Université de Lille
Crédit Agricole Consumer Finance - Inria Lille-Nord Europe

École doctorale **Sciences pour l'Ingénieur**
Unité de recherche **Équipe-projet MΘDAL**

Thèse présentée par **Adrien Ehrhardt**

Soutenue le ⟨*date de la soutenance*⟩

En vue de l'obtention du grade de docteur de l'Université de Lille

Discipline **Mathématiques et leurs interactions**
Spécialité **Statistique**

Titre de la thèse

Modèles prédictifs pour données volumineuses et biaisées

Application à l'amélioration du scoring en risque de crédit

Thèse dirigée par	Christophe Biernacki	directeur
	Philippe Heinrich	co-encadrant
	Vincent Vandewalle	co-encadrant

Composition du jury

Rapporteurs	⟨*Prénom*⟩ ⟨*Nom*⟩		
	⟨*Prénom*⟩ ⟨*Nom*⟩		
Examinateurs	⟨*Prénom*⟩ ⟨*Nom*⟩		président du jury
	⟨*Prénom*⟩ ⟨*Nom*⟩		
	⟨*Prénom*⟩ ⟨*Nom*⟩		
	⟨*Prénom*⟩ ⟨*Nom*⟩		
Invité	Jérôme Beclin		
Directeurs de thèse	Christophe Biernacki	professeur à l'Université de Lille	
	Philippe Heinrich	mcf à l'Université de Lille	
	Vincent Vandewalle	mcf à l'Université de Lille	

Université de Lille
Crédit Agricole Consumer Finance - Inria Lille-Nord Europe

École doctorale **Sciences pour l'Ingénieur**
Unité de recherche **Équipe-projet MΘDAL**

Thèse présentée par **Adrien Ehrhardt**

Soutenue le ⟨*date de la soutenance*⟩

En vue de l'obtention du grade de docteur de l'Université de Lille

Discipline **Mathématiques et leurs interactions**
Spécialité **Statistique**

Titre de la thèse

Modèles prédictifs pour données volumineuses et biaisées

Application à l'amélioration du scoring en risque de crédit

Thèse dirigée par	Christophe Biernacki	directeur
	Philippe Heinrich	co-encadrant
	Vincent Vandewalle	co-encadrant

Composition du jury

Rapporteurs	⟨*Prénom*⟩ ⟨*Nom*⟩		
	⟨*Prénom*⟩ ⟨*Nom*⟩		
Examinateurs	⟨*Prénom*⟩ ⟨*Nom*⟩		président du jury
	⟨*Prénom*⟩ ⟨*Nom*⟩		
	⟨*Prénom*⟩ ⟨*Nom*⟩		
	⟨*Prénom*⟩ ⟨*Nom*⟩		
Invité	Jérôme Beclin		
Directeurs de thèse	Christophe Biernacki	professeur à l'Université de Lille	
	Philippe Heinrich	mcf à l'Université de Lille	
	Vincent Vandewalle	mcf à l'Université de Lille	

Université de Lille
Crédit Agricole Consumer Finance - Inria Lille-Nord Europe

Doctoral School **Sciences pour l'Ingénieur**
University Department **Équipe-projet MΘDAL**

Thesis defended by **Adrien Ehrhardt**

Defended on ⟨*defense date*⟩

In order to become Doctor from Université de Lille

Academic Field **Applied Mathematics**
Speciality **Statistics**

Thesis Title

Predictive models for big and biased data

Application to Credit Scoring

Thesis supervised by	Christophe Biernacki	Supervisor
	Philippe Heinrich	Co-Monitor
	Vincent Vandewalle	Co-Monitor

Committee members

Referees	⟨*Prénom*⟩ ⟨*Nom*⟩		
	⟨*Prénom*⟩ ⟨*Nom*⟩		
Examiners	⟨*Prénom*⟩ ⟨*Nom*⟩		Committee President
	⟨*Prénom*⟩ ⟨*Nom*⟩		
	⟨*Prénom*⟩ ⟨*Nom*⟩		
	⟨*Prénom*⟩ ⟨*Nom*⟩		
Guest	Jérôme Beclin		
Supervisors	Christophe Biernacki	Professor at Université de Lille	
	Philippe Heinrich	Associate Professor at Université de Lille	
	Vincent Vandewalle	Associate Professor at Université de Lille	

L'Université de Lille n'entend donner aucune approbation ni improbation aux opinions émises dans les thèses : ces opinions devront être considérées comme propres à leurs auteurs.

Mots clés : scoring, risque, crédit, prédiction, discrétisation, segmentation
Keywords: scoring, credit, risk, predictin, discretization, clustering

Cette thèse a été préparée dans les laboratoires suivants.

Équipe-projet MΘDAL
Inria Lille Nord-Europe
40 Avenue Halley
59650 Villeneuve-d'Ascq

☎ +33 (0)3 59 57 78 00
+33 (0)3 59 57 78 50
contact-lille@inria.fr
Site https://www.inria.fr/centre/lille

Laboratoire Paul Painlevé
CNRS U.M.R. 8524
59655 Villeneuve d'Ascq Cedex
France

☎ (+33) 03 20 43 48 50
Site https://math.univ-lille1.fr/

À ma famille,

À mes amis,

À mes lecteurs.

The task of the human brain remains what it has always been ; that of discovering new data to be analyzed, and of devising new concepts to be tested.

Isaac Asimov, *I, Robot*

J'respecte R.

Damso

Modèles prédictifs pour données volumineuses et biaisées
Application à l'amélioration du scoring en risque de crédit

Résumé

Le ratio risque/récompense désigne en finance la logique selon laquelle un investissement peu risqué ne pourra être que faiblement rentable tandis qu'un investissement risqué a un rendement plus élevé mais est exposé à une perte. Les établissements financiers spécialisés en crédit à la consommation transposent ce principe en deux heuristiques : premièrement, le taux d'intérêt des crédits est adapté en fonction des clients et des produits ; deuxièmement, les clients demandeurs sont sélectionnés selon leur solvabilité. Ce mécanisme d'acceptation/rejet de la clientèle est composé de plusieurs règles de décision dont un score, c'est-à-dire une notation témoignant de la probabilité de défaut d'un nouveau client. La construction de ce score, qu'on désigne généralement par *Credit Scoring*, repose sur des techniques statistiques relativement anciennes et des heuristiques industrielles dont certaines ont été examinées dans cette thèse.
Après une première partie décrivant l'évolution et le contexte industriels actuels ainsi que la littérature académique associée, on s'intéressera dans une deuxième partie à une contribution importante de cette thèse : la "réintégration des refusés" ou comment tirer partie des informations collectées sur les clients refusés mais non utilisées. On verra ensuite en troisième partie l'apport de la méthode proposée dans cette thèse pour la discrétisation (resp. regroupement de modalités) des variables quantitatives (resp. qualitatives) constitutives du score ainsi que l'introduction d'interactions sur sa qualité. Enfin, la quatrième partie .
L'ensemble des travaux est illustré par des données réelles de Crédit Agricole Consumer Finance, établissement bancaire spécialiste du crédit à la consommation à l'origine de cette thèse CIFRE.

Mots clés : scoring, risque, crédit, prédiction, discrétisation, segmentation

---

Predictive models for big and biased data
Application to Credit Scoring

Abstract

The risk-reward is a well known finance paradigm: the higher the risk of an investment, the higher the expected reward . When it comes to consumer loans,

Keywords: scoring, credit, risk, predictin, discretization, clustering

---

Équipe-projet MΘDAL
Inria Lille Nord-Europe – 40 Avenue Halley – 59650 Villeneuve-d'Ascq

Remerciements

Aboutissement d'un travail personnel, cette thèse n'en est pas moins une réussite collective et la contribution de nombreuses personnes, injustement absente de la page de couverture de ce manuscrit, doit ici être extensivement mentionnée.

Tout d'abord, je suis persuadé que le principal facteur de succès d'une thèse CIFRE est l'implication de l'entreprise d'accueil, de la conception du sujet à l'usage des fruits du travail de recherche. A ce titre, je remercie Crédit Agricole Consumer Finance de m'avoir permis de réaliser cette thèse dans de très bonnes conditions. En particulier, j'ai eu la chance d'interagir avec des managers réceptifs à la démarche de recherche et qui m'ont fait confiance : un grand merci à Jérôme Beclin et Nicolas Borde. Je me dois également de saluer la probité intellectuelle de Sébastien Beben ; nos riches échanges de début de thèse constituent sans doute le carburant de ce doctorat.

Haut-lieu de la recherche publique française, Inria m'a permis, en acceptant d'être le laboratoire d'accueil de cette CIFRE, de compléter ma formation d'ingénieur généraliste centralien en tentant de combler le vide technique ressenti en fin de cursus, ce qui m'avait motivé à poursuivre en thèse. Je vous laisse le soin, chers lecteurs, d'apprécier l'éventuelle réussite de cet objectif initial. Je remercie chaleureusement le centre de Lille et plus particulièrement l'équipe-projet MΘDAL pour m'avoir permis de (re)connaître la beauté des mathématiques. Contributeurs directs et véritables artisans de ce travail de recherche, mes trois co-directeurs de thèse ont constitué le moteur de cette thèse ; merci à Christophe Biernacki dont j'espère garder la rigueur scientifique ; merci à Philippe Heinrich, pour m'avoir montré qu'un problème bien posé est déjà à moitié résolu ; merci à Vincent Vandewalle, dont les éclairages passionés, à grands coups de feutre virevoltant sur le tableau ou scripts R envoyés au milieu de la nuit, ont pour la plupart donné la vitesse initiale à chaque partie de ce manuscrit.

Aussi puissant et bien alimenté qu'il soit, un véhicule est peu de choses sans ses quatre roues. Les quelques mots qui suivront sont bien peu de choses en comparaison de la stabilité

Sommaire

Liste des tableaux

Table des figures

Glossaire

C | L

C

crédit affecté Le crédit affecté est accordé par un établissement de crédit ou une banque. Il est utilisé pour un achat déterminé : un bien mobilier (crédit automobile par exemple) ou une prestation. Il est souvent contracté directement sur le lieu de vente. Généralement, le défaut du crédit entraîne la récupération du bien sous-jacent par un huissier.. 5, 6

crédit classique Les conditions du prêt sont fixées à l'avance, lors de la signature du contrat. Le taux, la durée, et les mensualités du prêt sont fixes. Le coût total du financement est ainsi connu dès le début du prêt.. 5

crédit renouvelable Le crédit renouvelable, encore appelé crédit permanent, crédit revolving ou crédit reconstituable, consiste à mettre à la disposition d'un emprunteur une réserve d'argent qu'il pourra utiliser et reconstituer selon son gré. Ce crédit est proposé par un établissement financier ou une enseigne commerciale. Il peut être couplé avec une carte de crédit et peut être couvert par une assurance.. 6

L

location La location est elle-même commercialisée sous deux formes : la location avec option d'achat (L.O.A.), pour laquelle le client peut décider d'acquérir le bien loué en fin d'échéancier pour un montant d'option d'achat fixé à l'avance et la location longue-durée (L.L.D.) pour laquelle c'est le magasin / concessionnaire qui dispose d'une option d'achat.. 6

Symboles

$\mathbb{N}$	entiers naturels	7
$\mathbb{R}$	nombres réelles	7
$\boldsymbol{X}$	réalisation de X	7
$\boldsymbol{x}$	vecteur de caractéristiques d'un client	7

Introduction

Les cas d'application des travaux de ce manuscrit portent sur plusieurs problèmes connexes au *Credit Scoring*.

Pour un particulier, le recours au crédit, c'est-à-dire à l'emprunt d'argent en échange d'une promesse de remboursement étalé dans le temps et assorti d'un intérêt, est possible depuis très longtemps, les plus anciennes traces "modernes" de crédits bancaires se situant au XII^{ème} siècle en Italie [3]. De nos jours, l'emprunt immobilier ou automobile, c'est-à-dire pour financer un lieu de résidence ou l'achat d'un véhicule, est répandu [1]. Par opposition au crédit immobilier, on parle souvent de crédit à la consommation pour désigner le financement de biens et de services : automobile, électroménager, travaux, etc. De manière plus formelle, le crédit à la consommation est définie dans la loi №2010-737 du 1er juillet 2010 [2] comme une :

> Opération ou contrat de crédit, une opération ou un contrat par lequel un prêteur consent ou s'engage à consentir à l'emprunteur un crédit sous la forme d'un délai de paiement, d'un prêt, y compris sous forme de découvert ou de toute autre facilité de paiement similaire, à l'exception des contrats conclus en vue de la fourniture d'une prestation continue ou à exécution successive de services ou de biens de même nature et aux termes desquels l'emprunteur en règle le coût par paiements échelonnés pendant toute la durée de la fourniture.

De nombreux acteurs bancaires proposent des crédits à la consommation, si bien qu'en 2013 environ 26,6 % des ménages ont un crédit à la consommation []. Crédit Agricole Consumer Finance (CACF) est un acteur majeur du crédit à la consommation, à travers une marque spécialisée en France, Sofinco, et des partenaires distributeurs de crédit conso.

Parmi l'ensemble des demandeurs de crédit à la consommation, il est souhaitable, à plusieurs égards, de ne pas financer tous les crédits. Premièrement, si tant est que l'on puisse prêter un rôle sociétal à une entité bancaire, il paraît responsable de ne pas détériorer voire mettre en danger la santé financière de l'emprunteur. Pour ce faire, des contrôles automatiques permettent de refuser la clientèle dite fragile : taux d'endettement trop élevé, fichage bancaire pour incidents de paiements, ... Par ailleurs, d'un point de vue économique cette fois, un client se trouvant dans l'incapacité de rembourser le crédit souscrit sera vraisemblablement peu ou pas profitable pour l'institution financière du fait des coûts de traitements et de personnels de relance et procédure(s) judiciaire(s) qui peuvent aboutir à une annulation totale ou partielle de la dette du client engendrant une perte sèche pour l'organisme prêteur.

Dans ce cadre, le *Credit Scoring* vise à évaluer la propension d'un client à être "bon" ou "mauvais", selon des critères à définir ultérieurement, pour ainsi prendre une décision de financement ou de rejet de façon quantitative et objective. On donnera dans le chapitre 1 quelques éléments de contexte supplémentaires nécessaires à la bonne compréhension des cas d'application de cette thèse, un état de l'art de la pratique industrielle ainsi qu'un état de l'art académique des techniques d'apprentissage transposables au *Credit Scoring*.

Le chapitre 2 est consacré à l'étude du problème de "Réintégration des refusés" (ou *Reject Inference*) qui peut être réinterprété comme un biais d'échantillon

Ce problème d'échantillonnage résolu, il paraît naturel au statisticien de s'atteler à la modélisation : quelle relation existe-t-il entre les caractéristiques de l'emprunteur et la quantité de risque qu'il présente ? Le chapitre 1 aura mis en avant certaines faiblesses statistiques de la procédure actuelle : le chapitre **??** met en oeuvre une nouvelle méthode de recherche et sélection du meilleur modèle dans la famille imposée par le cas d'application.

à compléter avec le(s) dernier(s) chapitre(s)

Références de l'introduction

[1] *Les Français recourent toujours largement au crédit pour acheter leur voiture*. Oct. 2010. URL : https://www.latribune.fr/vos-finances/banques-credit/credit-auto-moto/20101007trib000556639/les-francais-recourent-toujours-largement-au-credit-pour-acheter-leur-voiture.html.

[2] *LOI n°2010-737 du 1er juillet 2010 portant réforme du crédit à la consommation (1)*. Juil. 2010. URL : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000022419094&categorieLien=id.

[3] H. Thomas. *The Wards of London : Comprising a Historical and Topographical Description of Every Object of Importance Within the Boundaries of the City. With an Account of All the Companies, Institutions, Buildings, Ancient Remains ... and Biographical Sketches of All Eminent Persons Connected Therewith*. The Wards of London : Comprising a Historical and Topographical Description of Every Object of Importance Within the Boundaries of the City. With an Account of All the Companies, Institutions, Buildings, Ancient Remains ... and Biographical Sketches of All Eminent Persons Connected Therewith vol. 1 à 2. J. Gifford, 1828. URL : https://books.google.fr/books?id=PDMQAAAAYAAJ.

Chapitre 1

Apprendre des demandes de crédit à la consommation

Ce chapitre est destiné à poser les bases de l'apprentissage statistique dans le cadre des crédits à la consommation. On introduira dans une première partie une partie de la terminologie consacrée aux crédits à la consommation avant de s'attarder plus en détails, dans une seconde partie, sur l'état de l'art industriel du *Credit Scoring* à travers une étude bibliographique et la pratique de CACF. On clotûrera le chapitre par une troisième partie, la plus traditionnelle pour débuter un manuscrit de thèse, à savoir l'état de l'art académique sur l'apprentissage statistique, en nous limitant bien entendu aux cas d'usage spécifiques aux crédits à la consommation mis en exergue dans les deux premières parties de ce chapitre.

1.1 Le marché du crédit à la consommation : quels enjeux ?

S'agissant d'une thèse CIFRE, il apparaît comme nécessaire de planter le décor industriel de la problématique. Dans cette première partie, on verra succintement le coeur du métier de CACF, les produits qu'elles proprosent et l'environnement dans lequel elle s'insert.

1.1.1 Qu'est-ce qu'un crédit à la consommation ?

La définition légale en a été donnée en 2. En pratique, on peut distinguer trois produits de crédit à la consommation.

Le premier d'entre eux, le crédit classique est le produit historique. De la même manière qu'un crédit immobilier, le client emprunte une somme fixe qui lui est attribuée au financement et qu'il rembourse selon un échéancier (taux et nombre de mensualités fixes) défini à l'avance. D'un point de vue statistique, le traitement est relativement simple : que ce soit à l'octroi, pour déterminer le risque du client, ou au cours de la vie du dossier, pour provisionner les pertes potentielles, tout est connu à l'avance. Il suffit en quelque sorte de vérifier le paiement de la mensualité à la date prévue. Il convient également de préciser que certains crédits classiques sont dits crédits affectés, c'est-à-dire qu'ils financent un bien précis et identifié, de sorte que le prêt transite directement de l'organisme prêteur au vendeur (concessionnaire par exemple). Par ailleurs, la mise en défaut du crédit entraîne généralement une procédure de recouvrement de la dette qui peut se solder, dans le cas d'un crédit affecté, par la récupération du bien par

un huissier. Là encore, d'un point de vue statistique, il paraît indispensable de consigner les caractéristiques du bien sous-jacent afin d'intégrer sa valeur résiduelle récupérable en cas de défaut.

Le second produit, développé à partir de 1965 en France et ayant connu une forte croissance depuis [2] mais néanmoins bien moins répandu en Europe qu'aux Etats-Unis par exemple [5], est le crédit renouvelable. Un capital dit accordé ou autorisé est attribué au demandeur qui peut utiliser tout ou partie de ce montant et le rembourse à un taux et par mensualités dépendants tous deux de la proportion du capital consommé. Au fur et à mesure du remboursement du capital emprunté, le capital "empruntable", c'est-à-dire la différence entre le capital accordé et le capital emprunté puis remboursé, se reconstitue et de nouvelles utilisations sont possibles, toujours dans la limite du capital accordé au départ. D'un point de vue statistique à nouveau, plusieurs problèmes se posent du fait du caractère intrinsèquement aléatoire de l'utilisation ou non de tout ou partie de la ligne de crédit accordée. Plus précisément, ce produit présente un risque important porté par deux facteurs : premièrement, le taux élevé attire des clients risqués, au taux de défaut plus élevé que pour un crédit classique par exemple ; deuxièmement, ces crédits portent un risque dit de hors-bilan très fort, puisqu'à tout moment, l'ensemble des crédits accordés mais non utilisés et donc non comptabilisés "au bilan" c'est-à-dire comme une dette du client envers l'établissement bancaire, peuvent être utilisés et faire défaut. La mauvaise quantification de ce risque est à présent reconnu comme un important catalyseur de la récente crise financière [3].

Enfin, la location a récemment connu un essor important [4]. D'abord concentrée sur le secteur automobile, elle se développe actuellement pour les produits électroniques (smartphones notamment) et même plus récemment pour des produits plus insolites comme les matelas [1]. Comme le crédit affecté, il est important de prendre en compte les données du bien loué afin d'évaluer le risque que porte ce produit, la difficulté supplémentaire reposant sur l'éventualité de l'exercice de l'option d'achat.

De cette partie, deux considérations statistiques doivent retenir notre attention : d'abord, ces différents produits nécessitent des traitements différents dans la mesure où leur risque est intrinsèquement différent ; ensuite, les données disponibles pour chacun de ces produits diffèrent : par exemple, les données du produit financé ne sont disponibles que pour les crédits affectés et les locations. Cette dernière notion de "blocs" de variables est au coeur du chapitre **??**.

1.1.2 Crédit Agricole Consumer Finance

CACF opère dans de nombreux pays. En France, c'est principalement à travers la marque Sofinco que sont commercialisés les crédits à la consommation pour lesquels il existe une relation directe entre CACF et le client (dite B2C), par exemple lorsqu'un demandeur se rend directement sur le site internet sofinco.fr.

Par ailleurs, de nombreux crédits à la consommation sont distribués à travers un réseau de partenaires, qui jouent le rôle d'intermédiaires (on parle alors de B2B) : concessionnaires automobile, distributeurs d'électroménager, etc.

Enfin, CACF faisant partie du groupe Crédit Agricole, de nombreuses Caisses Régionales distribuent des crédits à la consommation à leur clientèle bancarisé, par l'intermédiaire des gestionnaires de compte.

Là encore, on constate que les spécificités des canaux de distribution des crédits impactent grandement la collecte des données et leur traitement statistique. En effet, les informations collectées sur le client, le produit et éventuellement l'apporteur d'affaires sont différentes selon le canal.

Dans la partie suivante, la méthodologie présentée est spécifique à CACF ; il pourra néan-

moins être admis que, dans les grandes lignes, cette méthodologie est similaire à la concurrence d'une part, et à la pratique d'autres pays (européens du moins) puisque la législation sur la protection et le traitement des données est sensiblement similaire (du fait de l'entrée en vigueur récente de la GDPR) et les établissements bancaires possèdent généralement des filiales dans plusieurs pays d'Europe et y font appliquer la même méthodologie.

1.2 Le *Credit Scoring* : état de l'art de la pratique industrielle

Cette partie vise à présenter la pratique actuelle en matière de *Credit Scoring* et pose un certain nombre de questions statistiques dont certaines ont été traitées dans cette thèse, d'autres trouvent des réponses (parfois partielles) dans la litérature et dont certaines références sont données à titre informatif mais ne sont pas développées dans ce manuscrit ; enfin, certaines questions ne trouvent *a priori* pas de réponse immédiate dans la litérature et sont autant de matière à de futurs travaux dans le domaine !

1.2.1 Collecte des données

La partie précédente a mis en exergue la pluralité des sources de données. La figure 1.1 présente par exemple le formulaire de souscription en vigueur pour un crédit automobile auprès de Sofinco *via* leur site web. Dans cet exemple, des données socio-démographiques et du véhicule à financer sont demandées. Pour un client, elles sont notées $\boldsymbol{x} = (\boldsymbol{X}_j)_1^D$ dans la suite. Ces informations sont de nature continue $x_j \in \mathbb{R}$ ou catégorielle $x_j \in \mathbb{N}_{o_j}$.

1.2.2 Préparation des données

1.2.3 Critère à modéliser

1.2.4 Données d'apprentissage

1.2.5 L'apprentissage d'une règle de décision

1.2.6 La métrique de performance

1.2.7 Suivi temporel de la performance du score

1.3 Apprentissage statistique

1.3.1 Mécanisme de génération des données

1.3.2 Apprentissage semi-supervisé

1.3.3 Apprentissage supervisé

Sélection de variables

Variable cible

Choix de modèle

Ce chapitre a permis .

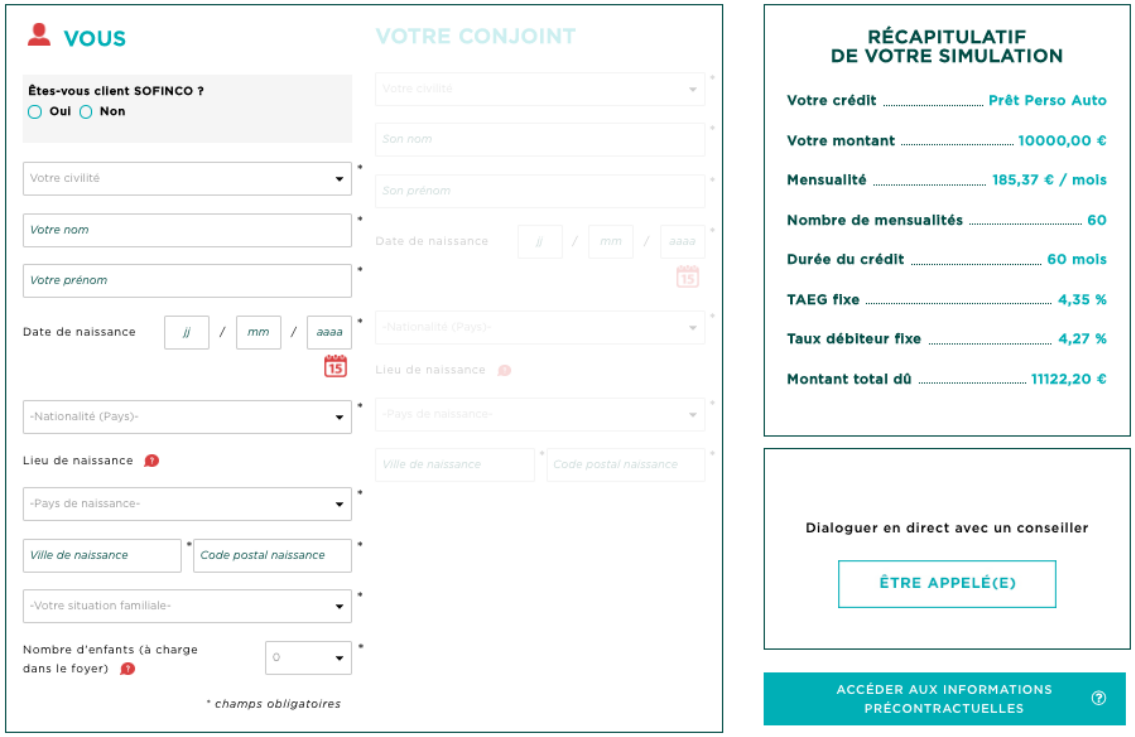

FIGURE 1.1 – Formulaire de souscription d'un crédit automobile Sofinco.

Références du chapitre 1

[1] Elsa Dicharry. *Maison de la literie lance la location avec option d'achat*. Oct. 2017. url : https://www.lesechos.fr/26/10/2017/lesechos.fr/030786701376_maison-de-la-literie-lance-la-location-avec-option-d-achat.htm.

[2] Hélène Ducourant. « Le crédit revolving, un succès populaire ». In : *Sociétés contemporaines* 4 (2009), p. 41–65.

[3] Dilruba Karim et al. « Off-balance sheet exposures and banking crises in OECD countries ». In : *Journal of Financial Stability* 9.4 (2013), p. 673–681.

[4] Jean-Philippe Peden. *VENTE DE VOITURES : LA PART DES FORMULES DE LOCATION A DECOLLE EN 2017*. Jan. 2018. url : https://news.autoplus.fr/Location-LLD-LOA-Vente-Marques-premium-1523494.html.

[5] Statista. *Credit cards per household by country in 2016*. 2016. url : https://www.statista.com/statistics/650858/credit-cards-per-household-by-country/.

Chapitre 2

Reject Inference

Nota Bene : ce chapitre s'inspire fortement de l'article [...]

2.1

Chapitre 3

Target feature in *Credit Scoring*

Chapitre 4

Supervised multivariate discretization and factor levels merging for logistic regression

Chapitre 5

Interaction discovery for logistic regression

Chapitre 6

Tree-structure segmentation for logistic regression

Chapitre 7

High dimensional data in *Credit Scoring*

Conclusion

Annexe A

Algorithms

A.1 Reject Inference methods

A.2 Discretization methods

A.3 Factor levels grouping methods

A.4 Interaction discovery methods

A.5 Logistic regression-based trees

Annexe B

Softwares

B.1 The R Statistical Software

B.1.1 The glmdisc package

B.2 The Python programming language

B.2.1 The glmdisc package

Table des matières

Modèles prédictifs pour données volumineuses et biaisées
Application à l'amélioration du scoring en risque de crédit

Résumé

Le ratio risque/récompense désigne en finance la logique selon laquelle un investissement peu risqué ne pourra être que faiblement rentable tandis qu'un investissement risqué a un rendement plus élevé mais est exposé à une perte. Les établissements financiers spécialisés en crédit à la consommation transposent ce principe en deux heuristiques : premièrement, le taux d'intérêt des crédits est adapté en fonction des clients et des produits ; deuxièmement, les clients demandeurs sont sélectionnés selon leur solvabilité. Ce mécanisme d'acceptation/rejet de la clientèle est composé de plusieurs règles de décision dont un score, c'est-à-dire une notation témoignant de la probabilité de défaut d'un nouveau client. La construction de ce score, qu'on désigne généralement par *Credit Scoring*, repose sur des techniques statistiques relativement anciennes et des heuristiques industrielles dont certaines ont été examinées dans cette thèse.
Après une première partie décrivant l'évolution et le contexte industriels actuels ainsi que la littérature académique associée, on s'intéressera dans une deuxième partie à une contribution importante de cette thèse : la "réintégration des refusés" ou comment tirer partie des informations collectées sur les clients refusés mais non utilisées. On verra ensuite en troisième partie l'apport de la méthode proposée dans cette thèse pour la discrétisation (resp. regroupement de modalités) des variables quantitatives (resp. qualitatives) constitutives du score ainsi que l'introduction d'interactions sur sa qualité. Enfin, la quatrième partie .
L'ensemble des travaux est illustré par des données réelles de Crédit Agricole Consumer Finance, établissement bancaire spécialiste du crédit à la consommation à l'origine de cette thèse CIFRE.

Mots clés : scoring, risque, crédit, prédiction, discrétisation, segmentation

---

Predictive models for big and biased data
Application to Credit Scoring

Abstract

The risk-reward is a well known finance paradigm: the higher the risk of an investment, the higher the expected reward . When it comes to consumer loans,

Keywords: scoring, credit, risk, predictin, discretization, clustering

---

Équipe-projet MΘDAL
Inria Lille Nord-Europe – 40 Avenue Halley – 59650 Villeneuve-d'Ascq