Mensualité avec frais d'assurance)*** *En gardant les mêmes notations définies dans la sous-section 2.1.1.1, pour un prêt immobilier, on notera* $\frac{t_1}{100}$ *le taux d'assurance fixe (mensuel) prise par une banque, la mensualité avec frais d'assurance vaut alors :*

$$\boxed{M = C \times \left[\frac{\frac{t}{100}}{1 - (1 + \frac{t}{100})^{-N}} + \frac{t_1}{100} \right]} \tag{2.4}$$

2.1.2.2 La Garantie

Lors d'un achat immobilier, en contrepartie du prêt immobilier que la banque accorde, elle exige une garantie. Elles permettent à l'établissement prêteur de se protéger en cas de défaillance de l'emprunteur.
Concrètement, en cas de problèmes, ce dispositif juridique permet aux établissements prêteurs d'appréhender et de faire vendre le bien financé pour récupérer les fonds octroyés.
Plusieurs formes de garantie existent en France :

- les sûretés réelles : hypothèque, Privilège de Prêteurs de Deniers (PPD réservé à l'ancien)
- le cautionnement : c'est la garantie proposée par la société Crédit Logement.

L'hypothèque

Cette garantie traditionnelle est très répandue. Si l'emprunteur n'honore pas ses mensualités, elle donne le droit au créancier de faire saisir le logement et de le vendre au enchères afin de récupérer le montant des sommes dues. Le plus souvent, elle est utilisée pour des crédits destinés à financer des travaux de construction.
Elle doit être inscrite à la conservation des hypothèques par un notaire. Pour l'emprunteur, cet acte notarié entraîne un certain nombre de frais.
Ces frais se décomposent de la façon suivante :

- Les émoluments proportionnels du notaire, notés k_1 (Il s'agit de la rémunération du notaire proprement dite, $\frac{1}{3}$ du barème suivant) :

Première série	avant le 01/05/2016	après le 01/05/2016
De 0 jusqu'à 6500 ϵ	4 %	3.945 %
Au-delà de 6500 ϵ jusqu'à 17000 ϵ	1.65 %	1.627 %
Au-delà de 17000 ϵ jusqu'à 60000 ϵ	1.1 %	1.085 %
Au-delà de 60000 ϵ	0.825 %	0.814 %

- Les frais de formalités, k_2 : environ 250 ϵ

- La contribution de sécurité immobilière k_3 : (0.6 % du prix de vente), (c'est l'ex salaire du conservateur des hypothèques égal à 0.05 % +0.01 % pour l'inscription au bureau des

hypothèques)

– Les droits d'enregistrement, k_4 : les frais d'enregistrement sont de 0.715 % du montant de prêt garanti.

Remarque 2.2. ***(Mensualité avec les frais d'hypothèques)*** *La mensualité avec frais de garantie (hypothèque) vaut :*

$$\boxed{M = C \times \left[\frac{\frac{t}{100}}{1-(1+\frac{t}{100})^{-N}}\right] + K} \tag{2.5}$$

Avec $K = k_1 + k_2 + k_3 + k_4$.

Privilège de Prêteurs de Deniers (PPD)

Le privilège du prêteur de deniers fonctionne sur le même principe que l'hypothèque.
Elle permet au préteur de saisir le bien et de le vendre si l'emprunteur ne rembourse pas ses échéances de prêt.
En revanche, le PPD ne peut porter que sur des biens existants (de l'ancien ou du neuf dont la construction est complètement terminée ou des terrains).
Le privilège de prêteur de deniers fait l'objet d'un acte notarié et doit être inscrit à la conservation des hypothèques dans les deux mois qui suivent la vente. L'intérêt pour l'emprunteur, c'est qu'il est exonéré de taxe de publicité foncière. Cette garantie reste donc moins coûteuse qu'une hypothèque. Ces frais sont encaissés par le notaire lors de la signature de l'acte de vente.

Cette somme comprend les émoluments du notaire, on le notera $p_1(\frac{1}{3}$ du barème dans le tableau ci-dessous), les frais de formalités, p_2 (environ 250 ϵ) ainsi que la contribution de sécurité immobilière p_3 (0, 05 % du prêt majoré de 20 % pour l'inscription au bureau des hypothèques).
Le calcul des émolument est établi à partir du nouveau barème des frais de notaire en vigueur depuis le 17 février 2011 suite à la parution du décret $2011 - 188$ qui a revalorisé les émoluments variables du notaire.
Les émoluments variables du notaire sont calculés selon le tableau ci-dessous.

Première série	en pourcentage (hors TVA)
De 0 jusqu'à 6500 ϵ	4
Au-delà de 6500 ϵ jusqu'à 17000 ϵ	1.65
Au-delà de 17000 ϵ jusqu'à 60000 ϵ	1.1
Au-delà de 60000 ϵ	0.825

Remarque 2.3. ***(Mensualité avec les frais de PPD)*** *Prise en compte des frais de la garantie PPD, le calcul de mensualité vaut alors :*

$$\boxed{M = C \times \left[\frac{\frac{t}{100}}{1-(1+\frac{t}{100})^{-N}}\right] + P} \tag{2.6}$$

Avec $P = p_1 + p_2 + p_3$.

Le cautionnement :

Les banques sont de plus en plus nombreuses à accepter, comme garantie, les engagements **des sociétés de cautionnement mutuel,** par exemple **Crédit Logement**, ces établissement pratiquent le mutualisation des risques. L'emprunteur verse une somme proportionnelle au montant de son crédit sur un fonds garanti et en contrepartie, la société s'engage à payer les échéances si l'emprunteur est défaillant.
La contribution versée à la société de caution se divise en deux montants : une commission qui est définitivement acquise par l'organisme de caution et une contribution versée au fonds mutuel de garantie, qui peut, selon les établissements, être partiellement ou totalement reversée à l'emprunteur à la fin du crédit.
Cette garantie est plutôt intéressante car elle peut être souscrite pour des biens neufs ou anciens et ne nécessite aucuns frais de notaire.

Il existe 2 barèmes des frais de garantie Crédit Logement : **CLASSIC** et **INITIO**.

2.1.2.3 Capital restant dû :

Le capital restant dû est le montant du capital restant à rembourser par un emprunteur, à son créancier, à une date donnée.
Pour obtenir son capital restant dû, l'emprunteur doit avoir la somme initiale empruntée noté C, la durée de remboursement N, le taux d'intérêt mensuel t % et le nombre de période de remboursement passées, n (c'est à dire le numéro de mensualité ou de période actuelle).
Sa formule de calcul est la suivante :

$$\boxed{R = \frac{M \times (1-(1+i)^{-(N-n)})}{t}} \tag{2.7}$$

2.1.2.4 Taux d'intérêt moyen :

Considérons deux capitaux C_1 et C_2 engagés respectivement pendant T_1 et T_2 à des taux d'intérêts t_1 et t_2, le taux d'intérêt moyen noté t_{moy} est celui qui, appliqué aux deux capitaux, pendant respectivement les durées T_1 et T_2, donne le même intérêt I :
Calcul de I avec t_1 et t_2 :

$$I = \frac{C_1 t_1 T_1 + C_2 t_2 T_2}{100},$$

de même, I est égale à :

$$I = \frac{C_1 t_{moy} T_1 + C_2 t_{moy} T_2}{100},$$

alors, le taux moyen t_{moy} est égale à :

$$\boxed{t_{moy} = \frac{C_1 t_1 T_1 + C_2 t_2 T_2}{C_1 T_1 + C_2 T_2}.} \tag{2.8}$$

2.2 Prêt à paliers (Lissage de crédits) :

· Outre son crédit immobilier, le porteur d'un projet immobilier peut avoir contracté d'autres prêts auprès du même établissement de crédit ou auprès de divers établissements financiers. Ces charges dites « d'emprunt » vont venir grever le taux d'endettement de l'emprunteur, réduisant par là sa capacité d'emprunt et donc, sa capacité de remboursement mensuelle. Pour éviter une telle situation de sur-endettement de l'emprunteur, les banques peuvent recourir à ce que l'on appelle un « crédit lissé ».

· Un crédit lissé est une technique de prêt qui consiste à « adoucir » (d'où le terme « lisser ») le coût des charges d'emprunt mensuelles d'un emprunteur en lui offrant la possibilité de payer une mensualité unique pour tous ses prêts. Le crédit lissé est également appelé « crédit à paliers ». Ainsi, au lieu de payer des mensualités distinctes pour chaque prêt, l'emprunteur paiera mensuellement une mensualité globale et unique.

Le mode opératoire d'un crédit lissé :

· Le crédit lissé est appelé « prêt à paliers » en raison du fait qu'il crée des paliers de remboursement, c'est-à-dire en quelque sorte des « ponts » entre les mensualités du prêt principal et celles des prêts complémentaires. Le prêt principal étant le crédit immobilier et les prêts complémentaires, les crédits divers contractés et à rembourser simultanément avec le crédit immobilier.

· Pour que l'emprunteur obtienne du banquier le lissage de son crédit immobilier principal, il est impératif que les prêts complémentaires aient une durée d'emprunt inférieure au prêt principal. En outre, l'établissement de crédit exige que le nombre des prêts à lisser (prêt immobilier principal et prêts complémentaires) soit inférieur à 6.

Juxtaposition de deux prêts :

On cherche à emprunter deux prêts qui s'emboitent en s'étalant sur des durées différentes : un emprunt (complémentaire) sur une durée plus courte N_c(c : complémentaire), d'un montant E_c et à un taux d'intérêt t_c, exprimé en pourcentage, et l'autre (prêt principale) d'un montant E_p (p :principal) sur une durée plus longue N_p à un taux d'intérêt t_l (en pourcentage).
On note E le montant global à emprunter, $E = E_c + E_p$.
On distingue deux phases de remboursement successives :
$\gg$ Première période de 0 à N_c :
Le prêt court est remboursé sur cette période comme un prêt à taux fixé (défini dans la sous-section 2.1.1), et on a donc le montant M_c des mensualités versées pour rembourser ce prêt.

$$\boxed{M_c = \frac{E_c t_c}{1-(1+t_c)^{-N_c}}.} \tag{2.9}$$

Pendant cette première période, on rembourse aussi chaque mois une partie du second prêt. On notera M_p les mensualité correspondantes.
Ainsi, durant cette première période le montant total des mensualités est $M = M_c + M_p$.

$\gg$ Deuxième période de N_c à N_p : Après N_c mois, le premier emprunt est totalement remboursé, il ne reste plus que le résidu du premier partie non encore remboursée, au taux t_p.
On notera M_1 les mensualités dans cette deuxième période.

Proposition 2.1. ***(Lissage)*** *On parle de lissage, ou de mensualités lissées (voir figure 2.3), lorsque les mensualités sont identiques sur toute la durée de l'emprunt global, c'est-à-dire pendant les N_p périodes. Le remboursement constant sur N_p de M est défini par*

$$\boxed{M = M_c + M_p} \tag{2.10}$$

Telles que M_c définie par (2.9) et M_p par :

$$\boxed{M_p = \frac{t_p E_p (1+t_p)^{N_p} - M_c((1+t_p)^{N_R}-1))}{(1+t_p)^{N_p}-1}} \tag{2.11}$$

Avec $N_R = N_p - N_c$.

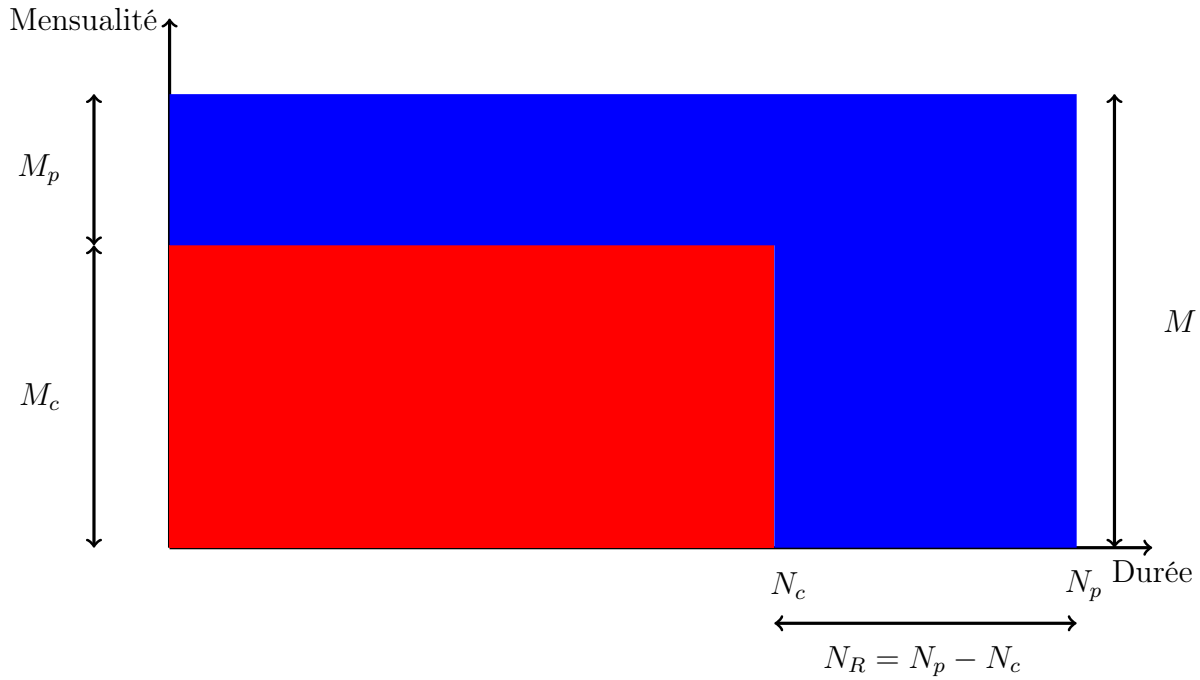


FIGURE 2.2 – lissage de deux prêts.

Démonstration Les mensualités pour le prêt court sont, comme détaillé dans sous-section 2.1.1 :

$$M_c = \frac{E_c t_c}{1 - (1 + t_c)^{-N_c}}.$$

Concernant le prêt long, la première question est de savoir, en versant des mensualités M_p, quelle sera la part de capital remboursée pendant la première partie d'une durée N_c.

On reprend pour cela le raisonnement et les calculs de la partie 2.1.1 (pour calcul de mensualité) :

· 1^{er} mois :

Le premier montant des intérêt versés est $I_1 = E_p t_p$, le montant remboursé est $M_p - I_1$ et donc le capital restant à rembourser est :

$$\begin{aligned} C_1 &= E_p - (M_p - I_1) \\ &= E_p - M_p + E_p t_p \\ &= E_p(1 + t_p) - M_p. \end{aligned}$$

· 2^{er} mois :

Le montant des intérêts versés est maintenant $I_2 = C_1 t_p$, et le montant restant à rembourser après le deuxième mois est

$$\begin{aligned} &= C_1 - M_p + C_1 t_p \\ &= C_1(1 + t_p) - M_p. \end{aligned}$$

En remplaçant C_1 par sa expression calculée précédemment, on aura :

$$\begin{aligned} C_2 &= C_1(1+t_p) - M_p \\ &= [E_p(1+t_p) - M_p](1+t_p) - M_p. \\ &= E_p(1+t_p)^2 - M_p(1+(1+t_p)). \end{aligned}$$

. . .

· $\underline{i^{eme} \text{ mois :}}$

en procédant par récurrence, on trouve que le capital restant du à la *ime* mois est

$$C_i = E_p(1+t_p)^i - M_p\left(1+(1+t_p)+(1+t_p)^2+...+(1+t_p)^i-1\right).$$

En utilisant la somme des termes d'une suite géométrique de raison $q = 1+t_p \neq 1 \Longleftrightarrow t_p \neq 0$, on obtient,

$$C_i = E_p(1+t_p)^i - M_p\frac{(1+t_p)^i-1}{t_p}. \tag{2.12}$$

Au mois N_c, on termine de rembourser le prêt court et à partir du mois N_c+1 on ne s'occupe plus que du l'autre prêt (principal). Pour celui-ci, il reste alors un capital au mois N_c,

$$C_{N_c} = E_p(1+t_p)^{N_c} - M_p\frac{(1+t_p)^{N_c}-1}{t_p}, \tag{2.13}$$

à rembourser pendant une durée $N_R = N_p - N_c$, évidement avec des mensualités M (de même que pour les mensualités M_c)

$$\boxed{M_{=}\frac{C_{N_c}t_p}{1-(1+t_p)^{-N_R}}.} \tag{2.14}$$

Il reste à déterminer l'expression de M_p, on a $M = M_p + M_c$, d'après l'expression (2.14) et (2.13), on aura :

$$\begin{aligned} M_p + M_c &= \frac{C_{N_c}t_p}{1-(1+t_p)^{-N_R}} \\ &= \frac{t_p}{1-(1+t_p)^{-N_R}}\left[E_p(1+t_p)^{N_c} - M_p\frac{(1+t_p)^{N_c}-1}{t_p}\right], \end{aligned}$$

donc

$$M_p\left[1+\frac{(1+t_p)^{N_c}-1}{t_p}\times\frac{t_p}{1-(1+t_p)^{-N_R}}\right] = \frac{t_p}{1-(1+t_p)^{-N_R}}E_p(1+t_p)^{N_c} - M_c.$$

Alors

$$M_p\left[\frac{(1+t_p)^{N_c}-(1+t_p)^{-N_R}}{1-(1+t_p)^{-N_R}}\right] = \frac{t_pE_l(1+t_p)^{N_c} - M_c(1-(1+t_p)^{-N_R})}{1-(1+t_p)^{-N_R}}.$$

D'où

$$M_p = \frac{t_p E_p (1+t_p)^{N_c} - M_c(1-(1+t_p)^{-N_R})}{(1+t_p)^{N_c} - (1+t_p)^{-N_R}}$$
$$M_p = \frac{t_p E_p (1+t_p)^{N_p} - M_c((1+t_p)^{N_R} - 1)}{(1+t_p)^{N_p} - 1}.$$

□

Généralisation : lissage de n prêts

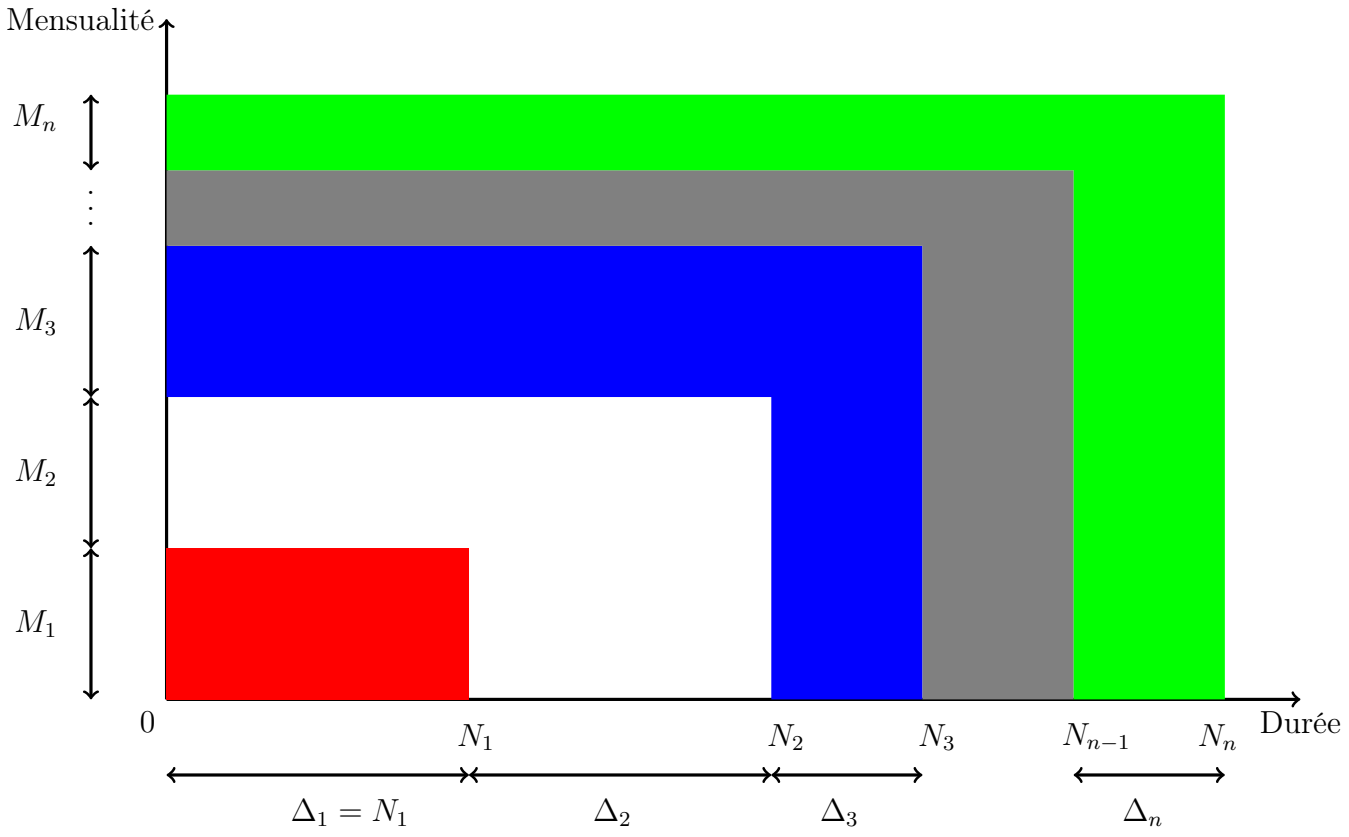


FIGURE 2.3 – lissage de n prêts.

On note pour chaque prêt,

– E_i, le montant du i-ème emprunt.
– t_i, le taux d'intérêts mensuel du i-ème emprunt.
– N_i, la durée, en mois, du i-ème emprunt.

En suivant la même démarche suivi dans la démonstration de la proposition 2.1, la mensualité pour le premier emprunt vaut :

$$M_1 = \frac{E_1 t_1}{1-(1+t_1)^{-N_1}}.$$

Et pour les $n-1$ prêts,

$$M_i = \frac{t_i E_i (1+t_i)^{N_i} - M'((1+t_i)^{-\Delta_i} - 1)}{(1+t_i)^{N_i} - 1},\ i \in \{2, ..., n\}$$

Telle que $M' = M_1 + M_2 + ... + M_i$ et $\Delta_i = N_i - N_{i-1}$.

Avantages et inconvénients du crédit lissé :

· Le grand avantage du crédit lissé est d'offrir une décote globale sur l'ensemble des mensualités et charges d'emprunt de l'emprunteur. Ceci permettra à l'emprunteur de pouvoir honorer ses dettes, dont le crédit immobilier.

· Le principal inconvénient du prêt à paliers et le sur-coût du crédit. Autrement dit, l'octroi du lissage de crédit immobilier entraîne au profit du banquier la perception d'intérêts supplémentaires.

Dans le paragraphe qui suit, on étudiera d' autres types de prêts, à savoir le prêt in fine, le prêt amortissable avec différé total et partiel, le prêt à taux variable, etc.

2.3 Prêt In fine

Contrairement au prêt amortissable où chaque mensualité permet de rembourser une partie de la somme emprunté, les mensualité d'un crédit in fine ne comporte que les intérêts (et l'éventuelle cotisation d'assurance). Le capital restant dû ne diminuent pas, les intérêts sont calculés sur le capital initial (emprunté), et ce montant reste le même pendant toute la période du l'emprunt. Ce n'est qu'à la dernière mensualité que le capital emprunté est intégralement remboursé.

Au final, toutes caractéristiques égales par ailleurs, un prêt in fine génèrera plus d'intérêts débiteurs qu'un crédit amortissable classique.

Conclusion 2.1. *Pour un prêt In fine noté C, a comme taux d'intérêt mensuel t en pourcentage, N comme nombre de mensualités et t_1 (en pourcentage) le taux d'assurance prise en compte alors le flux de remboursement et comme suit :*

· *Pour $1 \leq i \leq N-1$, La mensualité M_i vaut :*

$$M_i = \frac{C(tN + t_1)}{100}.$$

· *Pour la Nième mensualité,*

$$M_N = C\left[1 + \frac{(tN + t_1)}{100}\right].$$

• Les deux graphiques suivants représentent le flux des remboursements dans le cas d'un prêt amortissable et dans celui d'un prêt in fine.

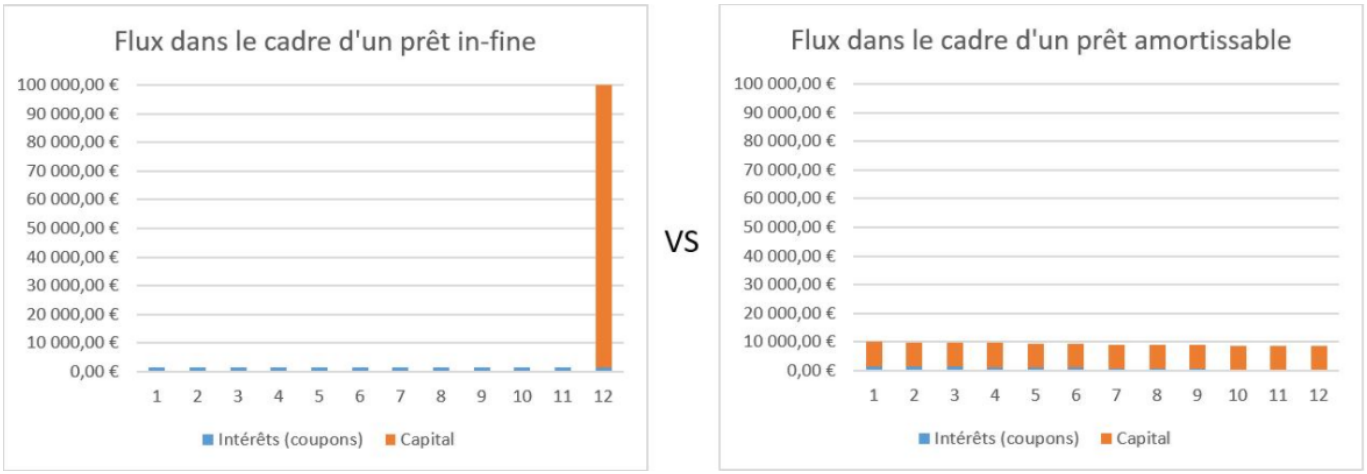


FIGURE 2.4 – Le flux de remboursement

2.4 Prêt à taux variable

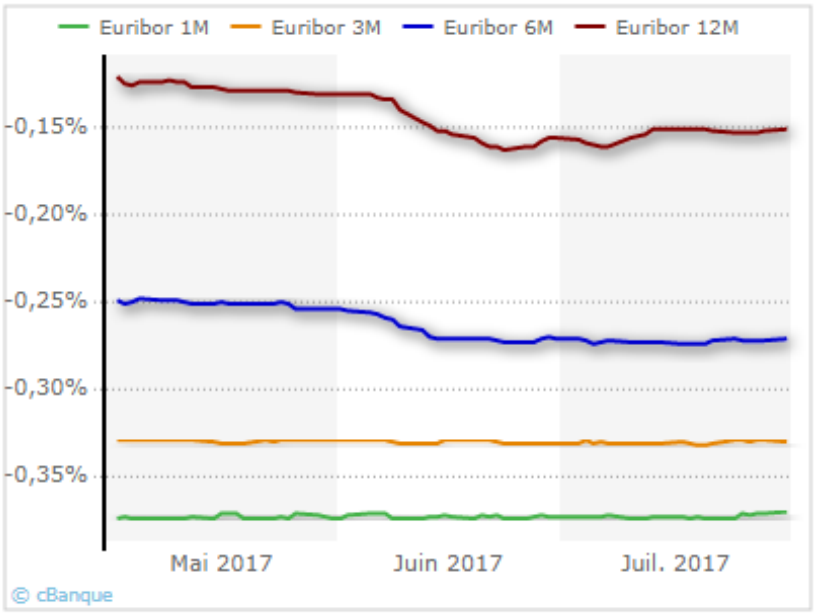
Le prêt à taux variable est révisé périodiquement, en général chaque année à la date de l'anniversaire du prêt, en fonction de l'évolution d'un indice de référence, cet indice est l'Euribor (Euro Interbank Offred Rate), qui est largement utilisé en Europe.

L'Euribor correspond au prix auquel les banques se prêtent des fonds à court terme sur le marché monétaire. Pour un emprunteur, les banques appliqueront cet indice de référence et elles y ajouteront une marge de 1 à 3 %, en fonction de sa situation personnelle. Bien entendu, plus il offre des garanties (revenus récurrents, bonne capacité d'épargne, emploi stable, qualité de votre acquisition,...), plus la marge que prendra la banque sera faible et meilleur sera le taux proposé.

L'Euribor comprenait initialement 13 durées de remboursement différentes ou 13 maturités (Euribor 1 mois, 2 mois, 3 mois ... jusqu'à 12 mois, ainsi que Euribor 1 semaine). A partir du 15 octobre 2001, deux nouvelles durées (Euribor 2 semaines et 3 semaines) sont venues s'ajouter.

Ce graphique présente l'évolution à court terme des principales maturités de l'Euribor en mois de mai, juin et juillet 2017.

Graphique de l'Euribor sur les principales maturités



2.5 Conclusion générale

L'étude présentée dans ce chapitre constitue une partie de travail mathématique élaboré pour Créditeo, nous nous restreignons dans cette présentation à ce niveau à cause de la conformité (de démarche de calcul) ente plusieurs formules. En revanche nous donnerons une idée sur le reste de travail fait pour Créditeo :

- **Le prêt amortissable avec différé total et partiel :**

  · Avec un différé de paiement partiel, l'emprunteur ne rembourse pas le capital pendant toute la durée du différé de paiement (appelé également « différé d'amortissement »). Seuls les intérêts, la prime d'assurance et d'éventuels frais seront prélevés.

  · Avec le différé total, l'emprunteur ne paiera plus de capital et intérêts durant la période de différé total. Les intérêts et les frais (assurance et garantie) restent néanmoins dus et s'ajouteront au capital emprunté.
- **Le prêt à taux zéro (ou zéro plus) :**

  Le prêt à taux zéro est un prêt destiné à financer un projet de premier résidence principal. Plus exactement, ce sont les personnes n'ayant pas été propriétaires de leur logement dans les deux ans précédent. Également avec des conditions demandés sur la somme des revenus fiscaux de référence des personnes destinées à occuper à titre principal le logement
- **le prêt progressif :**

Le prêt progressif, également appelé prêt à échéances progressives, désigne un crédit immobilier à taux fixe, dont les mensualités augmentent au fur et à mesure que le remboursement du prêt avance dans le temps.

– **le prêt à taux capé :**
Contrairement au taux révisable (variable) classique, le taux est dit capé ou sécurisé c'est-à-dire qu'il ne peut pas varier que dans les limites prévues contractuellement. Dans la majorité des cas, il est capé sur le taux à $+/-1\%$ ou 2%.

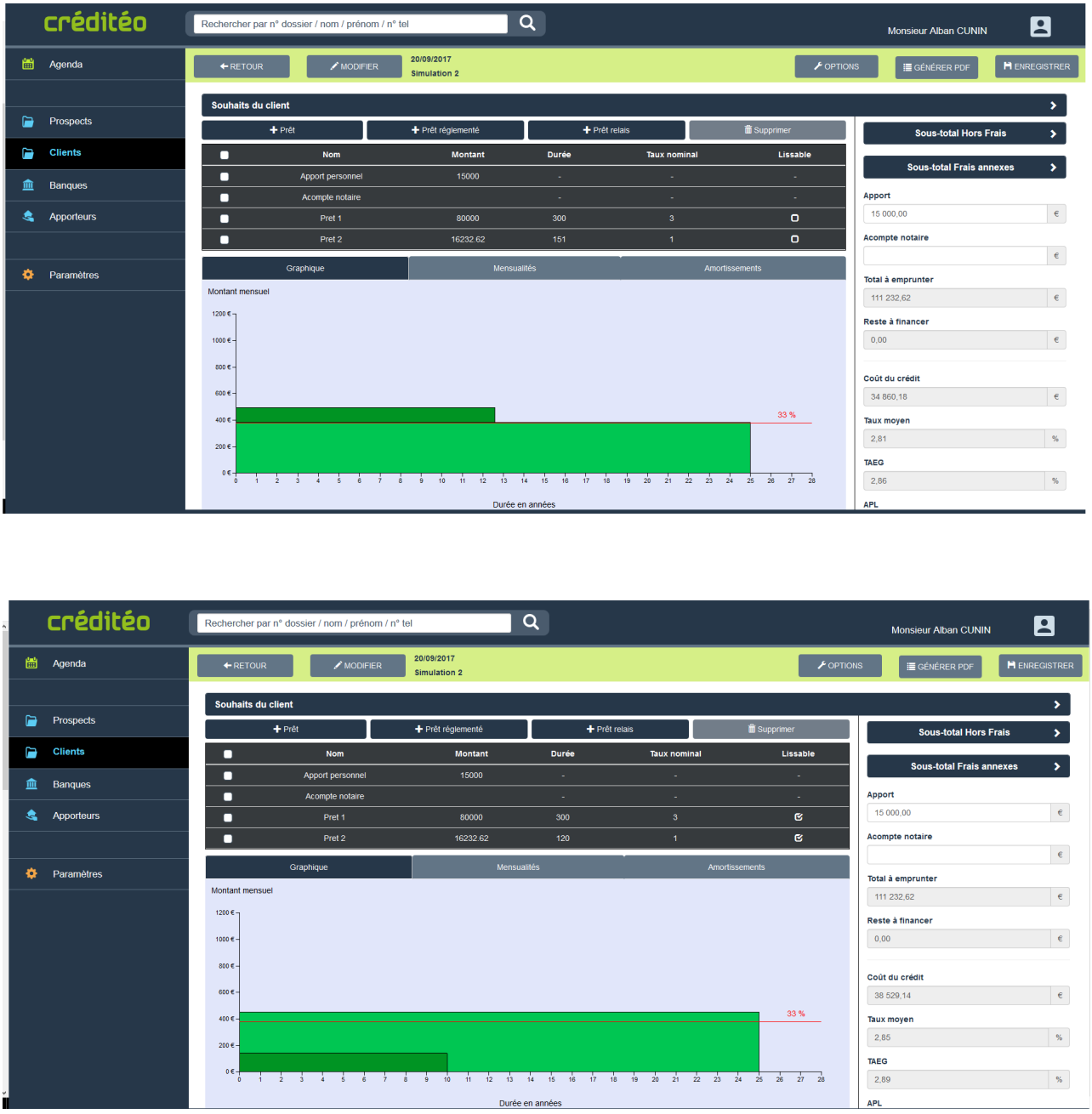


FIGURE 2.5 – Lissage des prêts avec créditéo

chapitre 3

Méthode de discrimination pour le crédit scoring

Introduction :

Pour résumé, Le crédit scoring est considéré comme une méthode d'évaluation du niveau du risque associé à un dossier de crédit. Cette méthode implique l'utilisation de différentes techniques statistiques pour aboutir à un modèle de scoring basé sur les caractéristiques du client. Il estime le risque de crédit en prévoyant la solvabilité du demandeur de crédit. Les institutions financières utilisent ce modèle pour estimer la probabilité de défaut qui va être utilisée pour affecter chaque client à la catégorie qui lui correspond le mieux : bon payeur ou mauvais payeur.

La construction d'un scoring fait appel à la modélisation prédictive, et l'on ne parle d'un scoring que lorsque la variable à prédire ne possède que deux modalités (cf.Saporta [14]). Dans ce chapitre, nous exposons une approche mathématique « classique » du crédit scoring par la modélisation du risque de crédit. Dans le cadre de notre étude, notre modèle sera construit à base d'un modèle paramétrique à savoir la régression logistique à cause de sa grande robustesse et sa facilité d'interprétation.

Pour nous, le but est d'identifier les signes du risque de crédit permettant de prévoir les défaillances, de construire pour cette méthodes un modèle en se basant sur une base de donnée qui contient tout un historique des caractéristiques des emprunteurs en crédit immobilier.

Dans ce qui suit, nous présentons la méthodologie de collecte des données et nous nous proposons simplement de produire via certains outils moins classiques mais efficaces et présents dans la plupart des logiciels statistiques comme le logiciel **R**. Cela nous permettra également d'illustrer les premières étapes exploratoires à réaliser sur notre jeux de données.

3.1 Méthodologie et collecte des données

La construction d'un échantillon pour notre étude statistique s'avère très couteuse en temps du point de vue de la collecte et du nettoyage de la base de données, en raisons des règles de confidentialité imposées par la nature des informations comptables et financières (personnelles).
La sélection s'est effectuée sur la base de données de M. CUNIN Alban (un courtier), qui date depuis 2014 (fichier électronique sous forme CSV), qui contient des informations complètes sur des dossiers de prêts immobiliers pour des emprunteurs.
Dans la sous-section suivante, on décrira notre base de données et on expliquera les contraintes survenus en travaillant avec, ainsi que la solution que j'ai proposé afin d'achever le travail.

3.1.1 Description des variables d'analyse

Le choix des variables d'analyse se doit d'obéir à la seule logique de couverture maximale, autant que faire se peut, de l'information susceptible d'aider à distinguer les bons dossiers de crédit des mauvais dossiers, ou les mauvais clients des bons clients. Les variables à retenir doivent donc contenir l'essentiel de l'information sur le client. La base de critères économiques et financiers comporte 72 observations (dont 6 données manquantes), 22 variables explicatives et une variable qualitative Y à expliquer dont les sélections sont faites selon les thèmes décrits dans le tableau 3.1.

Remarque 3.1. *Dans le tableau 3.1, les variables* ***CoutProjet***, ***TotalEmprunt***, ***Montant-Pret***, ***THA***, ***TH***, ***D***, ***Mensualité***, ***TotalCharges*** *et* ***Endettement*** *sont qualitatives et les* 12 *autres sont quantitatives. Il est à noter que la variable* Y *=statut d'un client est la variable qualitative binaire à prédire dont les modalités sont :*
0= *« bon client ou client non risqué »* et 1=*« mauvais client ou client risqué ».*

Nom	Variable	Codage
Co.emp	Présence d'un Co-emprunteur	1=Oui ; 0=Non
Age	âge du client	en années
SituationMatrimonial	Situation familiale	1=marié(e) ; 2=célibataire 3=union libre ; 4=Pacsé ; 5=Divorcé(e)
Rmensuels	Revenu mensuel	en euro
RevenusCoemp	Revenu du Co-emprunteur	en euro

CSP	Catégories socioprofessionnelles	2=artisans, commerçants et chefs d'entreprise ; 3=cadres et professions intellectuelles supérieures ; 4=professions intermédiaires ; 5=employés ; 6=ouvriers ; 7=retraités
SituationImmo	Situation immobilière	1=locataire ; 2=propriétaire ; 3 hébergé(e) à titre gratuit ; 4=Autre
NatureProjet	Nature du projet	1=rachat de prêt ; 2=acquisition ; 3=Acquisition+travaux ; 4=acquisition en VEEA ; 5=crédit professionnel
UsageProjet	Usage de projet	1=Résidence principale ; 2=Locative ; 3=Autres(consommation, résidence secondaire)
Etat	État de Projet	0=ancien ; 1=neuf
Zone	zone de projet	A ; B ; B2 ; C
CoutProjet	cout du projet	en euro
TotalEmprunt	Total emprunté	en euro
MontantPret	Montant du prêt	en euro
THA	Taux d'intérêt hors frais d'assurance	en pourcentage
TH	Taux d'assurance	en pourcentage
D	Durée du prêt	en mois
Mensualité	Mensualité	en euro
TotalCharges	Total des charges	en euro
Endettement	Taux d'endettement	en pourcentage
Nombre.d enfants	Nombre d'enfants	entier
Y	Statut d'un client	0=mauvais payeur ; 1=bon payeur

TABLE 3.1 – Codage des variables

Contrainte de l'étude :

La notation statistique des emprunteurs ou « scoring » fait recours à la modélisation prédictive qui se fond sur l'observation du passé : on connait pour un certains nombres de prêts attribués la qualité du payeur ainsi que les données recueillies lors du dépôt du dossier de prêt.

La plupart des organismes bancaires rencontrent deux problèmes majeurs dans l'étude de scoring, à savoir :

- Pas de stockage informatique des dossiers de prêts (il fallait alors retrouver les dossiers papiers).
- Une base de données de taille réduite (les observations), alors que le nombre de variables est importants.

Notre problème dans cette étude se résume par le deuxième point, en effet, à l'étape de la modélisation (la régression logistique binaire a été choisie à cause de sa robustesse et sa facilité d'interprétation, qu'on la détaillera dans la section 3.2), le logiciel **R** nous signale un message d'avertissement que l'algorithme d'optimisation pour l'estimation des paramètres de la régression logistique (par maximum de vraisemblance, voir la section 3.2), ne converge pas. Par conséquent, en utilisant cette base de données, notre travail statistique a été interrompu.

Solution proposée :

On remédie à ce problème, en utilisant une autre base de données sur le crédit allemand, qui contient des renseignements concernant 1000 clients ayant contracté un prêt à une banque. 700 des ces clients ont remboursé leur prêts sans difficulté, tandis que 300 ont eu des difficultés à rembourser leur prêts.

Source :

Le site UCI Machine Learning Repository est géré par l'université Irvine de l'état de Californie. Ce site propose des jeux de données, nous utiliserons le jeux de données appelé « German Crédit Data ». Il a été proposé par le professeur Hans Hofmann de l'université de Hambourg.

Les données :

Le jeux de données contient 1000 observations et 31 variables (qualitatives et quantitatives). La variable à expliquer est la variable **"RESPONSE"** associé au label "la notation du crédit est bonne", avec deux modalités : « 1= Oui, 0= Non ». Les variables explicatives sont toutes les autres exceptées le numéro de l'observation, elles sont décrites dans le catalogue si dessous (cf. tableau 3.2).

OBS	Nom du variable	Label	Modalité
1	CHK-ACCT	l'état du compte	0 :< 0 DM ; 1 : 0 < ... < 200 DM ; 2 :=> 200 DM ; 3 : aucun compte courant.
2	DURATION	durée de crédit	en mois.
3	HISTORY	l'historique des crédit	0 : pas de crédit attribué ; 1 tout les prêts accordés par cette banque ont été payés ; 2 : crédit existant remboursé en date voulu ; 3 retard dans le paiement ; 4 : compte critique.
4	NEW-CAR	un crédit voiture (nouveau)	0 : Non ; 1 : Oui.
5	USED-CAR	un crédit voiture (utilisé)	0 : Non ; 1 : Oui.
6	Fourniture	un crédit mobilier ou équipements	0 : Non ; 1 : Oui.
7	RADIO-TV	un crédit radio ou télévision	0 : Non ; 1 : Oui.
8	EDUCATION	un crédit éducation	0 : Non ; 1 : Oui.
9	RETRAINING	un crédit de recyclage	0 : Non ; 1 : Oui.
10	AMOUNT	montant du crédit	0 : Non ; 1 : Oui.
11	SAV-ACCT	solde moyen dans le compte d'épargne	0 :< 100 DM ; 1 : 100 <= ... < 500 DM ; 2 : 500 <= ... < 1000DM ; 3 :=> 1000 DM ; 4 : inconnu ou pas de compte d'épargne.
12	EMPLOYEMENT	emploi actuel depuis	0 : sans emploi ; 1 :< 1 ans ; 2 : 1 <= ... < 4 ans ; 3 : 4 <= ... < 7 ans ; 4 :>= 7 ans.
13	INSTALL-RATE	taux d'acompte en du revenu disponible	en pourcentage.
14	MALE-DIV	l'emprunteur est masculin et divorcé	0 : Non ; 1 : Oui.
15	MALE-SINGLE	l'emprunteur est masculin et célibataire	0 : Non ; 1 : Oui.
16	MALE-MAR-WID	l'emprunteur est masculin et marié ou veuf	0 : Non ; 1 : Oui.

17	CO-APPLICANT	l'emprunteur a un co-emprunteur	0 : Non ; 1 : Oui.
18	GUARANTOR	l'emprunteur a un garant	0 : Non ; 1 : Oui.
19	PRESENT-RESIDENT	la résidence présente est de puis	1 : <=ans ; 2 :1 < ... <= 2 ans ; 3 2 < ... <= 3 ans ; 4 : > 4 ans.
20	REAL-ESTATE	l'emprunteur a l'immobilier	0 : Non ; 1 : Oui.
21	PROP-UNKN-NONE	l'emprunteur ne possède aucun bien (ou inconnu)	0 : Non ; 1 : Oui.
22	AGE	l'age de l'emprunteur	en année.
23	OTHER-INSTALL		
24	RENT	possession des loyers	0 : Non ; 1 : Oui.
25	OWN-RES	l'emprunteur possède sa résidence	0 : Non ; 1 : Oui.
26	JOB	nature de l'emploi	0 : sans emploi/non qualifié/non résident ; 1 :résident non qualifié ; 2 : employé qualifié ; 3 :Gestion/travailleur indépendant/employée/officier bien qualifié.
27	NUM-DEPENDENT	nombre de personne en charge	numérique.
28	TELEPHONE	l'emprunteur à un numéro de téléphone en son nom	0 : Non ; 1 : Oui.
29	FOREIGN	travailleur étranger	0 : Non ; 1 : Oui.
30	RESPONSE	la note du crédit est bonne	0 : Non ; 1 : Oui.

TABLE 3.2 – Codage des variables d'étude

Dans la section suivante, nous décrivons la modélisation de notre variable aléatoire qualitative Y (cf. tableau 3.2).

3.2 Le modèle probabiliste de prédiction

Nous sommes à la présence d'un vecteur de variables aléatoires (X, Y).
Y est qualitative à deux modalités $\{0, 1\}$, « 1= la note du crédit est bonne, 0= la note du crédit est mauvaise » et $X = \{1, X_1, ..., X_p\}$ est qualitative ou quantitative, avec n observa-

tions $\{x_{i1}, ..., x_{ip}, y_i\}_{i=1:n}$.

Objectif : nous souhaitons expliquer une variable qualitative Y à partir de p variable quantitatives et qualitatives

Une nouvelle observation $x = (x_1, ..., x_p)$ arrive, nous mesurons les variables explicatives (cf. tableau 3.1), nous souhaitons prédire sa classe $Y = y_0 \in \{0, 1\}$ à partir de l'observation x, nous cherchons alors, $\mathbb{P}(Y = j|X = x)$. (Cf.Rouvière [11])

Définition 3.1. ***(régression logistique binaire)*** *Soit Y une variable à valeurs dans $\{0, 1\}$ à expliquer par p variables quantitatives $X = \{1, X_1, ..., X_p\}$. Le modèle logistique propose une modélisation de la loi $Y|X = x$ par une loi de Bernoulli de paramètre $\mathbb{P}(Y = 1|X = x)$ telle que :*

$$\boxed{\log \frac{\mathbb{P}(Y = 1|X = x)}{1 - \mathbb{P}(Y = 1|X = x)} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p} \tag{3.1}$$

ou encore

$$logit(\mathbb{P}(Y = 1|X = x)) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p. \tag{3.2}$$

Avec $\beta = (\beta_0, \beta_1, ..., \beta_p) \in \mathbb{R}^{p+1}$.
logit *désignant la fonction bijective et dérivable de $]0, 1[$ dans $\mathbb{R}$: $p \longmapsto \log(\frac{p}{1-p})$.*
L'égalité (3.1) peut également s'écrire,

$$\mathbb{P}(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}, \tag{3.3}$$

avec, $x = (x_1, ..., x_p)$.

Cas des variables qualitatives :(Mise sous forme disjonctive)

Pour un individu, une variable qualitative X_i à m catégories ou modalités, on définit les m variables indicatrices de ses modalités $\mathbb{1}^i = (\mathbb{1}_1, \mathbb{1}_2, ..., \mathbb{1}_m)$, telle que $\mathbb{1}_j$, $j \in \{1, .., m\}$, vaut 1 si on appartient à la modalité j, 0 sinon. Seule une des indicatrices vaut 1, celle qui correspond à la modalité prise. Les p indicatrices sont donc équivalentes à la variable qualitative.(cf.Saporta [15], Blayac [3])
Pour n individus, une variable qualitative X_i peut être représentée par le tableau de données binaires $\mathbb{X}$ suivant :

$$\mathbb{X} = \begin{pmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ . & & & & . \\ . & & & & . \\ 0 & 0 & 1 & ... & 0 \end{pmatrix}$$

On dit que X a été mise sous forme disjonctive.
Les variables qualitatives (explicatives) qui interviennent dans le modèle logistique (3.1) sont donc les indicatrices de toutes les variables.

Remarque 3.2. *Dans un modèle logistique, nous effectuons deux choix pour définir le modèle :*

1 Le choix d'une loi pour $Y|X = x$, ici la loi de Bernoulli ;
2 le choix de la modélisation de $\mathbb{P}(Y = 1|X = x)$ par : $logit(\mathbb{P}(Y = 1|X = x)) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$

La fonction logit est bijective et dérivable. Elle est appelée fonction de lien.
Remarquons également que

$$\mathbb{E}(Y|X = x_0) = \mathbb{P}(Y = 1|X = x_0).$$

3.3 Estimation des paramètres

3.3.1 L'estimateur du maximum de vraisemblance

L'estimation des paramètres se fait ici par maximum de vraisemblance. Nous sommes en présence de n observations des variables notées $\{1, X_1, ..., X_p, Y_i\}_{i=1:p}$, dont la i^e est notée (x_i, y_i), $y_i \in \{0, 1\}$. La vraisemblance conditionnelle de $Y|X = x_i$ associée à l'observation i s'écrit :

$$\nu(y_i, \beta) = \mathbb{P}(Y = 1|X = x_i)^{y_i}\mathbb{P}(Y = 0|X = x_i)^{1-y_i},\ \beta = (\beta_1, ..., \beta_p).$$

Et donc la vraisemblance conditionnelle pour l'échantillon $y = \{y_1, ..., y_n\}$ s'écrit :

$$\nu(y, \beta) = \prod_{i=0}^{n} \mathbb{P}(Y = 1|X = x_i)^{y_i}\mathbb{P}(Y = 0|X = x_i)^{1-y_i}. \tag{3.4}$$

Il est plus aisé de se servir de la log-vraisemblance notée $L(y, \beta)$. En passant au log dans (3.4), on aura :

$$L_n(y, \beta) = \sum_{i=1}^{n} \{y_i \ln(\frac{\mathbb{P}(Y = 1|X = x_i)}{\mathbb{P}(Y = 0|X = x_i)}) + \ln \mathbb{P}(Y = 0|X = x_i)\}. \tag{3.5}$$

En utilisant la définition du modèle logistique(3.1), on aura :

$$L_n(y, \beta) = \sum_{i=1}^{n} \{y_i({x_i}^T\beta) - \ln(1 + \exp({x_i}^T\beta))\}. \tag{3.6}$$

Définition 3.2. *On appelle estimation du maximum de vraisemblance une valeur $\hat{\beta}$, s'il existe une, telle que :*

$$L(y, \hat{\beta}) = \sup_{\beta} L_n(y, \beta). \tag{3.7}$$

Une telle solution dépend de la valeur de $x_1, ..., x_p$, ($\hat{\beta} = f(x_1, ..., x_p)$).

Théorème 3.1. *(Rappel) (cf.Rouvière [11])*
soit $\hat{\beta}$ l'estimateur du maximum de de vraisemblance défini ci dessus. sous certaines conditions de régularité, on a :
– $\hat{\beta}$ converge presque sûrement vers β.
– $\hat{\beta}$ est asymptotiquement normal :

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, [\mathcal{I}(\beta)]^{-1}) \text{ et } (\hat{\beta} - \beta)' n\mathcal{I}(\beta)(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{X}^2_{p+1}$$

.

Avec $\mathcal{I}[\beta]$ es la matrice d'information de Fisher (de dimension $(p+1) \times (p+1)$) au point β :

$$\mathcal{I}(\beta)_{kl} = -\mathbb{E}\left[\frac{\partial^2}{\partial\beta_k \partial\beta_l} L_1(y, \beta)\right], \ 0 \leq k, l \leq p. \tag{3.8}$$

Démonstration Pour une preuve détaillée, on pourra regarder Fahrmier Kaufmann [10], Cadre B [7]. □

Remarque 3.3. *La détermination d'un estimateur noté $\hat{\beta}$ de β, revient à maximiser la log-vraisemblance, en annulant le gradient de $L(y, \beta)$.*
Le vecteur gradient au point β défini par $\nabla L(\beta) = [\frac{\partial L}{\partial \beta_0}, ..., \frac{\partial L}{\partial \beta_p}]^T$ s'obtient par dérivation

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i x_{ij} - \frac{x_{ij} \exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \right].$$

*La valeur de β qui maximise la log-vraisemblance doit satisfaire l'équation (appelée équation du score) $\frac{\partial L(y, \hat{\beta})}{\partial \hat{\beta}} = 0$. Cette solution n'a pas de solution explicite. De point de vue pratique, on a recours à des méthodes numériques d'optimisation pour obtenir la valeur estimé de β (**l'algorithme de Newton-Raphson**), l'algorithme démarre avec une initialisation quelconque β^0, pour passer de l'étape (i) à l'étape $(i+1)$, il se rapproche de la solution finale $\hat{\beta}$ en utilisant la formule suivante :*

$$\boxed{\beta^{i+1} = \beta^i - \boldsymbol{H}^{-1}(\beta^i) h(\beta^i)}. \tag{3.9}$$

Ou

• *$h(\beta)$ est le vecteur des dérivées partielles de $L(y,\beta)$ par rapport à β, et qui est égale à :*

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n}\left[y_i x_{ij} - \frac{x_{ij}\exp\left(x_i^T\beta\right)}{1+\exp\left(x_i^T\beta\right)}\right]. \tag{3.10}$$

$$= \boldsymbol{X}(\mathbb{Y} - P_\beta). \tag{3.11}$$

Où, $\mathbb{Y} = (y_1, ..., y_n)'$, $P_\beta = (p_\beta(x_1), ..., p_\beta(x_n))$, avec $p_\beta(x_i) = \mathbb{P}((Y = 1|X = x_i)$ et $\boldsymbol{X}$: la matrice des observations,

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & .... & x_{1p} \\ 1 & . & .... & . \\ 1 & . & .... & . \\ 1 & . & .... & . \\ 1 & x_{n1} & .... & x_{np} \end{pmatrix}$$

• *H est la matrice des dérivées seconde $\boldsymbol{H}(\beta) = \frac{\partial^2}{\partial\beta_k\partial_l}L(y,\beta)$. Elle est égale à*

$$\boldsymbol{H}(\beta) = -\boldsymbol{X'VX}. \tag{3.12}$$

Où $\boldsymbol{V}$ la matrice diagonale dont les éléments sont $(p_\beta(x_i)(1-p_\beta(x_i)))$, $i = 1, ..., n$.
En effet

$$\begin{aligned}\frac{\partial^2}{\partial\beta_k\partial\beta_l}L(y,\beta) &= -\sum_{i=1}^{n} x_{ik}x_{il}\frac{\exp(x_i'\beta)}{(1+\exp(x_i'\beta))^2} \\ &= -\sum_{i=1}^{n} x_{ik}x_{il}p_\beta(x_i)(1-p_\beta(x_i)). \\ &= -\boldsymbol{X'VX}.\end{aligned}$$

Par suite, on peut exprimer β_{i+1} en fonction de β_i :

$$\boxed{\beta^{i+1} = \beta^i + (\boldsymbol{X'VX})\boldsymbol{X}(\mathbb{Y} - P_\beta)}. \tag{3.13}$$

• Plusieurs règles d'arrêt de l'algorithme sont possibles pour stopper le processus de recherche :

– On fixe à l'avance le nombre maximum d'itérations pour limiter le temps de calcul. Souvent bien utile pour éviter les boucles infinies faute de convergence.
– On stoppe les itérations lorsque l'écart entre les solutions $\hat{\beta}$ est faible d'une étape à l'autre.

Remarque 3.4. ***(Loi asymptotique)***

$$(\hat{\beta} - \beta)'(\boldsymbol{X'VX})(\hat{\beta} - \beta)\xrightarrow{\mathcal{L}}\mathcal{X}^2_{p+1}. \tag{3.14}$$

voir (Rouvière [11], page 24).

3.4 Choix de seuil pour la modélisation

Comme nous l'avons expliqué, le choix de la méthode de modélisations pour la prévision du notation d'emprunteur a été sur la régression logistique.
Pour choisir la valeur seuil notée s dans toute la suite, au-delà de la quelle un emprunteur va être considéré comme mauvais payeur, on fait appel à la règle « naïve » de Bayes.

Nous rappelons que nous sommes en présence de n observations d'un couple (X, Y), telle que $X = (X_1, ..., X_p)$ et Y prend deux modalités 0 ou 1.
Pour une nouvelle observation $x \in \mathbb{R}^p$, nous souhaitons prédire son groupe $Y = y \in \{0, 1\}$ à partir de l'observation de X avec une probabilité de se tromper dans cette prédiction aussi faible que possible (cf. cours [1]). Ceci revient à apprendre une fonction.

$$h : \mathbb{R}^p \longrightarrow \{0, 1\}.$$

Telle que l'erreur de prédiction qu'on le note, $\varepsilon(h) = \mathbb{P}(h(X) \neq Y)$ soit assez petite.
Donc, il faudra chercher une fonction $h^* : \mathbb{R}^p \longrightarrow \{0, 1\}$, vérifiant $\varepsilon(h^*) = \min_h \varepsilon(h)$.
Si une telle fonction h^* existe, le prédicateur $h^*(x)$ sera le meilleur.

3.4.1 Prédicateur de Bayes-Erreur de Bayes :

$h : \mathbb{R}^p \longrightarrow \{0, 1\}$, on :
$\varepsilon(h) = \mathbb{P}(h(X) \neq Y) = \mathbb{P}(h(X) = 0, Y = 1) + \mathbb{P}(h(X) = 1, Y = 0) = \mathbb{E}(\mathbb{1}_{h(X)\neq Y})$

On pose alors, $\varepsilon(h|X = x) = \mathbb{P}(h(X) \neq Y|X = x))$, c'est la probabilité de se tromper dans la prédiction de la valeur de Y, connaissant les valeurs prisent par X, pour un individu.

Définition 3.3. *Un prédicateur $h^*(X)$ de Y qui vérifie $\varepsilon(h^*|X = x) = \min \varepsilon(h); \forall h : \mathbb{R}^p \longrightarrow \{0, 1\}$ est appelé prédicateur de Bayes de $Y|X = x$.*
$\varepsilon(h^*)$ est appelé erreur de Bayes.

Choix de seuil

La théorème d Bayes nous permet d'écrire :

$$\mathbb{P}(Y = j|X = x) = \frac{\mathbb{P}(Y = j)\mathbb{P}(X = x|Y = j)}{\mathbb{P}(Y = 1)\mathbb{P}(X = x|Y = 1) + \mathbb{P}(Y = 0)\mathbb{P}(X = x|Y = 0)}, \forall j \in \{0, 1\}. \tag{3.15}$$

Le prédicateur de Bayes $h^*(x)$ est défini par

$$h^*(x) = \underset{k=0,1}{\operatorname{argmax}}\ \mathbb{P}(Y = k)\mathbb{P}(X = x|Y = k). \tag{3.16}$$

On aura alors la règle de décision finale, elle s'écrit comme suit :

1. si $\mathbb{P}(Y=1|X=x) < s = \frac{1}{2}$, alors $h^*(x) = 0$.
2. si $\mathbb{P}(Y=1|X=x) > s = \frac{1}{2}$, alors $h^*(x) = 1$.
3. si $\mathbb{P}(Y=1|X=x) = s = \frac{1}{2}$, alors $h^*(x) = 1$ ou $h^*(x) = 0$.

3.5 Première évaluation de la régression :

3.5.1 Tests d'hypothèses sur les paramètres du modèle :

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

$$H_0 : \beta_j = 0,\ j \in \{1, ..., p\}$$

– Évaluer la contribution d'un bloc de q variables, nous écrivons H_0 de la manière suivante :

$$H_0 : \beta_1 = ... = \beta_q = 0$$

– Évaluer l'apport de l'ensemble des variables explicatives (les p variables). Il s'agit d'une évaluation globale de la régression.

$$H_0 : \beta_1 = ... = \beta_p = 0.$$

Dans tous les cas, l'hypothèse alternative correspond à : $H_1 : \exists j \in \{1, ..., p\},\ \beta_j \neq 0$.
En utilisant une des hypothèses nulles présentées, on dispose deux tests :

• **Test de Wald : (évaluation individuelle des coefficients)**
Dans cette cas, nous réalisons le test suivant :

$$\begin{cases} H_0 : \beta_j = 0 \ \text{ (la variable n'est pas significative dans le modèle).} \\ H_1 : \beta_j \neq 0 \ \text{(la variable est significative dans le modèle).} \end{cases}$$

Le test de Wald est basé sur (3.14). On note $\beta_{0,...,q-1}$ le vecteur composé des q premières composantes de β et $(\mathbf{X'VX})_{0,...,q}$ la matrice composée des q premières lignes et colonnes de $\mathbf{(X'\ V\ X)}$. Donc sous H_0, on aura :

$$(\hat{\beta} - \beta)'_{0,...,q}(\mathbf{X'VX})_{0,...,q}(\hat{\beta} - \beta)_{0,...,q} \xrightarrow{\mathcal{L}} \mathcal{X}^2_{p+1}.$$

• **Test de rapport de vraisemblance ou de la déviance :**

La statistique de test est basé sur la différence des rapports de vraisemblance entre le modèle complet (avec tout le variables) et le modèle sous H_0.

On note $\hat{\beta}_{H_0}$ l'estimateur du maximum de vraisemblance contraint par H_0 (il s'obtient en supprimant les q premières variables du modèle). On aura sous H_0,

$$2(L_n(\hat{\beta}) - L_n(\hat{\beta}_{H_0})) \xrightarrow{\mathcal{L}} \mathcal{X}^2_q. \tag{3.17}$$

Pour ces deux tests, on rejette l'hypothèse nulle si la valeur observée de la statistique de test dépasse le quantile d'ordre $(1-\alpha)$ de la loi de $\mathcal{X}^2_q$, tel que α est le risque de première espèce (le risque de rejeter l'hypothèse nulle alors qu'elle est vraie).

3.6 Interprétation des paramètres et Odds ratio

L'évaluation de l'ajustement du modèle doit précéder toute tentative d'interprétation.

On suppose dans cette partie que le modèle de régression logistique a été trouvé, que les variables sont significatives au sens statistique et que le modèle est en adéquation avec les données. Les méthodes d'évaluation de l'adéquation sont assez techniques et sont ainsi exposées dans la section 3.7.

La régression logistique propose des outils qui permettent d'interpréter les résultats sous forme de risques, de chances, de rapports de chances. Une large documentation est dédiée à l'interprétation des sorties de la régression logistique dans les ouvrages qui font référence (Hastie [9], chapitre 3 ; S. Menard [13], chapitre 3).

On reprend l'équation de modèle logistique, défini par l'équation (3.3).

$$\mathbb{P}(Y=1|X=x) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)},$$

avec, $x = (x_1, ..., x_p)$.

On se rappelle que dans le modèle logistique binaire, la fonction de lien est la transformation logit :

$$f(x) = logit(\mathbb{P}(Y=1|X=x)) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p.$$

Avec $\beta = (\beta_0, \beta_1, ..., \beta_p) \in \mathbb{R}^{p+1}$.

3.6.1 Cas d'une variable explicative dichotomique

Pour une variable explicative binaire,

$$f(x) = logit(\mathbb{P}(Y = 1|X = x)) = \beta_0 + \beta_1 \mathbb{1}_{\{X=1\}} + \beta_2 \mathbb{1}_{\{X=0\}}.$$

la différence dans le logit pour une observation entre $x = 1$ et $x = 0$ est :

$$f(1) - f(0) = \beta_0 - \beta_1 - \beta_0 = \beta_1.$$

Afin d'interpréter ce résultat, nous avons besoin d'introduire une mesure d'association appelée odd ratio.
Les valeurs des probabilités quand la variable est dichotomique sont résumées dans le tableau suivant, on note $\pi_0 = \mathbb{P}(Y = 1|X = 0)$ et $\pi_1 = \mathbb{P}(Y = 1|X = 1)$.

	Présence du facteur ($X = 1$)	absence du facteur ($X = 0$)
Bon payeur ($Y = 1$)	$\pi_1 = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$	$\pi_0 \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)}$
Mauvais payeur ($Y = 0$)	$1 - \pi_1 = \frac{1}{1+\exp(\beta_0+\beta_1)}$	$1 - \pi_0 = \frac{1}{1+\exp(\beta_0+\beta_2)}$

• Le risque (ou chance ou odd) d'avoir $Y = 1$ quand le facteur est absent ($X = 0$) correspond au rapport $\frac{\pi_0}{1-\pi_0}$.
Le risque d'avoir $Y = 1$ lorsque $X = 1$ correspond au rapport $\frac{\pi_1}{1-\pi_1}$.
• L'odd ratio(ou rapport des chances ou rapport de cotes) noté OR, est défini par le rapport de odds pour $X = 1$ et $X = 0$:

$$\boxed{OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}} \tag{3.18}$$

$$\ln(\frac{\pi_1}{1-\pi_1}) = \ln(\frac{\pi_0}{1-\pi_0}) - \beta_2 + \beta_1 \;\Leftrightarrow\; \boxed{OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}} = \exp(\beta_2 - \beta_1)} \tag{3.19}$$

L'interprétation de l'odd ratio : L'odd ratio est facilement interprétable. Il s'agit d'un rapport de chances : la chance relative des observation présentant le facteur X d'avoir une défaillance (pour notre exemple) est OR (odd ratio) fois celle des observations ne la présentent pas. Si l'odd ratio est positif, les observations présentent le facteur X ont alors plus de chance d'avoir une défaillance que celles ne présentent pas le facteur.

3.6.2 Cas d'un modèle à une variable quantitative

Pour un modèle de régression logistique contient une variable explicative quantitative, l'interprétation du coefficient dépend de l'unité de la variable.

$$f(x) = logit(\mathbb{P}(Y = 1|X = x)) = \beta_0 + \beta_1 x$$

.

La valeur du coefficient β correspond au changement du logit pour une augmentation d'une unité x.

Si on s'intéresse au changement de $k \geq 1$ unités pour X, l'odd ratio vaut :

$$OR = \exp(k\beta_1). \tag{3.20}$$

3.6.3 Cas d'un modèle avec une variable dichotomiques et une quantitative

On considère un modèle de régression logistique contenant une variable X_1 dichotomique et une variable X_2 continue ou quantitative.

$$f(x_1, x_2) = logit\,[Y = 1|X = x] = \beta_0 + \beta_1 \mathbb{1}_{\{X_1=1\}} + \beta_2 \mathbb{1}_{\{X_1=0\}} + \beta_3 x_2.$$

avec $X = (X_1, X_2)$ et $x = (x_1, x_2)$.

L'odd ratio associé au passage de la modalité 1 à 0 pour la variable X_1 et à l'augmentation de k unités de X_2 vaut :

$$OR = \exp\left[\frac{logit(P(Y = 1|X_1 = 1, X_2 = x_2 + k))}{logit(P(Y = 1|X_1 = 0, X_2 = x_2))}\right] = \frac{\exp(\beta_0 + \beta_1 + \beta_2(x_2 + k))}{\exp(\beta_0 + \beta_2 x_2)} = \exp(\beta_1 + \beta_2 k). \tag{3.21}$$

3.6.4 Odd ratio et risque relatif

De façon analogue à la définition de l'odd ration défini par l'équation (3.18), le risque relatif (RR) pour est définie par :

$$RR = \frac{\pi_1}{\pi_0}. \tag{3.22}$$

On en déduit facilement la relation entre le risque relatif et l'odd ratio,

$$OR = \frac{\pi_1}{\pi_0} \times \frac{1 - \pi_0}{1 - \pi_1} = RR \times \frac{1 - \pi_0}{1\pi_1}.$$

Ainsi, lorsque les probabilités π_1 et π_0 sont petites, l'odd ratio est une approximation du risque relatif.

3.7 Sélection et validation de modèles

Cette partie se divise en deux parties :

1. **Sélection** : Étant donnée M modèles $\mathcal{M}_1, ..., \mathcal{M}_M$, comment sélectionner le "meilleur" à partir de l'échantillon dont on dispose.
2. **Validation** est ce que le modèle choisi est "bon" ?

Dans cette partie nous allons traiter ces questions à travers le modèle logistique.

3.7.1 Sélection ou choix de modèle

Pour le modèle logistique, sélection un modèle revient à choisir les variables qui vont constituer le modèle.

Outil spécifique : la déviance :

Pour la régression logistique, un outils spécifique est introduit : la déviance.

La déviance permet de comparer la vraisemblance du modèle étudié à celle d'un modèle parfait (complet ou saturé par définition) en terme d'adéquation aux données, il reconstitue parfaitement les valeurs de la variable dépendante (ou à expliquer), dont lequel on a tous les paramètres.

Modèle saturé en présence des données individuelles :

On note $\{(x_1, y_1), ..., (x_n, y_n)\}$ l'échantillon et $x = (x_1, ..., x_n)$, on dit que les observations $x_i,\ i \in \{1, ..., n\}$, sont individuelles lorsque tous les x_i sont différents.

Dans le modèle saturé, la prévision est parfaite et dans le cas des données individuelles, la probabilité estimée par le modèle au point $X = x_i$ est donc 1 pour le groupe observé et 0 sinon, c'est à dire $\mathbb{P}(Y = y_i|X = x_i) = y_i,\ y_i \in \{0, 1\}$

On note $\mathcal{L}_{sat}$ la log-vraisemblance du modèle saturé et la vraisemblance du modèle saturé est égale à :

$$\nu_{sat} = \prod_{i=1}^{n} y^{y_i}(1 - y_i)^{1-y_i} = 1.$$

Donc $\mathcal{L}_{sat} = 0$.

Modèle saturé en présence des données répétées :

En présence de données répétées $\{(x_1, n_1), ..., (x_T, n_T)\}$, tel que, $n_i\ i \in \{1, ..., T\}$ est le nombre de répétition de l'observation x_i, c'est le cas où plusieurs observations seraient disponibles au point $X = x_i$. Dans ce cas, le modèle saturé modélise $\mathbb{P}(Y = y_i|X = x_i) = \bar{y_i} = \frac{y_i}{n_i}$,

$i \in \{1, ..., T\}$. La vraisemblance du modèle saturé notée ν_{sat} est donné par (cf. Rouvière [11]) :

$$\nu_{sat} = \prod_{i=1}^{T} \binom{n_i}{y_i} \bar{y_i}^{y_i} (1 - \bar{y_i})^{n_i - y_i}.$$

On obtient la log-vraisemblance

$$\mathcal{L}_{sat} = \sum_{i=1}^{T} log \binom{n_i}{y_i} + \sum_{i=1}^{T} y_i log(\bar{y_i}) + (n_i - y_i) log(1 - \bar{y_i}). \tag{3.23}$$

Définition 3.4. ***(La déviance)*** *On note $\mathcal{L}_{sat}$ la log-vraisemblance du modèle saturé et $\mathcal{L}_n$ celle du modèle étudié. La déviance d'un modèle est défini par :*

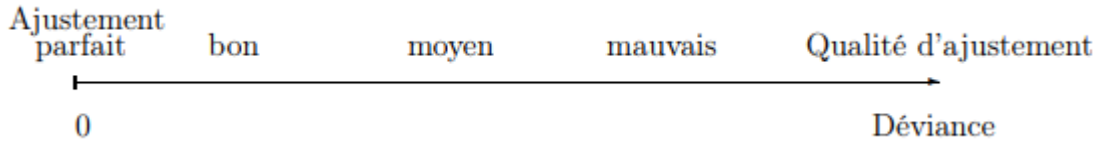
$$\boxed{D = 2\left[\mathcal{L}_{sat} - \mathcal{L}_n(y, \beta)\right]} \tag{3.24}$$

En présence de données individuelles on a, $D = -2\mathcal{L}_n(y, \beta)$ avec $\mathcal{L}_n(y, \beta)$ est donnée par l'équation (3.6) et en présence de données répétées, en utilisant les équations (3.6),(3.24) et (3.23), la déviance s'écrit

$$D = 2 \sum_{i=1}^{T} n_i \left[\bar{y_i} \log(\frac{\bar{y_i}}{p_\beta(x_i)}) + (1 - \bar{y_i}) \log(\frac{1 - \bar{y_i}}{1 - p_\beta(x_i)}) \right].$$

Où, $\bar{y_i} = \frac{y_i}{n_i}$.

La déviance constitue un écart en terme de log-vraisemblance entre le modèle saturé d'ajustement maximum et le modèle considéré



3.7.2 Sélection automatique

Une autre approche de la sélection de modèle consiste à chercher parmi les variables $X_1, ..., X_p$, celles qui "explique le mieux" Y.Par exemple, pour la régression logistique, nous pourrions nous poser le problème de chercher le meilleur sous ensemble des p variables explicatives pour un critère donné **(AIC, BIC...)**. nous sélectionnerions le modèle qui optimiserait le critère. Nous définissons ces deux critères :

Définition 3.5. ***(AIC (Akaike Informative Criterion))***
Pour un modèle à p variables, l'AIC est défini par :

$$AIC = -2\mathcal{L}_n + 2p.$$

Avec, $\mathcal{L}_n$ est la log-vraisemblance donnée par l'équation (3.6) du modèle logistique.

Définition 3.6. ***(BIC (Bayesian Informative Criterion))*** *Pour un modèle à p variables, le BIC est défini par :*

$$BIC = -2\mathcal{L}_n + p\log(n).$$

Avec, $\mathcal{L}_n$ est la log-vraisemblance donnée par l'équation (3.6) du modèle logistique.

On choisira le modèle qui possède le plus petit AIC ou BIC. L'utilisation de ces critères est simple. Pour chaque modèle concurrent le critère de choix de modèles est calculé et le modèle qui présente le plus faible est sélectionné.

Recherche pas à pas, méthode ascendante (forward selection)

A chaque pas, une variable est ajoutée au modèle de départ. Nous ajoutons la variable X_j dont l'ajout au modèle conduit à l'optimisation la plus grande du critère de choix (AIC ou BIC). Nous nous arrêtons lorsque toutes les variables sont intégrées ou lorsque qu'aucune variable ne permet pas l'optimisation du critère de choix.(voir aussi figure 3.1)

Recherche pas à pas, méthode descendante (backward selection)

A la première étape toutes les p variables sont intégrées au modèle.
Nous retirons la variable X_j dont le retrait du modèle conduit à l'augmentation la plus grande du critère de choix. Nous nous arrêtons lorsque toutes les variables sont retirées ou lorsque qu'aucune variable ne permet pas l'augmentation du critère de choix (AIC ou BIC).

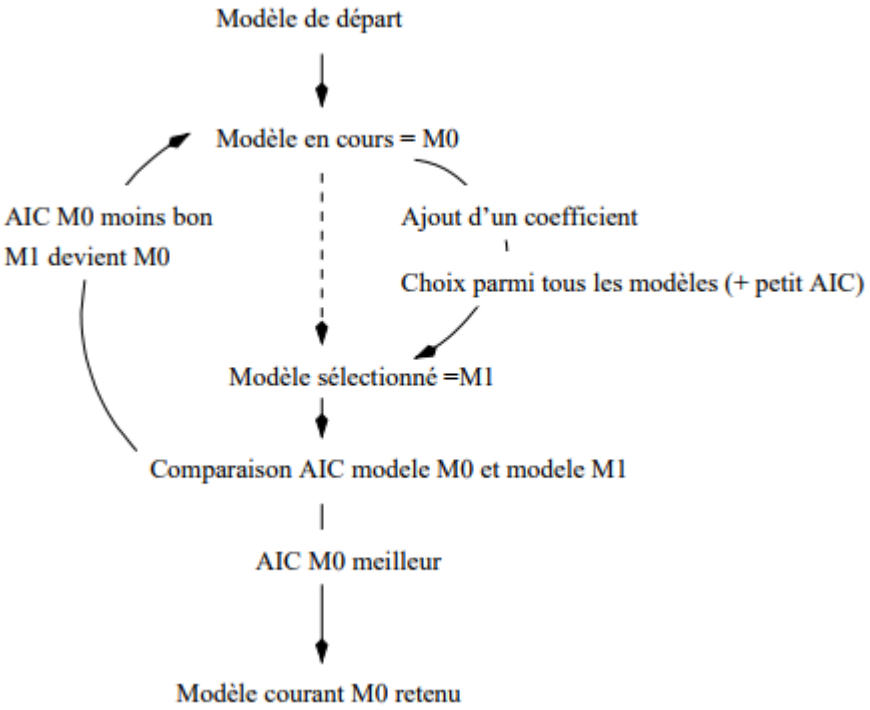


FIGURE 3.1 – Technique ascendante utilisant l'AIC.

3.8 Validation du modèle

Maintenant que nous avons construit un modèle de prédiction, il faut en évaluer l'efficacité. Nous pouvons le faire de différentes manières :
Confronter les valeurs observées de la variable dépendante Y avec les prédictions $\hat{Y}$. L'outil privilégié est la matrice de confusion

3.8.1 La matrice de confusion

Définition 3.7. *Une matrice de confusion ou tableau de contingence sert à évaluer la qualité d'un modèle de prédiction. Elle confronte toujours les valeurs observées de la variable dépendante Y avec celles qui sont prédite (voir tableau 4.1), puis comptabilise les bonnes et les mauvaises prédictions.*

$Y/\hat{Y}$	$\hat{1}$	$\hat{0}$	*Total*
1	a	b	$a+b$
0	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

TABLE 3.3 – Matrice de confusion - Forme générique

A partir de la forme générique de la matrice de confusion (Tableau 4.1), 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 observation qui ont été classées dans la classe* 1 *et qui le sont réellement.*
- ***d*** *sont les observation qui ont été classées dans la classe* 0 *et qui le sont réellement.*
- ***c*** *sont les observations classés* 1 *et qui sont en réalité dans la classe* 0*.*
- ***b*** *sont les observations classés* 0 *et qui sont en réalité dans la classe* 1*.*
- ***Le taux d'erreur*** *est égal au nombre de mauvais classement rapporté à l'effectif total c-à-d*

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

Il estime la probabilité de mauvais classement du modèle.

- ***Le taux de succès*** *correspond à la probabilité de bon classement du modèle*

$$\theta = \frac{a+d}{n} = 1 - \epsilon.$$

- ***La sensibilité*** *(ou le rappel, ou encore le taux de vrais classement à* 1 *[TVP]), indique la capacité du modèle à retrouver les classés à* 1*.*

$$S_c = \frac{a}{a+b}.$$

- ***La précision*** *indique la proportion de vrais classement à la classe* 1 *parmi les observation qui ont été classés* 1

$$précision = \frac{a}{a+c}.$$

Elle estime la probabilité d'un individu d'être réellement classé 1 *lorsque le modèle le classe comme tel.*

- ***La spécificité****, à l'inverse de la sensibilité, elle indique la proportion de* 0 *détectés.*

$$S_p = \frac{d}{c+d}.$$

Remarque 3.5. *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).*

3.8.2 Test de Hosmer-Lemeshow

Ce test permet de vérifier l'adéquation d'un modèle. Le test s'effectue de la manière suivante (voir Hosmer et Lemeshow [12], chapitre 5 pour plus de précisions).

1. Les probabilité $p_\beta(x_i) = \mathbb{P}(Y = 1|X = x_i)$ estimée par le modèle sont ordonnées par ordre croissant.
2. Ces probabilité ordonnées sont ensuite séparées en K groupes de taille égale (on prend souvent $K = 10$ si n est suffisamment grand). On note
   - m_k les effectifs du groupe k, $k \in \{1, ..., K\}$;
   - o_k le nombre de succès $(Y = 1)$ observés dans le groupe k ;
   - μ_k la moyenne des $p_\beta(x_i)$ dans le groupe k.

La statistique de test est alors

$$C^2 = \sum_{k=1}^{K} \frac{(o_k - m_k\mu_k)^2}{m_k\mu_k(1 - \mu_k)}. \tag{3.25}$$

Sous l'hypothèse nulle H_0 : le modèle est adéquat, la statistique C^2 suit approximativement une loi $\mathcal{X}^2$ à $K - 1$ degrés de liberté.
Lorsque la probabilité critique du test (p-value) est plus grand que le risque choisi (souvent c'est 0.05), le modèle issu de la régression logistique est accepté.

3.9 La courbe ROC et le critère AUC

Lorsqu'on fait varier le seuil s de décision (seuil de probabilité au-delà duquel on décidera de prédire une défaillance), la sensibilité et la spécificité (cf. section 4.2) changent puisque la règle de classement a été changée. Pour présenter les valeurs pour toutes les possibilités de seuils, on dessine une courbe qui décrit la sensibilité en fonction de la spécificité, c'est la courbe ROC.
La courbe ROC consiste à présenter la probabilité de prédire un vrai défaillance (sensibilité) d'un faux défaillance (1−spécificité) pour un ensemble de seuils possibles et à joindre par une courbe.
L'aire sous la courbe nommé **AUC** qui varie entre 0 et 1, fournit une mesure de la capacité du modèle à discriminer les bons payeurs (pas de défaillance) et les mauvais payeurs.

La règle générale :

- Si AUC= 0, on considère qu'il n' y pas de discrimination.
- Si $0.7 \leq$ AUC < 0.8, la discrimination est acceptable.
- Si $0.8 \leq$ AUC < 0.9, la discrimination est considérée excellente .
- Si AUC ≤ 0.9, la discrimination est parfaite.

chapitre 4

Modélisation

Dans ce chapitre, on s'intéresse à la modélisation du crédit scoring par un traitement de nos données, en utilisant la méthode paramétrique exposée dans le chapitre 3. Compte tenu de nombreux outils pouvant être utilisés, nous avons fait le choix d'insister sur la pratique de des méthodes étudiées et la compréhension des sorties par le logiciel **R** (la version 2.5.1). Nous allons estimer une fonction de scoring par la régression logistique et nous estimerons l'erreur de prédiction.

4.1 Régression Logistique

4.1.0.1 Échantillon d'apprentissage et de validation

Notre base d'étude (German Crédit Data) comporte 1000 observations et 31 variables (cf.tableau 3.2). Cette base ne comporte pas des valeurs manquantes.
La réalisation de scoring nécessite de diviser notre base (pour éviter l'overfitting ou le sur-apprentissage) en deux échantillon : l'échantillon apprentissage (nécessaire pour l'estimation) et l'échantillon test (utilisé pour l'analyse des performances). Nous avons effectué un tirage aléatoire simple sans remise, l'échantillon apprentissage représente 60% de la base de départ.

4.1.1 Estimation du modèle

4.1.1.1 Avant selection

* On remarque que les variables qualitatives ont été traitées comme des variables numériques : On remédie à cela en les transformant en type facteur.

L'estimation s'effectue sur notre échantillon d'apprentissage. On va construire le modèle complet noté **model-all** (contenant toutes les variables), en utilisant la fonction :
glm(formula, family=binomial(link='logit'), data=)) du logiciel **R**.
On testera si les variables contribuent significativement à la construction de la probabilité de ne pas être défaillant en regardant les p-values (p-values doivent être inférieures à notre seuil de confiance $\alpha = 0.05$.). Le tableau suivant résume les sorties numériques des variables qui sont significatives de la régression logistique du modèle complet.

Varibales	Coef.Estimate	Std.Error	Z value	p-value
CHK-ACCT3	2.040	0.3328	6.129	$8.82e-10$
History4	1.845	0.6276	2.940	0.003278
Duration	-0.04303	0.01295	-3.323	0.00089
Sav-ACCT3	1.305	0.06382	2.046	0.040790
SAV-ACCT4	0.09496	0.3552	2.673	0.00751
Employment	1.418	0.635	2.233	0.025556
Install-Rate	-0.353	0.1191	-2.816	0.004867
Male single1	0.5884	0.2969	1.982	0.047528
Num-credits	-0.5944	0.2496	-2.382	0.017231

TABLE 4.1 – Coefficients significatives du modèle logistique

En déterminant la matrice de corrélation entre les variables quantitatives de la base d'étude, on remarque qu'il y avait une corrélation moyenne (coefficient de corrélation= 0.62) entre les deux variables **Duration** et **AMOUNT**.

4.2 Sélection de modèle

La construction du modèle logistique se fait en utilisant la fonction **glm** du logiciel **R**, ensuite une sélection automatique des variables par minimisation de l'AIC est faite par la fonction **step** de la bibliothèque **Mass**. Pour la sélection des variables pertinentes du modèle, nous avons effectué une méthodes de sélection automatique descendante (backward selection) sur le jeu de donnée d'apprentissage,(voir la section 3.7).

backward selection :
Le modèle construit a retenu les variable **CHK-ACCT, Duration, History, New car, Used car, Education, Amount, Sav-Acct, Employement, Install-rate, Male single, Co-Applicant, Gurantor, Age, Other-install, Num credits, Foriegn** comme

significatives.
On effectuera la régression logistique sur le modèle sélectionné et on testera la performance du modèle, en utilisant les tests statistiques présentés dans la section 3.8.

4.3 Validation

Matrice de confusion :
Pour le modèle discriminant fournit , sur les données de test, on résume les erreurs de classement par un tableau de contingence :

$Y/\hat{Y}$	1	0
1	237	60
0	34	69

TABLE 4.2 – Matrice de confusion

L'erreur estimé sous R (nombre de mauvais classement $60 + 34 = 94$ rapporté à l'effectif total) vaut $\epsilon = 0.23$, nous jugeons l'erreur acceptable (faible), soit prés 78% des observations classées correctement ce qui est acceptable.
Test de rapport de vraisemblance :
On s'intéresse aussi à la déviance du modèle sélectionné, la p-value simulée vaut 1.2×10^{-28}. Alors le modèle construit est jugé significative par rapport au modèle réduit à une constante.

Courbe ROC : En examinant la performance du modèle logistique sélectionnée par la courbe ROC, avec la bibliothèque **ROCR**.
En estimant le critère AUC (l'aire sous la courbe ROC) pour les deux bases de données apprentissage et test, on aura AUC-Apprentissage= 0.83 et AUC-test=0.78, on en déduit alors que le modèle logistique construit est performant.

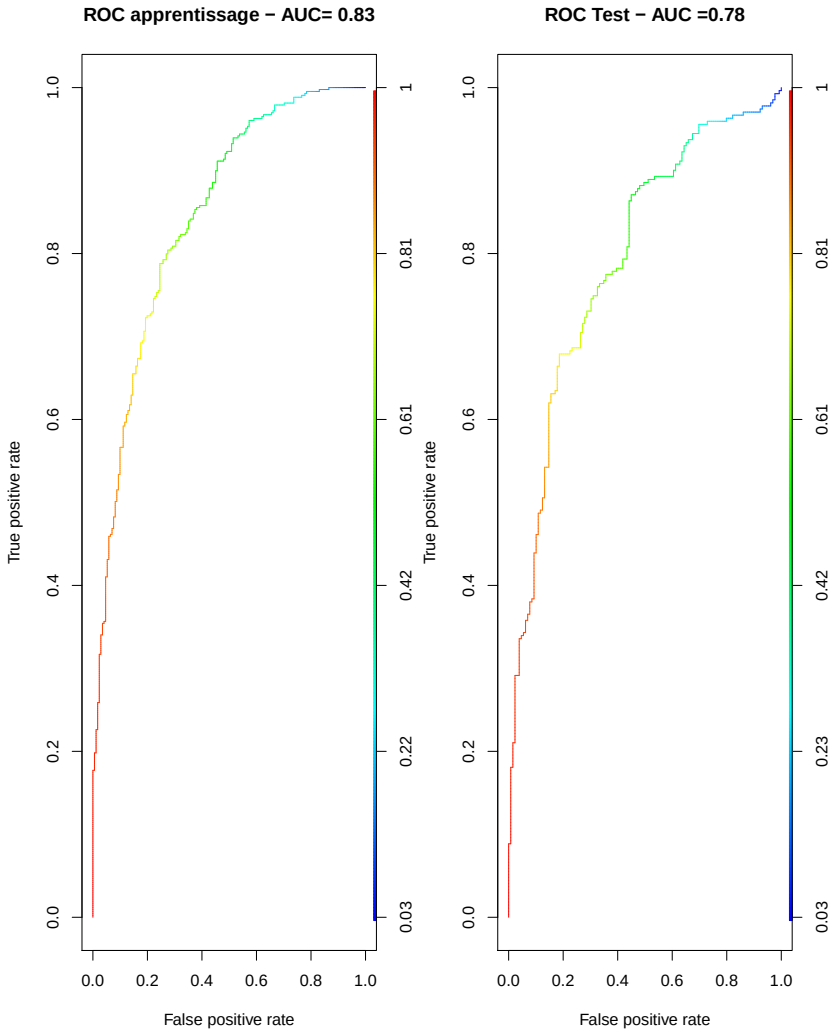


FIGURE 4.1 – Courbe ROC sur les deux base de données

Conclusion

Le scoring pour la régression logistique est :

$S(X) = 1.535 + 0.3495 \times \mathbb{1}_{\{CHK_ACCT=1\}} + 0.7336 \times \mathbb{1}_{\{CHK_ACCT=2\}} + 1.995 \times \mathbb{1}_{\{CHK_ACCT=3\}} - 0.04374 \times DURATION + 0.2437 \times \mathbb{1}_{\{HYSTORY=1\}} + 0.9159 \times \mathbb{1}_{\{HYSTORY=2\}} + 1.036 \times \mathbb{1}_{\{HYSTORY=3\}} + 1.722 \times \mathbb{1}_{\{HYSTORY=4\}} - 0.8 \times \mathbb{1}_{\{NEWCAR=1\}} + 0.853 \times \mathbb{1}_{\{USEDCAR=1\}} - 1.17\ 10^{-4} \times AMOUNT - 0.625 \times \mathbb{1}_{\{SAVACCT=1\}} + 0.6131 \times \mathbb{1}_{\{SAVACCT=2\}} + 1.296 \times \mathbb{1}_{\{SAVACCT=3\}} + 0.963 \times \mathbb{1}_{\{SAVACCT=4\}} - 0.348 \times \mathbb{1}_{\{EMPLOYMENT=1\}} + 0.285 \times \mathbb{1}_{\{EMPLOYMENT=2\}} + 0.852 \times \mathbb{1}_{\{EMPLOYMENT=3\}} + 0.231 \times \mathbb{1}_{\{EMPLOYMENT=4\}} - 0.3185 \times INSTALRATE + 0.599 \times \mathbb{1}_{\{MALE-SINGLE=1\}} - 0.755 \times \mathbb{1}_{\{CO-APPLICANT=1\}} + 0.977 \times \mathbb{1}_{\{GAURANTOR=1\}} - *0.386 \times \mathbb{1}_{\{PROP-UNKN-NONE=1\}} + 0.002 \times AGE - 0.478 \times \mathbb{1}_{\{OTHER-INSTALL=1\}} - 0.366 \times \mathbb{1}_{\{RENT=1\}} - 0.563 \times NUMCREDITS - 0.472 \times NUM -$

$DEPENDENTS + 1.31 \times \mathbb{1}_{\{FOREIGN=1\}}$.

Pour un nouvel emprunteur ou un client de crédit,

· On recueille auprès du client les variables retenues (les informations)

· On calcule son scoring $S(X)$ (combinaisons linéaires de ces variables) qui est une probabilité de défaut.

· On considère le seuil $s = \frac{1}{2}$, avec une erreur de 0.23 de mauvais classement, on a le choix d'affectation d'un nouvel emprunteur de la manière suivante :

1. Si $S(X) < \frac{1}{2}$ alors $Y = 0$, ie que l'emprunteur est non risqué (un bon payeur).
2. Si $S(X) > \frac{1}{2}$ alors $Y = 1$, ie que l'emprunteur est risqué, il est mauvais client.
3. Si $S(X) = \frac{1}{2}$ alors $Y = 1$ ou $Y = 0$, peu importe.

Annexe A

Codes R des fonctions utilisées

Importation des données :

```
DATA="http://www.math.unicaen.fr/~kauffman/data/"
X=read.table(paste(DATA,"GermanCredit.csv",sep=""),header=TRUE,sep=";")
str(X)
Attach(X)
install.packages(ROCR) # Pour la courbe ROCR
library(ROCR)
install.packages("MASS")
library(MASS)
```

Analyse exploratoire :

On remarque que les variables qualitatives ont été traitées comme des variables numériques :
On remédie à cela en les transformant en type facteur.

```
X$CHK_ACCT <- as.factor(CHK_ACCT)
X$HISTORY  <- as.factor(HISTORY)
X$NEW_CAR  <- as.factor(NEW_CAR)
X$USED_CAR <- as.factor(USED_CAR)
X$FURNITURE<- as.factor(FURNITURE)
X$RADIO_TV <- as.factor(RADIO_TV)
```

Les histogrammes :

```
par(mfrow = c(4,2))
hist(x=X$AMOUNT,col="green3", main="Montant", xlab="",ylab="")
hist(x=X$DURATION,col="red", main="Duree", xlab="",ylab="")
hist(x=X$NUM_CREDITS ,col="yellow", main="Nombre de credits",
```

```
xlab="",ylab=" ")
hist(x=X$NUM_DEPENDENTS ,col="blue", main="Nombre de personnes",
xlab="",ylab="")
hist(x=X$INSTALL_RATE, col="magenta",main="Pourcentage", xlab = "",
ylab="")
```

Présélection des variables et études des liaisons :

```
quanti=cbind(AGE,AMOUNT,DURATION,NUM_CREDITS,NUM_DEPENDENTS,INSTALL_RATE)
cor(quanti, use = "complete.obs")
acp<- dudi.pca(quanti)
s.corcircle(acp$co, xax = 1, yax = 2)
```

Data spliting : apprentissage vs test :

```
set.seed(111)
d = sort(sample(nrow(X), nrow(X) * 0.6))
appren=X[d, ]
test=X[-d, ]
summary(appren)
attach(test)
attach(appren)
```

Modélisation :

```
model_all <- "RESPONSE~."
MODEL<-glm(RESPONSE~.,data=appren,family = binomial())
summary(MODEL)
```

Sélection automatique "forward" et "backward" :

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

modele <- glm(RESPONSE ~., data = appren, family = binomial)
modele.backward<- stepAIC(modele, scope = list(lower = str_constant,
upper = str_all),trace = TRUE, data = appren, direction = "backward")
summary(modele.backward)
```

```
m.logit <-glm(RESPONSE~CHK_ACCT + DURATION + HISTORY + NEW_CAR +
USED_CAR + AMOUNT + SAV_ACCT + EMPLOYMENT + INSTALL_RATE +
MALE_SINGLE + CO_APPLICANT + GUARANTOR + PROP_UNKN_NONE +
AGE + OTHER_INSTALL + RENT + NUM_CREDITS + NUM_DEPENDENTS +
FOREIGN, family = binomial, data = appren)
summary(m.logit)
exp(cbind(OR = coef(m.logit), confint(m.logit)))
```

Validation du modèle :

```
par(mfrow = c(1, 1))
plot(rstudent(m.logit), type = "p", cex = 0.5, ylab =
 "Residus studentises",col = "springgreen2", ylim = c(-3, 3))
abline(h = c(-2, 2), col = "red")
# Test de rapport de vraisemblance
(chi2 <- with(m.logit, null.deviance - deviance))
(ddl <- with(m.logit, df.null - df.residual))
(pvalue <- pchisq(chi2, ddl, lower.tail = F))

# Matrice de confusion et estimation d'erreur
appren.p <- cbind(appren, predict(m.logit, newdata = appren,
 type = "link",  se = TRUE))
head(appren.p)
tail(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)
appren.p <- cbind(appren.p, pred.RESPONSE=factor
(ifelse
(appren.p$PredictedProb >  0.5, 1, 0)))
head(appren.p)
(m.confusion <- as.matrix(table(appren.p$pred.RESPONSE
, appren.p$RESPONSE)))
m.confusion <- unclass(m.confusion)
# Calcul du taux d'eureur.
```

```
Tx_err <- function(y, ypred)
{
  mc <- table(y, ypred)
  error <- (mc[1, 2] + mc[2, 1])/sum(mc)
  print(error)
}
Tx_err(appren.p$pred.RESPONSE ,appren.p$RESPONSE) # eureur=0.2

test.p <- cbind(test, predict(m.logit, newdata = test
, type = "response", se = TRUE))
test.p <- cbind(test.p, pred.RESPONSE <- factor(ifelse
(test.p$fit > 0.5, 1, 0)))
(m.confusiontest <- as.matrix(table(test.p$pred.RESPONSE
, test.p$RESPONSE)))
m.confusiontest <- unclass(m.confusiontest)
# Calcul du taux d'eureur.
Tx_err <- function(y, ypred)
 {
  mc <- table(y, ypred)
  error <- (mc[1, 2] + mc[2, 1])/sum(mc)
  print(error)
}
Tx_err(test.p$pred.RESPONSE, test.p$RESPONSE) # erreur= 0.235.
```

Courbe ROC :

```
Pred = prediction(appren.p$PredictedProb, appren.p$RESPONSE)
Perf = performance(Pred, "tpr", "fpr")
plot(Perf, colorize = TRUE, main = "ROC apprentissage")
perf <- performance(Pred, "auc")
perf@y.values[[1]]  # l'aire:0.84  (Discrimination excellente).
Predtest = prediction(test.p$fit, test.p$RESPONSE)
Perftest = performance(Predtest, "tpr", "fpr")
perftest <- performance(Predtest, "auc")
perftest@y.values[[1]]
par(mfrow = c(1, 2))
plot(Perf, colorize = TRUE, main = "ROC apprentissage - AUC= 0.83")
plot(Perftest, colorize = TRUE, main = "ROC Test - AUC =0.78")
```

Bibliographie

[1] Azais Romain, cours d'apprentissage (2016-2017). Faculté des sciences et technologies de Nancy, université de Lorraine.

[2] Anne. C, Chaigneau. G. Mathématiques financières. Ellipses, 2007.

[3] Blayac Thierry. cours de L3-sciences économiques. Université Montpellier I, 2011-2012. `http://eco.um1.free.fr/cours.php?id=52`.

[4] M. Bardos. Analyse discriminante, application au risque et scoring financier. Dunod, 2001.

[5] Cassabalian Jean-Louis. Mathématiques et calculs financiers sur tableur et internet. ESKA, 2000.

[6] Cornillon Pierre-André, Eric Matzner-Lober. Régression avec R. Springer, 2011.

[7] Cadre B. Statistique mathématique pour le master 1. cours de l'ENS Cachan Bretagne, 40 pages. `//w3.bretagne.enscachan.fr/math/people/benoit.cadre/`.

[8] Diouri Mohamed. Mathématiques Financières. L'IGA Institut supérieur du Génie Appliqué et les Éditions TOUBKAL, 2001.

[9] T. Hastie, R. Tibshirani, J. Friedman, The elements of Statistical Learning - Data Mining, Inference and Prediction, Springer, 2001.

[10] Fahrmier L., Kaufmann H. Consistency and asymptotic normality of the maximum liklihood estimator in generalized linear models. The Annals of statistics, **13**, 342-368.

[11] Rouvière Laurent. Régression logistique avec R. Université Rennes2, UFR sciences sociales. `https://perso.univ-rennes2.fr/system/files/users/rouviere_l/poly_logistique_web.pdf`.

[12] D.W. Hosmer, s.Lemshow. Applied Logistic Regression, Second Edition, Wiley, 2000.

[13] S. Menard, Applied Logistic Regression Analysis (Second Edition), Series : Quantitative Applications in the Social Sciences, Sage Publications, 2002.

[14] Saporta Gilbert. La Notation Statistique des Emprunteur ou « scoring ».

[15] Saporta Gilbert. Probabilités, analyse des données et statistique. Editions TECHNIP, 2011.

[16] W. David, Hosmer Stanley Lemeshow. Applied Logistic Regression, second edition, Wiley, 2000.