Institut des ACTUAIRES

ENSAE IP PARIS

Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 17/03/2022

Par : **Fallou NIAKH**

Titre : **Apport du Machine Learning dans l'analyse multivarié de la sinistralité des contrats emprunteur**

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de la filière

Christian ROBERT

Membres présents du jury de l'Institut des Actuaires

Arnaud COHEN (usso)

Anaelle LE BERRE (usso)

Entreprise : CNP Assurances

Nom : S. FRANC

Signature :

CNP Assurances Place Raoul Dautry 75015 PARIS

Directeur du mémoire en entreprise :

Nom : Cyrine GHARBI

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Secrétariat:

Signature du candidat

Bibliothèque:

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Liste des graphiques

Liste des tableaux

Résumé

Le contrat d'assurance emprunteur est un contrat temporaire, limité à la durée du prêt. Il permet de couvrir un emprunteur contre un risque d'insolvabilité à la suite de la survenance d'un sinistre tel que le Décès (DC), l'Incapacité de Travail (IT) ou la Perte d'Emploi (PE). Ainsi, une connaissance approfondie de ces risques est essentielle pour l'assureur afin que les primes collectées par ce dernier puissent couvrir les futurs sinistres inconnus. Dans ce mémoire, nous avons analysé et quantifié ces risques en maille fine en faisant appel aux méthodes de *Machine Learning*. Pour appréhender l'apport de ces nouvelles techniques, nous avons comparé leurs robustesse et performance avec celles des méthodes classiques à la lumière des deux plus grands risques en assurance emprunteur : le DC et l'IT.

Les données relatives aux emprunteurs ont été obtenues à partir du portefeuille de contrats collectifs d'un établissement de crédit. Nous avons commencé par segmenter notre portefeuille par scoring avec la régression logistique. Les variables binaires de DC et d'entrée en IT sont utilisées à cet effet. Ensuite, pour le risque DC, nous avons mis en place une loi d'expérience de maintien en vie dans le prêt. Nous avons utilisé trois approches de modélisation : *Cox*, *XGBoost Cox* et *XGBoost survival embedding (Xgbse)*. Pour ce qui est du risque IT, nous avons modélisé les prestations en adoptant trois approches : *Tweedie*, *XGBoost* et les *Réseaux de neurones*. En combinant les résultats issus de la modélisation sur le DC et l'IT, nous avons pu proposer des méthodes alternatives de provisionnement et de calcul de rentabilité dans le portefeuille.

Cette étude révèle que l'hypothèse du modèle de *Cox* selon laquelle les risques sont proportionnels n'est pas vérifiée par nos données. Ainsi, le modèle *Xgbse* est utilisé pour la construction de notre table de maintien. L'analyse des résultats de ce dernier laisse entrevoir que les variables les plus importantes pour l'explication du risque de mortalité sont : l'âge à la souscription, la durée du prêt et le sexe. Les résultats de la modélisation des prestations IT montrent que le modèle *XGBoost* affiche des meilleures performances. En outre, le capital assuré, la durée du prêt et la tranche d'âge [45 ; 55 ans) sont les facteurs les plus déterminants.

Mots clés : Assurance emprunteur, Cox, Xgbse, Tweedie, XGBoost, Réseaux de neurones

Abstract

Loan insurance contract is a temporary contract, limited to the duration of the loan. It covers a borrower against the risk of insolvency following the occurrence of a claim such as death, disability or loss of employment. Thus, a deep knowledge of these risks is essential for the insurer so that the premiums collected by the latter can cover future unknown claims. In this master thesis, we have analyzed and quantified these risks in fine mesh using *Machine Learning* methods. To understand the contribution of these new methods, we compared their robustness and performance with those of traditional methods using the two largest risks in loan insurance : death and work disability.

Borrower data were obtained from the group contracts portfolio of a credit institution. We started by segmenting our portfolio by scoring with logistic regression. The binary variables of death and disability entry are used for this purpose. Afterwards, for mortality risk, we established a survival law of experience in the loan. We used three modeling approaches : *Cox*, *XGBoost Cox* and *XGBoost survival embedding (Xgbse)*. For work disability risk, we modeled claims using three approaches : *Tweedie*, *XGBoost* and *Neural Networks*. By combining the results from death and disability modeling, we were able to propose alternative methods of reserving and calculating profitability in the portfolio.

This study shows that the assumption of the *Cox* model regarding proportional risks is not verified by our data. As a result, the model *Xgbse* is used to construct our survival law of experience in the loan. Analysis of the results of the latter suggests that the most important variables in explaining mortality risk are : age at underwriting, loan duration and gender. The results of the modeling of work disability claims show that the model *XGBoost* performs better. In addition, the sum insured, the loan duration and the age range [45 ; 55 years] are the most important contributors.

Mots clés : Loan insurance, Cox, Xgbse, Tweedie, XGBoost, Neural Networks

Remerciements

La justesse et la rigueur des méthodes utilisées, ainsi que des conditions de travail favorables et un bon relationnel au sein de la structure d'accueil sont des facteurs décisifs pour la réussite du stage de fin d'études sanctionné par ce présent mémoire.

C'est l'occasion pour moi de remercier tous ceux qui, de près ou de loin, ont contribué à sa réalisation.

Je tiens tout d'abord à remercier Cédric ATCHAMA de m'avoir donné l'opportunité de travailler au sein de l'équipe ainsi que pour ses conseils, sa générosité et son grand sens de partage. Je remercie également mon maître de stage Cyrine GHARBI qui n'a ménagé aucun effort pour mettre à ma disposition les outils nécessaires ainsi que pour son suivi et sa grande disponibilité.

Je ne saurais terminer sans remercier tout le personnel du département emprunteur de CNP Assurances pour l'accueil et leur sens de partage.

Enfin, je remercie tout le corps professoral de l'ENSAE Paris pour les sacrifices réalisées afin de nous offrir une formation de qualité. Je témoigne ici ma reconnaissance à mon encadreur M. Christian-Yann ROBERT, pour non seulement sa grande disponibilité mais également tout ce qu'il m'a apporté durant ma formation.

Table des matières

Introduction

Les contrats d'assurance emprunteur consistent à garantir le remboursement d'un prêt à l'organisme prêteur en cas de réalisation d'un ou de plusieurs risques durant la durée du prêt. Les risques couverts par ces contrats sont généralement le décès (obligatoire), l'arrêt de travail et la perte d'emploi. L'assureur rembourse l'intégralité du capital restant dû en cas de décès ou s'engage à verser une rente durant la durée d'arrêt de travail ou de perte d'emploi à l'organisme prêteur. Ainsi, une connaissance approfondie de ces risques est essentielle pour l'assureur afin que les primes collectées par ce dernier puissent couvrir les futurs sinistres inconnus. Dans le même sillage, une telle expertise permet aussi à l'assureur de proposer une prime précise et équitable aux emprunteurs.

La couverture de ces risques est au cœur de l'activité de CNP Assurances qui est d'ailleurs le premier assureur emprunteur en France. Nous avons mis à notre disposition, dans le cadre de notre stage, un portefeuille d'assurance emprunteur collectif d'un établissement de crédit dont les risques assurés sont le décès, l'arrêt de travail et la perte d'emploi. La tarification existante sur ce portefeuille collectif est du type tranche d'âge à la souscription $\times$ tranche de capital emprunté. Cependant, outre ces deux variables de tarification, nous avons aussi à notre disposition des variables comme la catégorie socioprofessionnelle, la durée totale du prêt, le capital assuré, la quotité d'assurance, etc.

De ce fait, il convient de se demander si ces variables non tarifaires contribuent à l'explication de la sinistralité du portefeuille ?

Depuis déjà plusieurs décennies, les actuaires ont développé des méthodes statistiques traditionnelles pour estimer ces risques. Récemment, face à la grande quantité des données en entreprise, l'apport considérable du *Machine Learning* dans la résolution des problèmes d'apprentissage supervisé est aujourd'hui un fait. De plus, les algorithmes de *Machine Learning* permettent de capturer la structure de l'information sans recourir à des hypothèses fortes sur les distributions des variables, contrairement aux méthodes statistiques traditionnelles. Par conséquent, les algorithmes de *Machine Learning* peuvent être plus efficaces pour capturer et modéliser des phénomènes complexes.

Dans l'optique d'optimiser son processus de suivi de risque en assurance emprunteur, CNP Assurances souhaite donc emboîter le pas des méthodes *Machine Learning*. Cependant, il ne

s'agit pas d'un parti pris, mais plus concrètement, il est ici question d'éprouver cette deuxième piste.

Cette étude naît ainsi de cette préoccupation et se focalise sur deux garanties en contrats emprunteur : le décès et l'arrêt de travail. Elle analyse et quantifie ces deux risques en maille fine. Nous avons comparé à cet effet la robustesse et la performance des méthodes classiques et *Machine Learning*.

Pour aborder cette problématique, le mémoire est organisé autour de cinq parties. La première partie est consacrée à la présentation de l'assurance emprunteur, des données et des statistiques descriptives. La deuxième partie, quant à elle, porte sur la segmentation de notre portefeuille en niveau de risque homogène. Ensuite, la troisième partie expose la modélisation du risque de décès avec la construction de la table d'expérience de maintien en vie dans le prêt. La quatrième partie est destinée à la modélisation coût / fréquence du risque d'arrêt de travail. Enfin, la cinquième partie présente les impacts des résultats obtenus sur les méthodes de provisionnement et de calcul de la rentabilité dans le portefeuille d'étude.

1) Cadre général de l'étude

DANS ce chapitre, nous présentons le marché de l'assurance emprunteur en mettant en exergue le vocabulaire utilisé ainsi que le principe de tarification. Ensuite, il sera question d'exposer les sources des données qui ont permis la réalisation de ce mémoire et la construction de la base de données de modélisation. Enfin, nous réalisons des statistiques descriptives afin d'avoir une vue globale de la répartition de notre portefeuille d'étude.

1.1 Le marché de l'assurance emprunteur

1.1.1 Définition de l'assurance emprunteur

Le contrat d'assurance emprunteur est un contrat temporaire, limité à la durée du prêt. Il permet de couvrir un emprunteur contre un risque d'insolvabilité à la suite de la survenance d'un sinistre tel que le décès, l'arrêt de travail ou la perte d'emploi. Cette couverture est importante dans la mesure où la survenance d'un de ces sinistres pourrait compromettre les capacités de l'emprunteur à rembourser ses futures échéances de prêt.

Cette assurance est généralement une condition nécessaire à l'obtention d'un prêt. Bien qu'il n'y ait pas d'obligation légale, les établissements de crédit conditionnent souvent l'accès au crédit à l'adhésion à la souscription d'un contrat d'assurance emprunteur. En outre, même si cette garantie offre une réelle sécurité à l'emprunteur et à sa famille, l'assurance emprunteur permet également et avant tout de protéger le prêteur.

1.1.2 Vocabulaire d'un contrat d'assurance emprunteur

Les parties d'un contrat d'assurance emprunteur sont usuellement les suivantes :

- **Assuré** : correspond à la personne qui sollicite le crédit à la banque, c'est elle qui est sous risque (emprunteur ou co-emprunteur) ;
- **Assureur** : correspond à la compagnie d'assurance ;
- **Bénéficiaire** : correspond à l'établissement de crédit ;
- **Souscripteur** : dans le cas d'un contrat individuel pur c'est l'assuré, si c'est un contrat de groupe c'est le bénéficiaire.

Il existe deux types de contrats d'assurance : les contrats de groupe et les contrats individuels.

— **Contrat d'assurance de groupe** : "contrat souscrit par une personne morale ou un chef d'entreprise en vue de l'adhésion d'un ensemble de personnes répondant à des conditions définies au contrat, pour la couverture des risques dépendant de la durée de la vie humaine, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail ou d'invalidité ou du risque de chômage. Les adhérents doivent avoir un lien de même nature avec le souscripteur [1]."

— **Contrat individuel** : la loi Lagarde de 2010 est venue faciliter la non-adhésion du client au contrat de groupe proposé par la banque. Il est possible dans ce cadre de souscrire une assurance individuelle, dans laquelle tous les critères propres à l'individu sont pris en compte dans le tarif (âge, sexe, état de santé, profession, etc.). De fait, ce type de contrat est plus intéressant pour les assurés jeunes et en bonne santé, qui bénéficieront d'un tarif adapté à leur risque.

Sur le marché de l'assurance emprunteur trois catégories de contrats existent :

— **Contrats de groupe** : il s'agit de contrats collectifs, très peu segmentés, qui sont systématiquement proposés par l'établissement prêteur lors de la souscription du prêt. La tarification est établie en général en fonction du capital emprunté et la prime est maintenue constante dans le temps. Le processus de segmentation tient généralement compte de la tranche d'âge de l'assuré au moment de la souscription.

— **Contrats défensifs dits "individuels"** : comme les contrats de groupe, ils sont proposés par l'établissement prêteur, mais sous forme d'une "offre défensive". Leur tarification se fait généralement sur la base du capital restant dû. La tarification tient compte des caractéristiques de l'assuré, notamment de son âge au moment de la souscription, de son état de santé, et parfois de sa catégorie professionnelle, et peut également dépendre du fait que celui-ci soit fumeur ou non. Ce sont donc les emprunteurs présentant un profil à faible risque qui constituent les cibles principales de ces produits.

— **Contrats individuels purs** : ce sont des contrats conclus directement entre l'assuré et l'assureur. Très segmentés, ils sont donc avantageux pour les bons risques. Les primes peuvent prendre la forme d'une prime unique, d'une prime constante ou d'une prime basée sur le capital restant dû. Sur le marché français, le plus souvent, la tarification de ces contrats se fait en fonction du capital restant dû et de l'âge atteint.

Lors des différentes phases de souscription et de tarification, d'autres mots de vocabulaire

1. Code des assurances - Article L141-1

apparaissent. Ces nouveaux éléments sont les suivants [2] :

— **Délai de carence** : il s'agit de la période durant laquelle les garanties souscrites par l'assuré (décès, arrêt travail, perte d'emploi...) ne sont pas encore effectives à la suite de la survenance d'un sinistre mentionné dans son contrat d'assurance de prêt.
C'est donc la période durant laquelle les garanties ne s'appliqueront pas même si l'assuré est, durant ce laps de temps, confronté à une difficulté concernant le remboursement du prêt, l'assurance ne jouera alors pas le rôle de relais.

— **Franchise** : c'est la période durant laquelle les prestations ne sont pas dues par l'assureur.

— **Quotité d'assurance** : elle définit la part du capital empruntée garantie par la compagnie d'assurance. Autrement dit, elle décrit la répartition de la couverture qui sera fournie par l'assureur en cas de sinistres.

— **Sélection médicale** : dans le cadre des contrats d'assurance emprunteur proposés par les assureurs en France, un questionnaire de santé peut être rempli afin d'aider l'assureur à évaluer les risques qu'il prend. Lequel est visé par l'article L113-2 du Code des assurances. À partir de ces informations, l'assureur se réserve le droit de proposer un tarif avec une surprime et d'ajouter des exclusions au contrat.

1.1.3 Les garanties d'un contrat d'assurance emprunteur

Les garanties sont en général acquises une fois le prêt validé par l'établissement de crédit, dès lors que l'affiliation a été acceptée par l'assureur. Une des conditions d'adhésion à un contrat d'assurance emprunteur est d'avoir un âge à la souscription compris entre 18 et 75 ans.

Le produit d'assurance emprunteur permet de protéger l'emprunteur contre la survenance des évènements suivants :

— **Décès** : c'est la seule garantie qui est obligatoire. Elle entraîne le remboursement intégral du capital restant dû par l'assureur en cas de décès de l'emprunteur.

— **Incapacité de travail** : ce risque est réalisé lorsque l'assuré se trouve, à la suite d'un accident ou d'une maladie, dans l'incapacité reconnue médicalement d'exercer son activité professionnelle, même à temps partiel, et que cette incapacité persiste au-delà d'une période de franchise.

— **Perte totale et irréversible d'autonomie (PTIA)** : un assuré est en état de PTIA lorsque l'invalidité dont il est atteint le place dans l'impossibilité définitive de se livrer

2. Glossaire assurance emprunteur

à toute occupation et / ou toute activité rémunérée ou lui procurant un gain ou profit, et le met définitivement dans l'obligation de recourir à l'assistance totale et constante d'une tierce personne pour l'ensemble des actes ordinaires de la vie (se laver, s'habiller, se nourrir, se déplacer).

— **Perte d'emploi** : elle couvre le risque de licenciement et ne concerne que les emprunteurs salariés en contrat à durée indéterminée ayant renseigné une déclaration de Plein Emploi ; les salariés en CDD, les artisans, les professions libérales et les commerçants n'étant pas pris en charge. L'assureur peut soit prendre en charge totalement ou partiellement les échéances de crédit (montant plafonné) ou soit permettre leurs reports sur la période prévue de la garantie (limitée à 18 mois).

1.1.4 Principe de tarification en assurance emprunteur collectif

Les contrats collectifs reposent sur la *mutualisation des risques*. Le tarif est identique par groupe d'emprunteurs. Par conséquent, il peut être le même pour un ensemble d'emprunteurs quels que soit leur âge, et leur catégorie socioprofessionnelle.

Calcul de la prime pure

Aux termes d'un contrat emprunteur, l'assureur se charge de rembourser le capital restant dû ou les échéances du prêt en cas de réalisation d'un risque garanti. En retour, l'emprunteur accepte de payer une prime. Définie dès le départ et pour toute la durée du prêt, cette prime correspond au prix de la protection contre un ou plusieurs risques définis dans le contrat.

Le principe de tarification est de définir le montant de la prime commerciale demandée à l'emprunteur. La prime appliquée par l'assureur à l'emprunteur, appelée "prime commerciale", intègre le prix de la couverture du risque ainsi que les chargements.

La prime commerciale peut être ventilée comme suit :

Prime Commerciale = Prime Pure + Chargements assureur + Commissions sur prime + Taxe

— La prime pure évalue le coût du risque seul ;

— Les chargements couvrent les frais engagés par l'assureur ;

— Les commissions rémunèrent le partenaire bancaire (il s'agit ici de commissions "garanties", assises sur les primes hors taxe et versées au partenaire quel que soit l'état technique du compte) ;

— La taxe incluse dans la tarification (Taxe sur les Conventions d'Assurance) concerne la part de prime relative au risque arrêt de travail (et perte d'emploi).

Le montant de la prime pure est défini comme étant la multiplication entre le taux de prime pure et la somme assurée. On détermine le taux de prime en égalisant les valeurs actuelles probables des engagements respectifs de l'assureur et de l'emprunteur en $t = 0$:

$$VAP\left(\text{Engagements\_Assureur}\right)_{t=0} = VAP\left(\text{Engagements\_Assuré}\right)_{t=0}$$

Pour rappel, le contrat emprunteur est un contrat temporaire, limité à la durée du prêt. Il garantit le remboursement :

— du Capital Restant Dû (CRD) en cas de décès ;

— des mensualités du crédit en cas d'arrêt de travail.

Pour calculer l'engagement de l'assureur, il faut d'abord construire le tableau d'amortissement pour obtenir le CRD et la mensualité pour chaque période.

Tableau d'amortissement

Dans le cas de remboursement par mensualité constante où les versements périodiques (amortissement + intérêt) sont constants pendant toute la durée de vie de l'emprunt nous avons :

1. Détermination de la mensualité du prêt

   La séquence des n paiements mensuels (m) envers le prêteur doit avoir une valeur actuelle égale au capital initial.

   $$CI = m \times \frac{1-(1+t)^{-n}}{t}$$

   où t désigne le taux mensuel du prêt.

   Ainsi, on en déduit le montant de la mensualité :

   $$m = \frac{CI \times t}{1-(1+t)^{-n}}$$

2. Construction du tableau d'amortissement

   Tout paiement mensuel constant correspond à la somme d'une partie intérêt (I) et d'une partie amortissement (A) qui correspond au remboursement d'une fraction du capital initial. L'intérêt est calculé sur le capital restant dû en début de période. L'amortissement peut être calculé en soustrayant les intérêts de la mensualité. L'ensemble du tableau d'amortissement peut être rempli ligne par ligne :

Mois	Capital dû en début de période	Amortissement	Intérêt	Mensualité
1	CI	$A_1 = m - I_1$	$I_1 = t \times CI$	m
2	$CRD_1 = CI - A_1$	$A_2 = m - I_2$	$I_2 = t \times CRD_1$	m
:	:	:	:	m
p	$CRD_{p-1} = CRD_{p-2} - A_{p-1}$	$A_P = m - I_p$	$I_p = t \times CRD_{p-1}$	m
$p+1$	$CRD_p = CRD_{p-1} - A_p$	$A_{P+1} = m - I_{p+1}$	$I_{p+1} = t \times CRD_p$	m
:	:	:	:	m
n	$CRD_{n-1} = CRD_{n-2} - A_{n-1}$	$A_n = CRD_{n-1}$	$t \times CRD_{n-1}$	m

TABLEAU 1.1 – Tableau d'amortissement.

On obtient la formule suivante de calcul du capital restant dû à l'échéance p :

$$CRD_p = CI \times \frac{(1+i)^n - (1+i)^p}{(1+i)^n - 1}$$

Engagement de l'assureur en cas de décès

Nous partons du principe que les décès surviennent en moyenne au milieu de mois, de même que l'assureur se chargerait de rembourser la totalité du capital restant dû après le dernier versement à l'établissement financier en cas de décès.

En ce qui concerne les tables de mortalité TH00-02 et TF00-02, elles fournissent le nombre de survivants ainsi que les taux de mortalité annuels pour les individus d'âge entier x, nous avons cherché à exprimer ces données en mensuel.

On note $x(0)$ l'âge de l'assuré au début de la projection ; ainsi, la valeur $x(t)$ - exprimée en années - se traduit par l'âge de l'assuré au bout de t mois de projection : $x(t) = x(0) + \frac{t}{12}$.
On définit par conséquent l'âge atteint en années entières au bout de t mois projetés, noté $y(t)$, comme suit :

$$y(t) = Ent[x(t)] = x(0) + Ent[t/12]$$

où la fonction *Ent* représente la partie entière.
Il est ainsi possible de déterminer les probabilités de décès sur une base mensuelle à l'aide de la formule suivante :

$$q\_mensuel_{x(t)} = 1 - (1 - q\_annuel_{x(0) + Ent[t/12]})^{\frac{1}{12}}$$

où :

q_mensuel : représente le taux de mortalité mensuel

q_annuel : représente le taux de mortalité annuel

La formule ainsi établie fait apparaître que la probabilité annuelle de survie jusqu'à l'âge $y(t)$ résulte du produit des probabilités mensuelles.
On trouve ci-dessous la valeur actuelle probable des engagements de l'assureur :

$$VAP\_\text{Assureur} = \sum_{k=1}^{\text{Dass}} CRD_k \times \frac{l_{x(k)}}{l_{x(0)}} \times \frac{l_{x(k)} - l_{x(k+1)}}{l_{x(k)}} \times (1 + i_{\text{mens}})^{-(k+0,5)}$$

Avec :

- CRD_k désigne le capital restant dû du $k^{\text{ème}}$ mois c'est-à-dire le montant de la prestation à la charge de l'assureur dans le cadre de cette garantie
- $\frac{l_{x(k)}}{l_{x(0)}}$ correspond à la probabilité de survie de l'assuré, d'âge x à la souscription, au $k^{\text{ème}}$ mois. Les $l_{x(k)}$ sont déduits des taux de décès mensuels présentés ci-dessus.
- Dass représente la durée de la garantie exprimée en mois ;
- i_{mens} correspond au taux d'actualisation mensuel définit par : $i_{\text{mens}} = (1+i)^{\frac{1}{12}} - 1$ avec i le taux d'actualisation annuel.

Engagement de l'assureur en cas d'arrêt de travail

Dans la garantie arrêt de travail, il y a une période de franchise, qui est généralement de 90 jours. Une condition d'âge limite étant fixée contractuellement, la garantie arrêt de travail prend fin au-delà de cet âge bien que la durée du prêt n'ait pas été entièrement consommée.

Nous admettons que les arrêts de travail surviennent en moyenne au milieu du mois.

Dans le cas d'un arrêt de travail, l'assureur se charge des mensualités de remboursement du prêt pour la totalité de la période assurée. Comme pour les tables de mortalité, les taux d'entrée en arrêt de travail se définissent par âges entiers. Nous procéderons donc de même à des interpolations en vue de déterminer les taux d'entrée en arrêt de travail mensuel. Par conséquent, le calcul appliqué est le suivant :

$$\text{Entree\_arrêt}_{x(t)} = 1 - \left(1 - \text{ Entree\_arrêt\_annuel }_{y(t)}\right)^{\frac{1}{12}}$$

- $\text{Entree\_arrêt}_{x(t)}$ représente la probabilité d'entrée en incapacité au $t^{ème}$ mois pour un individu d'âge $x(0)$ en début de projection ;
- Entree_arrêt_annuel $_{y(t)}$ représente la probabilité d'entrée en incapacité au $t^{ème}$ mois pour un individu d'âge entier $y(t)$ à cette date.

La valeur actuelle probable à laquelle s'attend l'assureur au moment de la souscription de la garantie arrêt de travail est alors déterminée comme suit :

$$VAP\_\text{Assureur} = \sum_{k=1}^{Dass} \frac{l_{x(k)}}{l_{x(0)}} \times (1+i_{\text{mens}})^{-(k+0,5)} \times \text{ Entree\_arrêt\_mensuel}_{x(k)} \times \text{ prestation}_k$$

— Entree_arrêt_mensuel$_{x(k)}$ représente la probabilité d'entrée en incapacité au $k^{\text{ème}}$ mois pour un individu d'âge $x(0)$;

— Le montant de la prestation correspond aux versements des remboursements à l'établissement de crédit tant que l'individu est en état d'arrêt de travail ; cette durée étant limitée au maximum entre la durée de prêt et la durée de couverture d'assurance :

$$\text{prestation}_k = \sum_{j=1}^{Dass-k} REMB_{j+k} \times (1+i_{mens})^{-(j+0,5)} \times \frac{l_{x(k)}^{maintien\_incapacite_{j+k}}}{l_{x(k)}^{maintien\_incapacite_k}}$$

Avec :

— $l_{x(k)}^{maintien\_incapacite_j}$ le nombre d'individus entrés en incapacité à l'âge $x+\frac{k}{12}$ et ayant j mois d'ancienneté (fournis par la table de maintien en arrêt de travail)

— $REMB_{k+j}$: le $(j+k)^{\text{ème}}$ remboursement du prêt.

Au moment de la tarification du risque de décès et/ou d'arrêt de travail, l'assureur pourrait aussi prendre en compte un taux estimé de remboursement anticipé du prêt.

Engagement de l'assuré

Nous faisons l'hypothèse que le taux de prime reste inchangé tout au long de la durée du prêt et que la tarification est basée sur le capital initial.

La valeur actuelle probable des engagements de l'assuré à la souscription est définie par :

$$VAP(Engagements\_Assuré)_{t=0} = T_a \times CI \times \sum_{j=1}^{\text{Dass}} {}_jP_x \times \left(1+i_f\right)^{-j}$$

Avec :

— T_a : taux d'assurance (il traduit la prime pure suivant la garantie, décès ou arrêt de travail, laquelle sera établie selon les engagements de l'assureur sur chacune des garanties : voir formule finale ci-dessous.)

— CI : le Capital Initial

— $Dass$: le nombre de versements de primes (durée du prêt en mois)

— ${}_jP_x$: probabilité qu'un individu d'âge x survive à l'âge $(x+j)$

— x : âge actuariel moyen des adhérents au contrat collectif

— i_f : le taux d'intérêt technique (dépendant du fractionnement)

À partir du taux d'intérêt annuel, on peut obtenir le taux d'intérêt périodique adjoint au moyen de la formule suivante :

$$\text{Taux\_périodique} = \Big(1 + \text{Taux\_annuel}\Big)^{\frac{1}{\text{Nombre de période}}} - 1$$

À titre d'exemple : dans le cas d'une échelle mensuelle, le nombre de périodes est de 12.

En égalisant les valeurs actuelles probables de l'assuré et de l'assureur, le taux d'assurance est déterminé par la formule suivante :

$$T_a = \frac{VAP\Big(\text{Engagements\_Assureur}\Big)_{t=0}}{CI \times \sum_{j=1}^{\text{Dass}} {}_jP_x \times \Big(1+i_f\Big)^{-j}}$$

1.2 Présentation des données

Dans cette section, nous allons présenter les sources des données qui ont permis la réalisation de ce mémoire ainsi que la construction de notre base de données de modélisation.

1.2.1 Objectifs de modélisation

Pour rappel, nous cherchons à effectuer une analyse multivariée de la sinistralité des contrats emprunteurs. Les contrats concernés sont ceux des assurés de contrats Emprunteur d'un établissement de crédit.

Le portefeuille est ainsi constitué de contrats liés à des prêts pour financer des achats immobiliers ou des prêts de consommation couvrant les risques suivants : Décès (DC), Incapacité de travail (IT) et Chômage (CH).

Dans les parties ci-dessous, nous présentons le processus de construction et de qualification des données afin de valider leur utilisation pour :

— construire une table de maintien en vie dans le prêt ;

— modéliser les montants des prestations pour la garantie IT ;

— proposer une méthode alternative de provisionnement pour les deux garanties DC et IT ;

— et apprécier la rentabilité du portefeuille.

1.2.2 Base de données des assurés

Nous disposons d'une base de données de 84 millions de lignes et 95 variables sous SAS. Toutes les opérations de prétraitement ont été réalisées avec ce logiciel. Cette base regroupe entre autres les variables caractérisant l'assuré et le prêt.

Après identification des variables pertinentes et informatives pour notre étude, nous décidons de sélectionner les caractéristiques ci-dessous :

TABLEAU 1.2 – Les caractéristiques du prêt

Variable	Description
MIP_ID_PRET	L'identifiant du prêt
MIP_NAT_PRET	Immobilier, Consommation
MIP_DT_SOUSCRIPTION	Date de souscription du prêt
MIP_DUR_TOTALE_PRET	Durée initiale du prêt
MIP_TX_EMPRUNT	Taux d'intérêt du prêt
MIP_CAP_INIT	Capital initial
MIP_QUOTITE_GLOBALE	Quotité assuré pour le prêt par emprunteur
MIP_RISQ_DC	Indicatrice spécifiant la présence de la garantie DC pour le prêt.
MIP_RISQ_IT	Indicatrice spécifiant la présence de la garantie IT pour le prêt.
MIP_RISQ_CH	Indicatrice spécifiant la présence de la garantie CH pour le prêt.
MIP_DT_FIN_PRET	La date de fin du prêt
MIP_DT_DERN_RA	La date du dernier remboursement anticipé
MIP_TARIF_APP_GLOBAL	Le tarif appliqué pour la couverture du prêt

TABLEAU 1.3 – Les caractéristiques de l'assuré

Variable	Description
MIP_ID_ASSURE	L'identifiant de l'assuré
MIP_DT_NAISS	Date de naissance de l'assuré
MIP_CD_SEXE	Sexe de l'assuré
MIP_CSP	Catégorie socio-professionnelle de l'assuré

A l'aide de ces variables, nous pouvons définir de nouvelles variables qui nous seront utiles lors de l'analyse statistique et la fiabilisation de notre base :

— MIP_PLUSIEURS_PRETS : correspond à une variable indicatrice qui permet d'identifier si un assureur détient plusieurs prêts. Elle possède deux modalités et vaut "OUI" si NOMBRE_PRETS>1 et "NON" sinon ;

— MIP_PLUSIEURS_ASSURES : correspond à une variable indicatrice qui permet d'identifier si plusieurs assurés sont présents pour le même prêt ;

— MIP_TYPE_GARANTIE : elle permet de récupérer simultanément l'information fournie par les variables indicatrices de garantie couverte pour le crédit. Elle a deux modalités : "DC" et "DC+IT" ;

— MIP_CAPITAL_ASSURE=$\frac{MIP\_CAP\_INIT * MIP\_QUOTITE\_GLOBALE}{100}$.

1.2.3 Base de données des sinistres

En ce qui concerne la base des sinistres, nous disposons de différentes tables dans ORACLE, issues d'un univers de données de l'entreprise et visualisables dans SAS. L'extraction de données que nous cherchons à effectuer ici, a pour but de transformer et d'intégrer les données disponibles dans le modèle relationnel de cet univers, au sein d'une table à plat. Cette dernière va nous permettre par la suite, d'une part de stocker toutes les informations dans une seule et unique table mais également d'effectuer diverses études et requêtes directement à partir de celle-ci.
L'univers de données se compose de différentes tables de sinistres tête par tête, comportant chacune des informations différentes. De plus, notons que les tables gèrent à la fois des sinistres emprunteurs et prévoyance. Les informations que nous allons trouver dans les tables ne concerneront donc pas exclusivement l'emprunteur. Nous avons ainsi utilisé les numéros de contrats de l'établissement de crédit afin de retenir uniquement ceux liés à l'emprunteur.
Les différentes tables sinistres sont :

— ODD_INFO_SINIS_PREV : informations relatives aux sinistres ;

— ODD_FLUX_CID : informations relatives aux flux ;

— ODD_PREST_TPT_SERVIES : informations concernant les risques et les règlements ;

— ODD_PREST_FSS_AGREG : informations concernant les montants versés ;

— ODD_INFO_PRET_SINIS_PREV : informations sur les prêts ;

— ODD_INFO_PREST_PERIOD : informations sur les périodes d'indemnisation des sinistres ;

— ODD_PERS_SINIS : informations sur les assurés ;

— ODD_BENEF_ASSU : informations sur le bénéficiaire de l'assurance ;

— ODD_BENEF_PREST : informations sur le bénéficiaire des prestations.

Ces tables nous permettent d'avoir les informations concernant les assurés, leur prêt, leur sinistre mais également les flux et règlements qui en découlent.
En tout, nous disposons de cinq variables communes à toutes les tables. Nous pouvons les classer en deux groupes :

— les variables identiques, à savoir, « NUM_VERSION » et « NUM_INTEGRATION » ;

— et les variables dont l'acronyme change : « NUM REF SINIS », « ID_FLUX » et «NUM_CCOLTE » correspondant à la concaténation du numéro de contrat et du numéro de collectivité.

On aura par exemple pour la table ODD_INFO_SINIS_PREV, la variable « ISP_NUM_CCOLTE » qui servira de clé. Concernant les flux, nous faisons face au même cas de figure puisque la variable « ID_FLUX » nous servira directement de clé reliant les différents flux entre eux. De même pour identifier et lier les sinistres entre eux, la variable « NUM_REF_SINIS », qui fonctionne comme les deux variables précédentes du même « groupe » nous servira elle aussi de clé.
Grâce à ces variables communes, il sera désormais possible de lier les tables entre elles dans le but de les fusionner pour en créer une seule et unique.

1.2.3.1 Modélisation des données

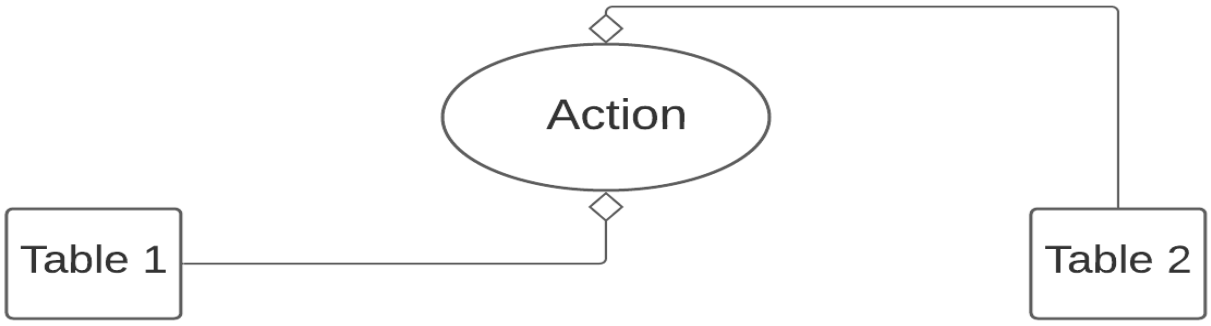
Tables utilisées

Dans le but de concevoir le Modèle Conceptuel des Données (MCD), nous allons utiliser les neuf tables citées ci-dessus. Ces dernières devront rassembler toutes les caractéristiques propres à l'assuré (date de naissance, capital emprunté . . .), au prêt (capital, quotité . . .), mais aussi aux sinistres (règlements, durée . . .). Pour cela, nous partons du modèle relationnel déjà existant dans l'univers des données. En effet, les neuf (9) tables ci-dessus, regroupent toutes ces caractéristiques. Nous devons également noter les cardinalités existantes entre les tables. Celles-ci vont nous permettre de relier ces tables et les informations qu'elles contiennent. La cardinalité sera utilisée pour compter le nombre minimum et maximum de possibilités présentes entre chaque classe, reliées à un ou plusieurs objets.
Notons, que parmi nos neuf tables, seuls cinq seront utilisées. En effet, lorsque nous affectons les numéros de contrats issus du périmètre d'étude aux quatre tables restantes nous constatons qu'elles ne contiennent aucune observation.

Clés et jointures entre les tables

Afin de lier les tables entres elles, nous devons trouver des clés de fusion. Ces dernières doivent être présentes à la fois dans la table à « gauche » de l'action et dans celle à « droite » comme ci-dessous :

GRAPHIQUE 1.1 – Exemple de jointure

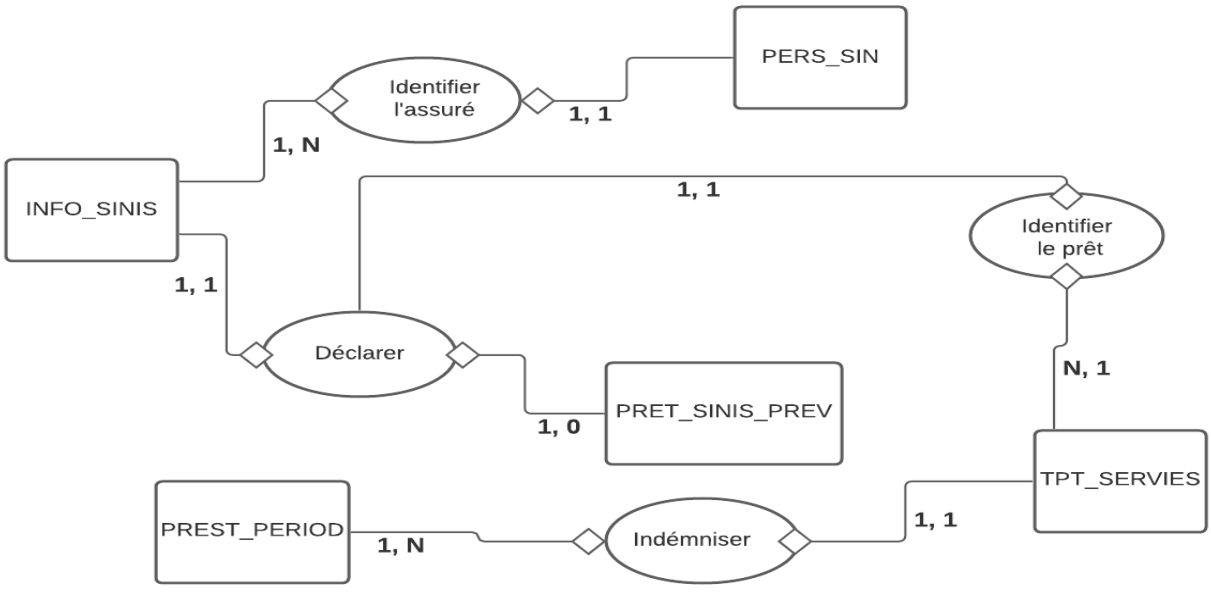


Si une variable commune existe entre ces tables, elle servira de clé de fusion, sinon selon les informations disponibles, une variable commune sera créée afin de réaliser cette fusion. Pour cette étude, des clés sont déjà disponibles et permettent donc de lier les tables modulo un travail préalable sur les données.

Schéma MCD et interprétation

Ainsi, le schéma MCD que nous obtenons est donc le suivant :

GRAPHIQUE 1.2 – Schéma MCD



Ce modèle nous permet de visualiser les liens entre les tables et les différents objets (actions). Ainsi dans la table INFO_SINIS, l'assuré déclare son sinistre. On a donc une relation entre un seul assuré qui peut déclarer un ou plusieurs sinistres, d'où la cardinalité (1, N).
Ensuite entre cette action et la table PRET_SINIS_PREV, un seul sinistre déclaré est relié à un seul identifiant de sinistre, donc la cardinalité est ici (1, 1). Concernant l'identification du

prêt, pour chaque numéro de sinistre il y a un seul et unique prêt d'où la cardinalité (N, 1) avec la possibilité d'avoir plusieurs sinistres sur un même prêt et la (1, N) car un prêt peut être relié à plusieurs sinistres.
En ce qui concerne l'identification de l'assuré, pour chaque prêt nous avons un seul et unique assuré, et inversement, la cardinalité sera donc (1, 1). De l'autre côté de cet objet, un assuré sera identifié par prêt et un prêt par assuré : (1, 1). Les autres cardinalités en (1, 1) suivent la même logique que précédemment.
Pour finir, avec l'indemnisation, un même assuré peut bénéficier de plusieurs règlements, en effet, il peut avoir différents sinistres, ou alors un sinistre sur plusieurs mois, d'où (1, N).

Travail préalable sur SAS

En analysant les variables qui composent les différentes tables, on peut voir qu'aucune d'entre elles ne peut servir de clé de fusion telle quelle. Ainsi, afin de lier au mieux les cinq tables dont nous disposons, nous allons agir comme suit :

- $1^{ère}$ étape : renommer les variables communes « ..._id_flux », « ..._num_ccolte » et « ..._num_ref_sinis » de façon à avoir un seul et unique nom de variable commun à toutes les tables
- $2^{ème}$ étape : trier les tables de la même façon et dans l'ordre de la première étape (« id_flux », « num_ccolte » et « num_ref_sinis »), ce qui nous permettra par la suite de les fusionner
- $3^{ème}$: fusionner les tables avec ces trois variables, renommées et triées

Sur SAS, on distinguera bien ces trois étapes, de telle sorte qu'à la fin on puisse avoir une seule et unique table regroupant toutes les caractéristiques des cinq tables sur lesquelles nous travaillons.

Clés de fusion

Pour lier les tables INFO_SINIS et PRET_SINIS_PREV du schéma MCD vu auparavant, nous aurons besoin d'une clé primaire et de deux clés secondaires. En effet, pour chaque numéro de contrat et de collectivité (NUM_CCOLTE), on dispose d'un numéro de référence de sinistre. Un même numéro peut donc correspondre à deux collectivités et contrat différents. Il faudra donc fusionner les tables selon ces deux critères. La même remarque sur le numéro de référence de sinistre est valable pour l'identifiant du flux. La clé primaire à toutes les variables sera donc la variable « NUM_REF_SINIS » et les deux clés secondaires seront : « ID_FLUX » et « NUM_CCOLTE ». Notons que pour relier les tables TPT_SERVIES et PREST_PERIOD,

une autre variable nous servira de clé qui sera « NUM_REF_PREST ». Elle permet d'identifier le nombre de règlements par sinistre, et ainsi, renforce la précision lorsqu'on va fusionner les tables entre elles. Il faudra donc, comme pour les variables communes à toutes les tables, la renommer dans TPT_SERVIE et PREST_PERIOD.

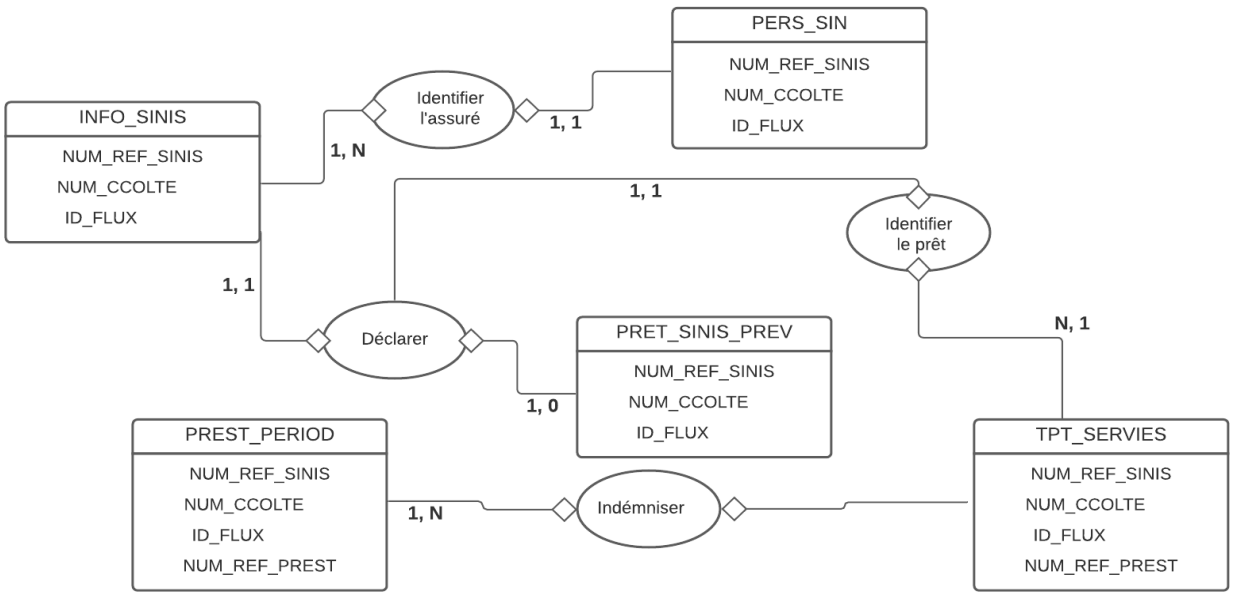
Table finale

L'objectif final est d'obtenir une table à plat regroupant toutes les informations issues des tables du MCD. Pour ce faire, on passe par un MLD avec les clés que nous avons énoncées précédemment.

Pour construire la table, il nous faut prendre en compte le fait qu'un assuré puisse avoir plusieurs sinistres et également le fait qu'un sinistre puisse toucher deux prêts différents (contractés) par un même assuré. Dans le cas où ce dernier compte deux prêts, on doublera la ligne mais exclusivement dans ce cas de figure. En effet, il y aura deux prêts différents à indemniser.

La table finale suivra le schéma du MLD suivant, avec les bonnes clés primaires et secondaires. Ainsi, on obtiendra une table avec toutes les données des sinistres, des assurés et les informations relatives aux prêts en assurance emprunteur.

GRAPHIQUE 1.3 – Schéma MLD



1.2.3.2 Les variables constituant la base

En réalisant les jointures entre les différentes tables représentées ci-dessus, on obtient la table finale qui compte 1 829 767 observations et 126 variables. Par ailleurs, en regroupant tous les sinistres appartenant au même prêt, on en obtient par la suite 17 508. Ce regroupement par

sinistre nous permet de calculer le nombre de prestations par prêt mais également leur coût total. Par ailleurs, nous obtenons également une seule et unique date de début de sinistre et une de fin de sinistre.

Dans un second temps, nous supprimons les variables qui sont constamment vide et qui ne sont donc pas exploitable. On passe alors de 126 à 64 variables grâce à un retraitement sous SAS. Après étude de leurs pertinences, nous avons choisi de garder les caractéristiques des sinistres suivantes :

TABLEAU 1.4 – Les caractéristiques du sinistre

Variable	Définition
NUM_VERSION	Numéro de version
NUM_INTEGRATION	Numéro intégration
NUM_REF_SINISTRE	Numéro de référence du sinistre
NUM_CCOLTE	Numéro de collectivité
ID_FLUX	Identifiant du flux
ISP_ID_SINIS	Identifiant du sinistre
ISP_DT_SURVENANCE_SINIS	Date de survenance du sinistre
PTS_CD_RISQ_PRIM_REF	Code du risque couvert
PTS_DT_RSGL_PREST	Date de règlement de la prestation
PTS_MNT_RSGL_PREST	Montant du règlement de la prestation
PTS_NUM_REF_PREST	Numéro de référence de la prestation
PTS_TOP_PERIO_INDEMNISEE	Période indemnisée
IPP_BASE_CALC_INDEMN	Base de calcul de l'indemnité
IPP_CD_BASE_INDEMN	Code base de l'indemnité
IPP_DT_DEB_PERIO	Date de début de période
IPP_DT_FIN_PERIO	Date de fin de période
IPP_MNT_TOT_RSGL_PERIO	Montant total règlement période

Au moment de l'enregistrement d'un sinistre, certaines caractéristiques du prêt et de l'assuré sont reprises dans la base de données des sinistres. Il s'agit du capital initial, de la quotité, les dates de souscription et fin du prêt, le sexe et la date de naissance. Nous avons ainsi testé la fiabilité de ces informations en les comparant avec celles enregistrées dans la base des assurés. Les résultats du test de fiabilisation sont renseignés dans le tableau 1.5.

Nbre d'enreg. base sinistrés	22 000
Nbre d'enreg. base sinistrés sans correspondance avec la base assuré	5 300
Nbre d'enreg. avec un CI <>	3 147
Nbre d'enreg. avec une quotité <>	4 000
Nbre d'enreg. avec une DT_SCRP <>	3 147
Nbre d'enreg. avec une DT_FIN <>	3 907
Nbre d'enreg. avec un code sexe <>	300

TABLEAU 1.5 – Fiabilisation de la base de données des sinistres.

1.2.4 Fusion des deux bases : assurés et sinistres

Le but de cette partie est de fusionner les données des sinistres avec les données des assurés. Cette fusion nous permet d'affecter à chaque couple (prêt, assuré) de la base des assurés, les informations sur leurs sinistres. Afin de réaliser cet exercice, nous avons choisi comme clés, le numéro de prêt et la date de naissance. Nous avons constaté lors de cette fusion que 5 300 enregistrements étaient introuvables. 4 200 d'entre eux correspondent à des prêts clôturés avant 2007 (date de début d'observation des assurés) ce qui justifie leur absence. Les 1 100 restants correspondent à des erreurs de saisie soit sur l'identifiant de prêt ou sur la date de naissance dans la base de données des sinistres.

1.3 Tests de fiabilisation de la base

Après la sélection des variables pertinentes pour notre étude, nous avons effectué une série de tests de cohérence et de retraitements afin de rendre notre base utilisable pour les travaux actuariels.

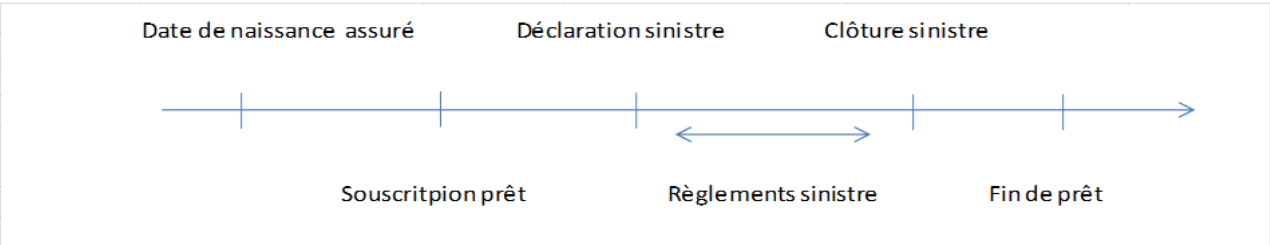
1.3.1 Unicité de l'emprunteur

Comme la base de données empile des observations sur plusieurs dates d'arrêtés, nous avons commencé par supprimer les duplications en utilisant le couple prêt/assuré. La base de données contient dorénavant 2,8 millions de lignes. Ensuite, nous avons testé l'existence de valeurs manquantes sur l'ID de l'assuré et celui du prêt. Il en ressort que 64 260 lignes ont des IDs assurés vides, nous les avons donc supprimés dans la base. On observe la présence à tort de 663 couples prêts/assurés SENIOR (externe du portefeuille d'étude).

1.3.2 Cohérence des dates

Cette partie a pour objectif de vérifier l'exactitude de la date de souscription du prêt, de la date de naissance de l'assuré et de la date de fin du prêt.
Le schéma d'une observation dans la table fusionnée doit être le suivant :

GRAPHIQUE 1.4 – Schéma d'une observation



Ce schéma nous a permis de réaliser les tests de cohérences ci-dessous.

Check	Naissance < Adhésion	Adhésion < Décla sinistre	Décla sinistre < Règlement	Règlement < Clôture	Clôture < Fin prêt
OK	100%	100%	100%	100%	100%
NOT OK	0%	0%	0%	0%	0%
MISS	0%	0%	0%	0%	0%

TABLEAU 1.6 – Test de cohérence entre les dates.

Après cette étape de vérification de la cohérence des dates, nous pouvons ainsi les utiliser pour créer de nouvelles variables qui s'avèreront utiles par la suite.

— MIP_AGE_ASSU_DT_SCRP=Year(MIP_DT_SCRP) - Year(MIP_DT_NAISS) ;

— $MIP\_AGE\_PRET = \begin{cases} min(MIP\_DT\_ARRETE; MIP\_DT\_FIN) - MIP\_DT\_SCRP \\ \text{si } non\ DC\ et\ RA \\ MIP\_DT\_RA - MIP\_DT\_SOUSCRIPTION \\ \text{si } RA \\ MIP\_DT\_SINISTRE - MIP\_DT\_SOUSCRIPTION \\ \text{si } DC \end{cases}$

Enfin, l'analyse des valeurs aberrantes des variables de notre portefeuille est aussi réalisée.

1.4 Statistiques descriptives

Dans cette section, nous avons effectué une description de notre portefeuille d'étude à la suite des retraitement opérés. Cette exploration est importante puisqu'elle nous permet d'avoir une vue globale de notre échantillon et nous guide à choisir des méthodes adaptées.

1.4.1 Analyse uni-variée

1.4.1.1 Analyse des variables quantitatives

Le tableau A.1 en annexe résume les statistiques élémentaires des variables quantitatives de notre portefeuille.

Il en ressort de l'analyse de ce dernier que :

— L'âge moyen des emprunteurs de notre portefeuille à la souscription est de 42 ans et ils ont souscrit en moyenne leurs prêts en 2005 ;

— Plus de 75% des emprunteurs assurent leurs prêts à 100% ;

— Moins de 1% des prêts ont une durée réelle quasi-nulle (moins d'un mois). Ce résultat pourrait être expliqué par le fait que les emprunteurs de ces prêts se sont trouvés dans des conditions particulières pour racheter leurs emprunts ;

— Le capital emprunté par nos individus est en moyenne égal à 24 910 euros alors que le capital assuré est à l'ordre de 20 922 euros en moyenne. Le tarif d'assurance associé à ces capitaux s'élève à 0,0032 conduisant à une prime annuelle de 66,88 euros en moyenne ;

— Les prêts à taux zéro représentent environ 1% du portefeuille.

En outre, la même analyse effectuée au sein de la sous-population des emprunteurs décédés (voir tableau A.2 en annexe) donne les résultats ci-dessous :

— 25% des décès concernent les assurés âgés de plus de 64 ans alors qu'ils ne représentent que 10% de la population assurée ;

— Plus de 75% de la population sinistrée possède une quotité égale à 100% ;

— L'âge maximal de ces prêts est de 22 ans. De plus, l'âge minimale est 0, cela veut dire qu'il existe des emprunteurs dont le sinistre est survenu moins d'un mois après la souscription du prêt. En pratique, une telle situation ne doit normalement pas déclencher une indemnisation car un délai de carence est généralement appliqué. Nous gardons quand même ces emprunteurs dans l'apprentissage de nos modèles.

En somme, nous constatons que les variables quantitatives changent de distribution quand on passe de l'échantillon global au sous-échantillon des décès.

1.4.1.2 Analyse des variables qualitatives

Les graphiques de A.2 à A.6 en annexe représentent la composition de notre portefeuille suivant les différents caractères qualitatifs. Il en ressort de l'analyse de ces graphes que la population assurée est composée d'un peu plus d'hommes (51,86%). La répartition suivant la catégorie socio-professionnelle nous donne une représentativité plus élevée chez les employés (37,19% de la population d'étude). Ils sont suivis par les cadres (21,25%) et les professions intermédiaires (15,28%). Les agriculteurs exploitants sont les moins représentés avec seulement 0,42%. Pour ce qui est de la nature du prêt, nous constatons que la majorité d'entre eux sont des prêts pour financer des achats immobiliers (71,72%) contre 28,28% des prêts pour financer des achats de consommation. La répartition suivant le type de garantie nous montre que le risque "Décès" est rarement couvert seul. En effet, seuls 12,59% des emprunteurs prennent la garantie "Décès" contre 87,41% pour la garantie "Décès + IT". La grande majorité des prêts présents dans notre portefeuille d'étude sont des prêts dont le risque assuré est porté par plusieurs individus (68,79%). D'un autre côté, les multi-emprunteurs (ceux qui détiennent plus d'un prêt) sont plus représentés avec une proportion de 65,18%.

1.4.2 Analyse bi-variée

1.4.2.1 Matrice de corrélation des variables quantitatives

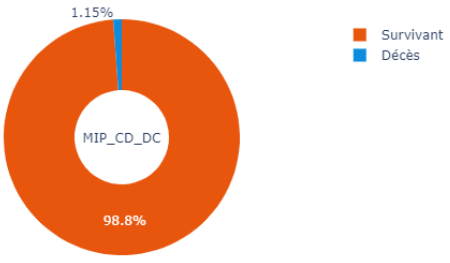
Nous avons représenté au niveau de la figure A.7 en annexe la matrice de corrélation de nos variables quantitatives en utilisant le coefficient de corrélation de *Pearson*. L'analyse de cette figure montre une forte corrélation positive entre le capital initial, le capital assuré et la prime annuelle globale. Nous constatons aussi une forte corrélation positive entre la durée totale du prêt et l'âge du prêt. La génération (année de souscription du prêt) est très négativement corrélée à la durée théorique et l'âge du prêt. Une légère corrélation négative est notée entre l'âge à la souscription et les variables durée théorique et âge du prêt.

1.4.2.2 Matrice de corrélation des variables qualitatives

Le graphique A.8 en annexe représente la matrice de corrélation de nos variables qualitatives en utilisant le *V de Cramer*. L'analyse de cette figure montre une corrélation moyenne entre la catégorie socio-professionnelle et le choix de couverture ("DC" ou "DC + IT"). Une légère corrélation est notée entre le sexe de l'emprunteur et sa catégorie socio-professionnelle.

1.4.3 Analyse des variables d'intérêts en fonction des variables explicatives

1.4.3.1 Distribution de la variable de décès



GRAPHIQUE 1.5 – Distribution de la variable de décès.

Le graphique 1.5 ci-dessus nous permet de voir en bleu la proportion de couple (prêt, assuré) ayant subi l'événement décès durant la durée de leur prêt. Nous constatons à quel point notre jeu de données original est déséquilibré. En effet, la plupart des couples (prêt, assuré) ne subissent pas le décès. Si nous utilisons cet ensemble de données comme base pour nos modèles prédictifs

et nos analyses, nous risquons d'obtenir un grand nombre d'erreurs et nos algorithmes seront probablement sur-ajustés puisqu'ils "supposeront" que la plupart des prêts ne subissent pas le décès. Mais nous cherchons ici à apprendre nos modèles à détecter les caractéristiques qui informent sur le risque de décès d'un assuré. Nous allons utiliser à cet effet des techniques de rééchantillonnage adaptées pour contourner ce problème.

Les graphiques A.10 à A.20 en annexe représentent la distribution du taux de décès en fonction des modalités des différentes variables explicatives. L'analyse de ces graphes laissent entrevoir un taux de décès plus important chez les hommes, les prêts pour achats immobiliers, les emprunteurs individuels, les retraités, les quotités de plus de 0.8. Nous constatons aussi que les couples (prêts, assurés) qui font appels à des tranches de capitaux élevés ou à des durées de prêts de 15 à 23 mois sont les plus exposés aux décès. En outre, les assurés qui appartiennent à la tranche d'âge 55 ans et plus au moment de la souscription sont ceux qui subissent le plus le décès.

1.4.3.2 Analyse descriptive de l'arrêt de travail

Nous avons réalisé des statistiques descriptives sur les bases de données de fréquence et de coût de l'arrêt travail.

Tout d'abord, nous constatons une surreprésentation de la valeur 0 pour le nombre de sinistres IT. En effet, le quantile d'ordre 75% est égal à 0 ce qui veut dire qu'au moins 75% des nombres de sinistres sont nuls. Les figures A.21 à A.27 tracent les diagrammes en fréquence relative et en proportion des nombres de sinistres IT par variable explicative. Il ressort de l'analyse du nombre de sinistres en fonction de l'âge à la souscription que l'effet de l'âge diffère suivant la plage d'âge considéré. En effet, les emprunteurs dont les âges sont compris entre 43 et 55 ans sont ceux qui subissent plus l'IT. De même, les ouvriers présentent une fréquence d'arrêt travail plus importante par rapport aux autres catégories socioprofessionnelles. En ce qui concerne le sexe, nous ne constatons pas une différence très significative de la répartition de la fréquence IT chez les hommes et les femmes.

Pour l'analyse du coût des sinistres IT, nous avons commencé par représenter la fonction de répartition empirique du montant des sinistres. Nous notons une queue de distribution élevée puis que la convergence vers 1 est particulièrement lente. Ensuite, l'analyse de la figure A.30 laisse entrevoir que la tranche d'âge [30, 45) est la plus risquée en termes de variabilité. De même pour les caractéristiques du prêt, on observe un effet de classe sur la répartition de la variabilité.

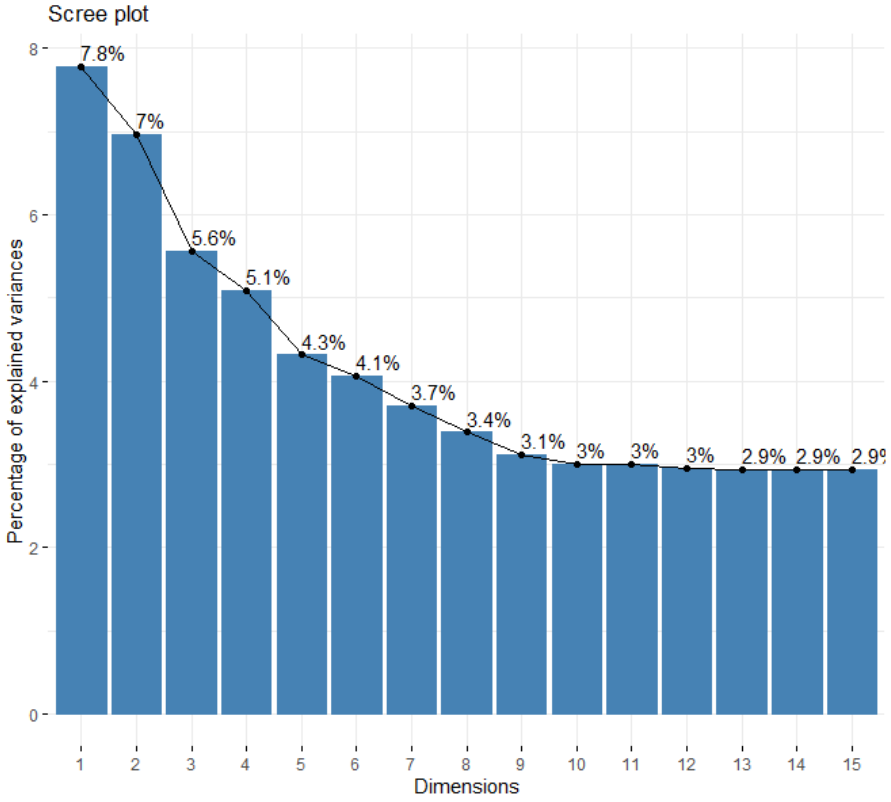
Enfin, nous avons représenté au niveau des figures A.31 à A.36 les boxplots du montant des sinistres en fonction des variables explicatives. L'analyse de ces figures laisse entrevoir que

les niveaux médians des montants de sinistres sont quasiment identiques quel que soit la classe considérée mais les premiers et troisièmes quartiles sont très différents, ainsi que les écart-types.

1.4.4 Analyse multivariée : Analyse des correspondances multiples (ACM)

L'objectif de cette partie est de visualiser simultanément les relations qui existent entre les différentes variables de notre portefeuille. Comme l'ACM ne s'applique que sur des variables qualitatives, nous avons commencé par discrétiser nos variables quantitatives. La méthode CART (Classification And Regression Tree) a été utilisée à cet effet (voir annexe A.2 pour plus de détails sur la discrétisation avec CART).

La première étape consiste à choisir le nombre d'axes pour l'interprétation de nos résultats. Pour ce faire, nous avons appliqué la règle du *coude* qui consiste à choisir le nombre d'axes à partir du point où on observe une décroissance très rapide de l'inertie. Nous avons représenté le graphique 1.6 de décroissance de l'inertie ci-dessous. En appliquant cette règle, nous choisissons de garder les quatre premiers axes.

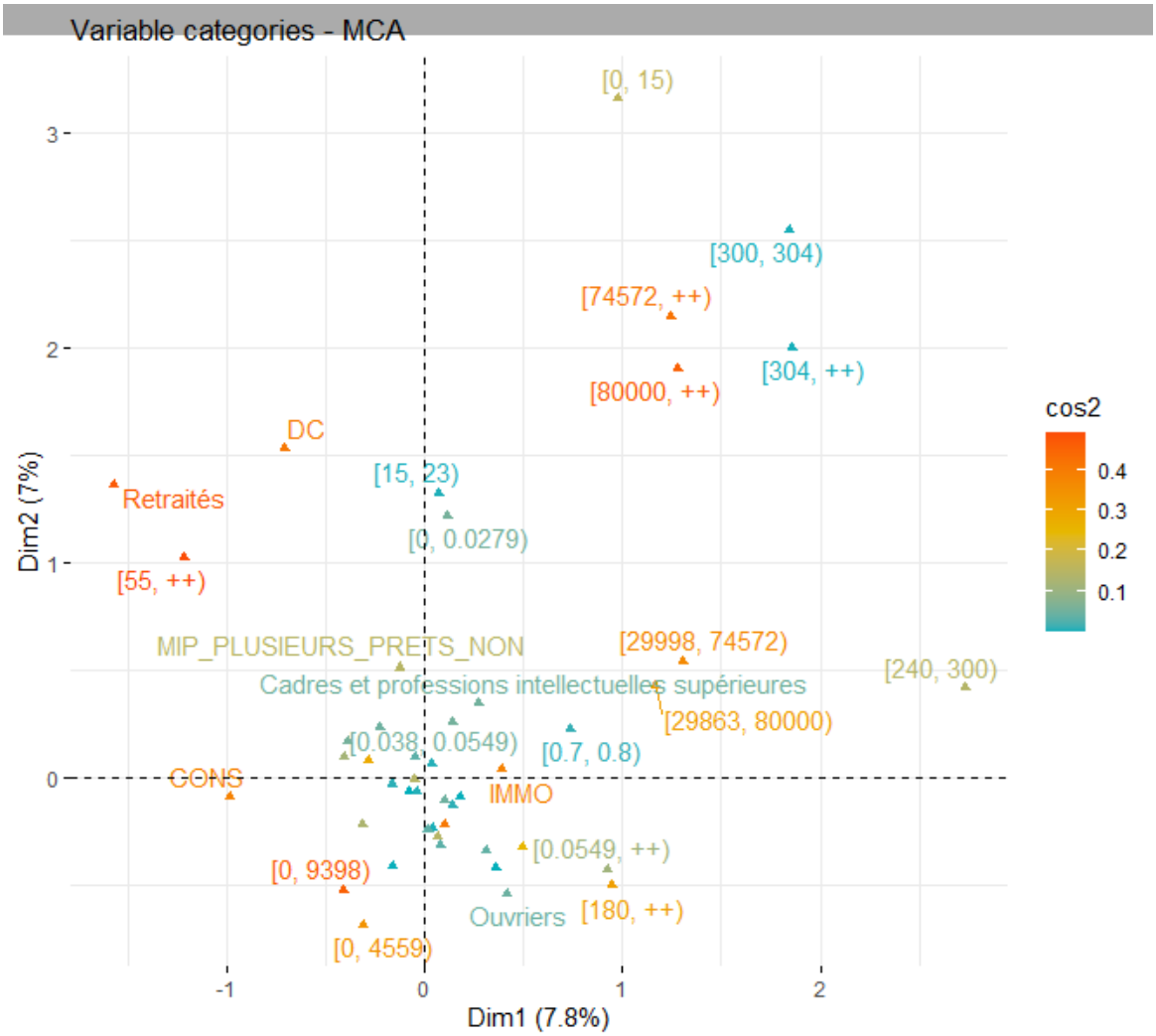


GRAPHIQUE 1.6 – Pourcentage d'inertie suivant les axes.

La figure 1.7, ci-dessous, représente le graphe des liaisons ou nuage de points des modalités. La coordonnée sur un axe d'une modalité est le rapport de corrélation au carré entre la modalité et

la dimension. Il ressort de l'analyse de ce graphe que le premier axe est positivement corrélé aux prêts destinés à des achats immobilliers ("IMMO") et négativement corrélé aux prêts destinés à des achats de consommation ("CONSO"). Pour ce qui est du deuxième axe, nous remarquons qu'il est positivement corrélé aux prêts qui mobilisent des capitaux élévés et négativement corrélé aux prêts qui mobilisent moins de capitaux.

D'autre part, nous remarquons à partir du graphe que les modalités "Retraités" et "[55, ++)" sont proches : on parle d'association entre ces deux modalités. Les emprunteurs "Retraités" appartiennent généralement à la tranche d'âge "[55, ++)". En outre, nous observons aussi que les assurés qui empruntent des capitaux élevés appartiennent généralement à des tranches de durées importantes. Enfin, les ouvriers sont ceux qui empruntent les capitaux les moins importants.



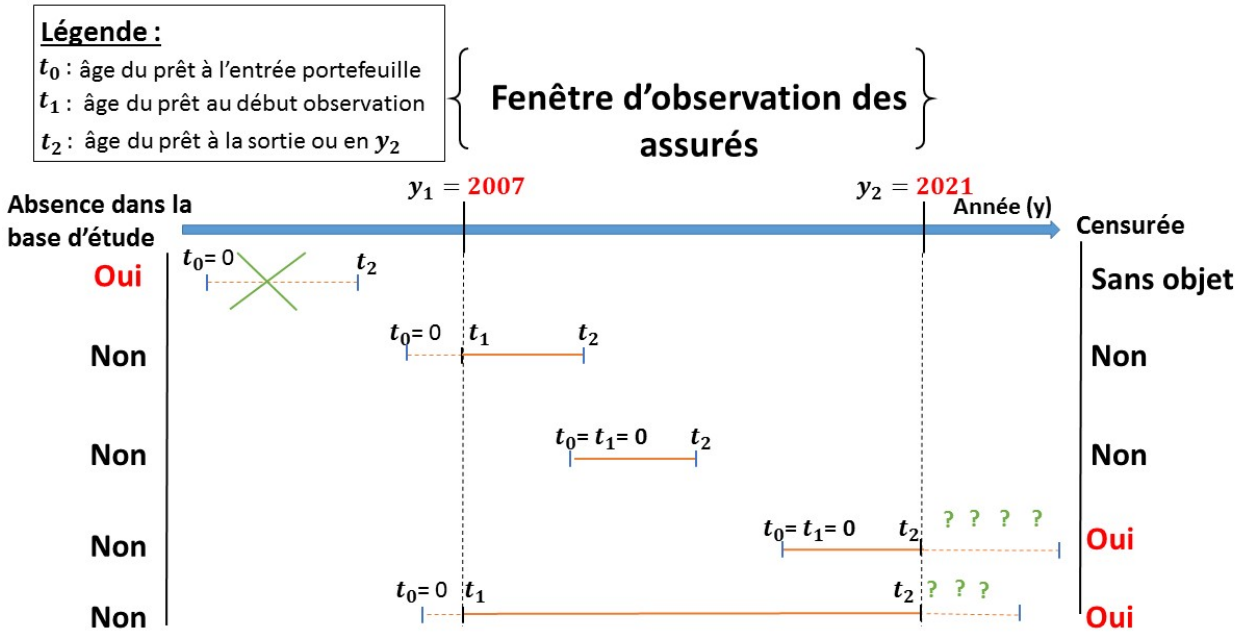
GRAPHIQUE 1.7 – Nuage de points des modalités.

2) Segmentation du portefeuille par scoring

LES travaux décrits dans le premier chapitre ont permis de stabiliser notre base d'étude en l'enrichissant des variables, a priori, pertinentes pour notre étude.
Ce deuxième chapitre consiste à caractériser les variables discriminantes sur les garanties décès et arrêt de travail afin de constituer des groupes homogènes de risque. Pour ce faire, nous commençons par présenter la segmentation du portefeuille avec le risque Décès. Ensuite, nous faisons de même pour le risque Incapacité de Travail. Enfin, nous mettons en place une segmentation globale intégrant les deux risques.

2.1 Segmentation des emprunteurs pour la garantie DC

Nous allons faire, tout d'abord, un point sur la typologie des observations (prêt, assuré) qu'on a dans notre portefeuille et les données qui seront retenues dans la segmentation.
Le graphique 2.1 ci-dessous représente les schémas possibles d'une observation (prêt, assuré). Ce graphique est important pour la suite dans la mesure où la segmentation mise en place n'utilise que les observations qui sont non censurées.



GRAPHIQUE 2.1 – Les schémas possible d'une observation (prêt, assuré).

2.1.1 Modèle *GLM* et régression logistique

La réalisation du décès est codifiée dans notre base d'étude par une variable binaire. La méthode usuelle pour modéliser ce phénomène est la régression logistique. Elle appartient à la classe des modèles linéaires généralisés (*GLM*) et permet de modéliser la probabilité d'occurrence d'un événement binaire à partir de données observées catégorielles ou continues.
Nous nous sommes intéressés dans ce cas à la recherche des covariables optimales à retenir dans le modèle final. Pour ce faire, différents types de modèles de régression logistique sont construits à partir de différentes combinaisons de variables explicatives. Le choix du modèle optimal sera fait en utilisant les critères de performances présentés dans cette section.

La variable binaire modélisée se définit comme suit :

$$Y_i = \begin{cases} 1 \text{ si l'emprunteur } i \text{ décède durant sa durée de prêt} \\ 0 \text{ sinon} \end{cases}$$

Un modèle *GLM* est défini sous la forme :

$$h(\mathbb{E}[Y \mid X = x]) = x\beta$$

où h désigne la fonction de lien et β le paramètre du modèle.
Le principe consiste à se ramener à une régression linéaire via la fonction de lien.

La régression logistique constitue un cas particulier du modèle *GLM* où la fonction de lien correspond à la fonction logit définie par :

$$\text{logit}(z) = \log\left(\frac{z}{1-z}\right)$$

La fonction logistique est très utile car elle permet d'obtenir une image $\Phi(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$ dans $[0,1]$ à partir d'un antécédent $z \in R$ et respecte la propriété de non-décroissance des fonctions de répartition. Cette propriété nous ramène à la définition d'une probabilité.
La variable z représente l'exposition à un ensemble de facteurs de risque et est appelée prédicteur linéaire. Elle est donnée par l'équation de régression classique

$$z = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

où les X_i sont les covariables (explicatives), par exemple l'âge. Ainsi $\forall\, i \in \{1, ..., k\}$, β_i représente le coefficient de régression associé au facteur de risque X_i. Nous noterons les coefficients de régression $\beta = (\beta_0, \ldots, \beta_k)^T$ et les variables $X = (X_1, \ldots, X_k)^T$. En considérant une approche de régression stricte, le principe est de transformer la sortie d'une régression linéaire classique afin d'obtenir une probabilité en utilisant une fonction de lien.

On estime les paramètres du modèle par la méthode du maximum de vraisemblance. La fonction de vraisemblance pour une distribution binomiale est, par définition, la suivante

$$L(\beta, X) = \prod_{i=1}^{n} \Phi\left(X_i\beta'\right)^{Y_i} \left(1 - \Phi\left(X_i\beta'\right)\right)^{1-Y}$$

La log-vraisemblance est donc :

$$\ln(L(\beta, X)) = \sum_{i=1}^{n} Y_i \left(X_i\beta'\right) - \ln\left(1 + e^{X_i\beta'}\right)$$

L'estimateur du maximum de vraisemblance satisfait $\frac{\partial \ln(L)}{\partial \hat{\beta}} = 0$. Il en découle l'estimation des coefficients $\hat{\beta}_i$ de la régression en utilisant les algorithmes d'optimisation numériques (*Newton Raphson* par exemple).

2.1.2 Critères de Comparaison

Après cette première phase de modélisation, plusieurs modèles de régression logistique avec différentes combinaisons de variables explicatives peuvent être retenus. Dès lors il faudra choisir le plus performant. Il existe plusieurs critères de comparaison de modèles dans la littérature pour la régression logistique. Dans le cadre de cette étude, nous avons choisi d'utiliser un critère généralement mentionné à cet effet : la **courbe de ROC**. Ce choix est motivé par le succès de ce critère dans la littérature mais aussi par sa facilité d'interprétation. Nous allons décrire ci-dessous le fonctionnement de ce dernier.

La courbe de ROC (Receiver Operating Characteristic)

C'est une courbe qui place en ordonnée, la sensitivité (capacité à prédire le décès) et en abscisse la spécificité (capacité à prédire la survie). Comme pour le taux de bien classés, on regroupe les scores mais cette fois-ci en deux classes suivant un seuil s que l'on se fixe. On suppose que les clients dont le score est en-dessous de s ne pourront pas réaliser l'évènement (appelons cette classe : les négatifs) et ceux dont le score est plus élevé le réaliseront (les positifs). On notera que dans chacune de ces classes, il pourrait y avoir des décès (Positifs) et des survivants (Négatifs).

Dès lors, la sensitivité est la proportion des Positifs classées positifs et la spécificité, la proportion des Négatifs classés positifs. Le modèle parfait serait celui pour lequel la sensitivité est toujours 1 (ce qui impliquerait que la spécificité est toujours nulle). Par conséquent, on choisira le modèle pour lequel l'aire entre la courbe de ROC et la première bissectrice est maximale.

2.1.3 Eléments d'interprétation du modèle final avec les Odds-ratios

Ils correspondent à l'évolution du rapport de cotes lorsque l'on augmente la variable explicative d'une unité par rapport à la modalité de référence. Les variables ou modalités présentant un Odds-Ratio supérieur à 1 agissent positivement sur le décès. Inversement si l'Odds-Ratio est inférieur à 1, on exercerait une influence négative sur le décès. On n'interprète que les Odds-Ratios dont l'intervalle de confiance ne contient pas la valeur "1".
Exemple : Pour la variable "Sexe" (hommes =1, femmes = 2) on a par exemple un Odds-Ratio de 1,35 pour la modalité "Homme". Cela signifie donc que les hommes ont 35% de chance de plus de réaliser le décès par rapport aux femmes.

2.1.4 Construction de la grille de score

Après cette étape, il s'en suit la construction du score de risque pour chaque emprunteur. Celle-ci nous permet de séparer au mieux les "bons" emprunteurs des "mauvais" emprunteurs par rapport au risque de décès. La fonction de score ou note de risque, se calcule comme une combinaison linéaire des variables explicatives :

$$S(X_i) = \beta_0 + \beta_1 \times x_i^1 + \ldots + \beta_k \times x_i^k.$$

Dans notre cas, le score du risque de décès ou "note de risque" est une mesure de la probabilité qu'un emprunteur décède sachant toutes ses spécificités. Nous normalisons cette combinaison de telle sorte que le score soit compris entre 0 et 100. Il s'agit de calculer une note par modalité. Le principe est qu'on calcule une note pour chacune des modalités de chaque variable. Cette note est normalisée puisqu'on la réduit par la somme de toutes les notes. On peut ensuite additionner pour un emprunteur chaque note en fonction de son profil et on obtient finalement un score "normalisé" entre 0 et 100. Concrètement, la note se calcule comme suit :

$$\text{Note } = \frac{\text{Coef}_{\text{Modalité}} - \text{ Coef}_{\text{Min de la variable}}}{\sum_{\text{chaque variable}} \left(\text{ Coef}_{\text{Max de la variable}} - \text{ Coef}_{\text{Min de la variable}} \right)} * 100$$

Le coefficient de la modalité (respectivement maximum et minimum de la variable) est le produit entre le numéro de la modalité au niveau de la variable (respectivement numéro maximal et

minimal) et le coefficient estimé de la variable.

2.1.5 Application à notre portefeuille

Présentation du modèle

Nous avons utilisé la régression logistique pour segmenter les emprunteurs suivant le risque Décès. Pour ce faire, nous avons modélisé la probabilité de décès (MIP_CD_DC="OUI") durant la durée du prêt en fonction des caractéristiques de l'emprunteur et du prêt. L'échantillon utilisé pour la régression logistique est obtenu avec la méthode "under sampling" à cause du déséquilibre de notre base de données. Une fois les données équilibrées, nous avons découpé la base en 75% d'apprentissage et 25% de test. L'estimation s'effectue sur notre échantillon d'apprentissage. L'ensemble des variables explicatives sont qualitatives et nous avons donc spécifié une modalité de référence pour chacune d'entre elles. Cette référence représente la modalité la moins risquée de manière à obtenir des coefficients positifs facilitant l'interprétation et la construction de la grille de score.

Nous avons utilisé trois démarches pour la sélection de nos variables explicatives : backward, forward et stepwise. Le modèle optimal est obtenu en comparant les statistiques d'ajustement du modèle et les critères de qualité de prévision. Les critères AIC (minimisation) et AUC (maximisation) sont utilisés à cet effet.

Tests des effets croisés

Avant de lancer la sélection automatique des variables explicatives, nous avons commencé par tester l'existence d'effets croisés entre nos différentes variables explicatives. Pour ce faire, nous avons lancé la régression dans les deux sous-populations celles des hommes et des femmes. Cette opération nous a permis de ne pas effectuer la construction de deux scores de risque selon le genre. En effet, l'impact des variables explicatives reste fixe dans les deux sous-populations. Ensuite, nous avons testé l'apport des croisements entre le sexe et la CSP d'une part et l'âge et la CSP d'autre part. Ces effets croisés ne se sont pas révélés significatifs pour la construction de notre score de risque.

Analyse des résultats

Le modèle choisi est bien spécifié selon les trois tests de significativité globale (Wald, Likelihood Ratio et Score). D'un autre côté, le R^2 ajusté du modèle confirme également cette hypothèse.

La sélection automatique a éliminé les variables : MIP_CAP_INIT, MIP_PLUSIEURS_ASSURES, MIP_PLUSIEURS_PRETS, et MIP_QUOTITE_GLOBALE. L'analyse des effets de type 3

ou bien test de significativité individuelle des variables explicatives permet de déterminer si chaque variable prise individuellement contribue significativement à la probabilité de décéder durant la durée du prêt. Toutes les variables ont des p-valeurs inférieures à notre seuil de confiance, elles sont donc significatives.

L'analyse du khi-2 de Wald permet de savoir quelles modalités sont significatives. Cette statistique montre que toutes les modalités sont significatives (pvalue $< 0,05$).
Nous avons aussi testé l'adéquation du modèle aux données avec le test d'Hosmer et Lemeshow. Ce test permet de vérifier si la probabilité prédite est proche de la probabilité observée. La statistique du test est de $8,2$ qui est inférieure à $15,5$ (quantile du khi-2 à huit degrés de liberté), alors les données ne contredisent pas l'hypothèse H_0 ; le modèle est adéquat et les probabilités de décès estimées correspondent aux probabilités de décès observées.

Backtesting du modèle

Le backtesting du modèle retenu est indispensable pour la validation du score. En effet, ce dernier a été construit sur notre échantillon d'apprentissage, on doit donc contrôler sa robustesse sur un échantillon qui n'a pas servi à sa construction. Pour ce faire, nous avons utilisé l'échantillon test obtenu à partir des données déséquilibrées qui est représentatif de notre portefeuille global.

La courbe de ROC représentée dans la figure 2.2 indique que l'aire sous la courbe (AUC) obtenue avec cet échantillon test est de 0,79 donc la qualité de discrimination de nos emprunteurs est satisfaisante.

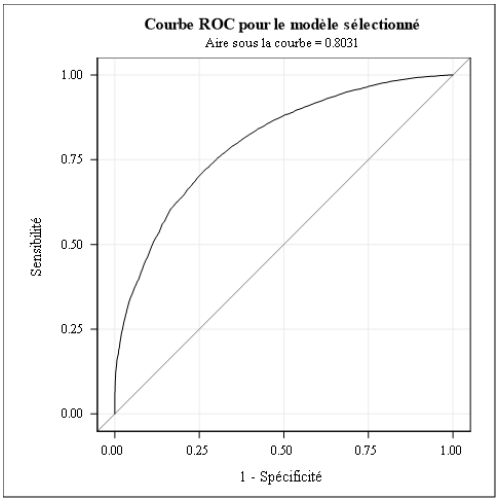
En outre, nous avons représenté, d'une part, la matrice de confusion en utilisant l'échantillon d'apprentissage issu de notre base équilibrée. D'autre part, nous avons représenté la matrice de confusion en utilisant l'échantillon de validation issu de la base initiale déséquilibrée. En effet, c'est ce dernier échantillon qui est représentatif de notre population de départ. Les métriques obtenues sur un échantillon de validation autre que celui-ci doivent être corrigées pour tenir compte du rééquilibrage.
Il en ressort de l'analyse de ce tableau que pour un nouvel emprunteur, le modèle détecte correctement 7 sur 10 s'il va effectivement décéder sur la période de prêt ou pas.

		Prédits	
		DC	Pas DC
Observés	DC	71,8%	28,2%
	Pas DC	27,8%	72,2%

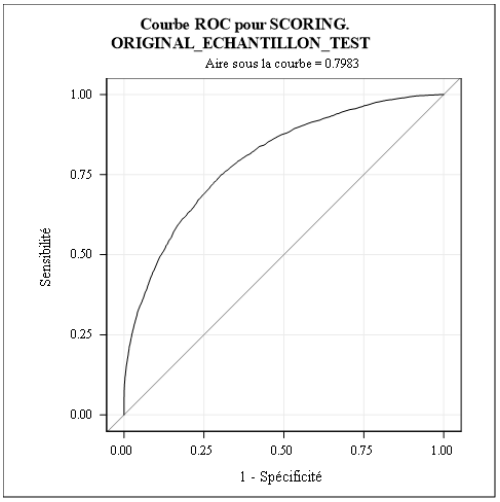
TABLEAU 2.1 – Matrice de confusion-échantillon d'apprentissage.

		Prédits	
		DC	Pas DC
Observés	DC	70,8%	29,2%
	Pas DC	27,8%	73,3%

TABLEAU 2.2 – Matrice de confusion-échantillon de validation original.



GRAPHIQUE 2.2 – Courbe de ROC-échantillon d'apprentissage.



GRAPHIQUE 2.3 – Courbe de ROC-échantillon validation original.

Scoring des emprunteurs

La grille de score est construite à partir des estimations sur l'échantillon d'apprentissage. Les coefficients des modalités de référence sont fixés par définition à 0. Plus la note affectée est importante plus le risque de décès est élevé.

Les résultats de la grille de score sont présentés dans le tableau 2.3.

Une fois la grille de score établie, nous découpons notre score en différentes classes de risque afin d'avoir des taux de décès différents pour chaque classe. Nous avons utilisé la commande *proc rank* de SAS pour la construction des classes de risques. Nous avons ainsi regroupé les emprunteurs en six classes. Le numéro de classe est croissant avec le risque de décès.

L'emprunteur le moins risqué est une femme qui a entre 18 et 38 ans, avec une durée totale de prêt comprise entre 0 et 118 mois. Cet emprunteur détient un crédit dont le capital assuré est compris entre 0 et 3 994 euros et il fait partie des cadres et professions intellectuelles supérieures. Enfin, la garantie choisie par ce dernier est "DC + IT".

Variables	Modalités	Estimation	Contributions	Notes	Niveau risque
MIP_AGE_ASSU_DT_SCRP_D	[18, 38)	0	21	0	très peu risqué
MIP_AGE_ASSU_DT_SCRP_D	[38, 45)	0,9521	21	9	Intermédiaire
MIP_AGE_ASSU_DT_SCRP_D	[45, 55)	1,5979	21	15	risqué
MIP_AGE_ASSU_DT_SCRP_D	[55, ++)	2,2891	21	21	très risqué
MIP_CD_SEXE_D	Femme	0	7	0	très peu risqué
MIP_CD_SEXE_D	Homme	0,7174	7	7	risqué
MIP_DUR_TOTALE_PRET_D	[0, 118)	0	56	0	très peu risqué
MIP_DUR_TOTALE_PRET_D	[118, 194)	1,3365	56	12	Intermédiare
MIP_DUR_TOTALE_PRET_D	[194, 258)	2,8456	56	26	risqué
MIP_DUR_TOTALE_PRET_D	[258, ++)	6,1368	56	56	très risqué
MIP_CI_ASSU_GLOB_D	[0, 3994)	0	5	0	très peu risqué
MIP_CI_ASSU_GLOB_D	[3994, 29769)	0,3670	5	3	Intermédiaire
MIP_CI_ASSU_GLOB_D	[29769, 75110)	0,3011	5	3	Intermédiaire
MIP_CI_ASSU_GLOB_D	[75110, ++)	0,6039	5	5	risqué
MIP_CSP_D	Ouv/Agr/Art/NR	0,5665	8	5	risqué
MIP_CSP_D	Retraités	0,4525	8	4	risqué
MIP_CSP_D	Cadres	0	8	0	très peu risqué
MIP_CSP_D	Employés	0,4110	8	4	risqué
MIP_CSP_D	Sans act Pro	0,9319	8	8	très risqué
MIP_CSP_D	Pro Inter	0,2355	8	2	Intermédiaire
MIP_TYPE_GARANTIE	DC + IT	0	3	0	très peu risqué
MIP_TYPE_GARANTIE	DC	0,3178	3	3	risqué

TABLEAU 2.3 – Grille de score pour le risque DC.

Segmentation	Score	Effectif (%)	Taux de décès observé
Classe 1	0 à 14 points	18,02%	0,12%
Classe 2	15 à 19 points	14,96%	0,30%
Classe 3	20 à 24 points	15,25%	0,47%
Classe 4	25 à 28 points	17,97%	0,80%
Classe 5	29 à 34 points	16,75%	1,27%
Classe 6	35 à 100 points	17,06%	4,07%

TABLEAU 2.4 – Segmentation du portefeuille suivant le risque DC.

2.2 Segmentation des emprunteurs pour la garantie IT

Nous avons réitéré le processus de segmentation de la garantie DC ci-dessous pour la garantie IT. De la même manière que pour le DC, nous avons obtenu le score de risque en IT, compris en 0 et 100, pour chaque emprunteur. La variable binaire modélisée ici se définit comme suit :

$$Y_i = \begin{cases} 1 \text{ si l'emprunteur } i \text{ a été au moins une fois en IT durant son prêt} \\ 0 \text{ sinon} \end{cases}$$

Les résultats de la grille de scoring pour cette partie sont renseignés en annexes A.3.

Une analyse des résultats montre que le profil le moins risqué pour la garantie IT est un emprunteur âgé de plus de 55 ans. Il détient un prêt destiné à des achats consommation dont

la durée totale est comprise entre 0 et 118 mois. Enfin, cet emprunteur fait partie des cadres.

Segmentation	Effectif (%)	Taux d'entrée en IT
Classe 1	16,70%	0,60%
Classe 2	15,42%	1,38%
Classe 3	17,92%	1,62%
Classe 4	11,77%	2,81%
Classe 5	18,97%	3,54%
Classe 6	19,23%	6,62%

TABLEAU 2.5 – Segmentation du portefeuille suivant le risque IT.

2.3 Segmentation intégrant les deux risques

Les deux sections ci-dessous nous ont permis d'avoir un score de risque par emprunteur pour chacune des garanties IT et DC. Dans cette section, nous allons sommer, pour chaque emprunteur, son score de DC et IT. Plus précisément, pour un emprunteur i, son score global est obtenu comme suit :

$$Score\_GLOB\_i = \begin{cases} Score\_i\_DC + Score\_i\_IT \text{ si l'emprunteur } i \text{ choisit la garantie "DC + IT"} \\ Score\_i\_DC \text{ sinon} \end{cases}$$

La formule de calcul du score global ci-dessus suppose que les deux risques DC et IT sont indépendants ce qui n'est réellement pas le cas en pratique. En outre, en attribuant le même poids aux deux risques en faisant une somme simple, nous supposons qu'ils ont le même coût pour l'assureur.

La segmentation globale s'obtient directement en transformant en classes de risques la variable $Score\_GLOB$.

Les résultats obtenus sont présentés dans le tableau ci-dessous.

Segmentation	Effectif (%)	Taux de décès	Taux d'entrée en IT
Classe 1	17,31%	0,45%	0,79%
Classe 2	15,76%	1,06%	1,12%
Classe 3	16,82%	1,04%	1,44%
Classe 4	17,35%	0,65%	2,81%
Classe 5	17,29%	1,11%	3,63%
Classe 6	15,47%	3,09%	6,31%

TABLEAU 2.6 – Segmentation du portefeuille suivant les risques DC et IT

3) Modélisation de la garantie Décès

DANS le chapitre précédent, nous avons constitué des groupes homogènes de risque de notre portefeuille d'étude.

La première partie de notre étude consiste à quantifier le coût individuel de la garantie DC en mettant en place une table d'expérience de maintien en vie dans le prêt. Pour ce faire, il sera d'abord question de présenter les méthodes classiquement utilisées et leurs limites. Puis dans un second temps, nous décrirons l'apport que pourrait avoir des méthodes alternatives d'apprentissage dites *Machine Learning*.

3.1 Table de maintien en vie dans le prêt

Nous allons avoir recours aux modèles de durée en ce sens pour estimer la probabilité de décès d'un emprunteur à x années de son prêt. Dans ce cas, la variable de durée va correspondre à la durée passée en vie dans le contrat de prêt. Nous allons modéliser cette durée en fonction de l'âge à la souscription, la CSP, la durée totale, etc. en utilisant les méthodes classiques et *Machine Learning*.

3.1.1 Modèles classiques : Modèle de *Cox*

L'analyse des données de survie ou encore analyse de survie consiste à étudier le délai de survenance d'un événement au cours du temps (comme le décès). Ainsi, elle permet la modélisation du facteur temps dans la probabilité d'occurrence des événements. Les modèles les plus utilisés à cet effet sont le Cox (1972) et le Kaplan and Meier (1958).

Le modèle de *Cox* à la différence du modèle basique de *Kaplan-Meier*, permet d'avoir un aperçu sur le comportement général des individus qui n'appartiennent pas forcément à des groupes identiques. Il fournit une technique statistique pour explorer la relation entre la survie d'un individu et plusieurs variables explicatives (continues ou non). En actuariat, ces modèles sont généralement utilisés pour construire les tables de mortalité, calibrer les lois de maintien en incapacité/invalidité, les lois de rachat, etc. Dans le cadre de ce mémoire, nous l'utilisons pour construire une table de maintien en vie dans le prêt avec comme variable d'intérêt, la durée passée en vie d'un assuré dans le contrat emprunteur. Un autre aspect important de ce modèle est qu'il permet de prendre en compte les phénomènes de censure fréquemment rencontrés dans

les problématiques d'études de durées.

Fonctionnement du modèle de *Cox*

Le modèle se formule à l'aide d'une relation paramétrique entre la fonction de risque instantanée $h(t,x)$ (en anglais, hazard function) et les facteurs de risque (covariables X) qui sont supposées agir de façon multiplicative sur la fonction de risque de base $h_0(t)$, non explicitée (non paramétrique) au temps t :

$$h(t,x) = h_0(t)\exp\left(\sum_k \beta_k x_k\right) = h_0(t)\exp\left(\beta' X\right)$$

où $X = (X_1, X_2, \ldots, X_p)$ est un vecteur de p variables explicatives et β' est un vecteur de p coefficients de régression à estimer. Ainsi, le modèle s'écrit comme le produit de deux fonctions, l'une $h_0(t)$ étant dépendante du temps mais pas des caractéristiques individuelles, non explicitée, et l'autre $\exp\left(\sum_k \beta_k x_k\right)$ ne dépendant pas du temps mais uniquement des caractéristiques des individus, de forme paramétrique, d'où le nom de modèle semi-paramétrique.

En considérant qu'à chaque instant t_i correspond un évènement, la vraisemblance partielle de *Cox* considère alors la probabilité qu'en t_i, l'individu i subisse l'évènement plutôt qu'un autre individu exposé au risque au même instant. La contribution à la vraisemblance de l'individu i est donnée par :

$$LP_i = \frac{h\left(t, x_i\right)}{\sum_{j\geq i} h\left(t, x_j\right)} = \frac{\exp\left(x_i'\beta\right)}{\sum_{j\geq i}\exp\left(x_j'\beta\right)}$$

La fonction de vraisemblance partielle à maximiser est obtenue en considérant l'ensemble des individus non censurés et prend la forme suivante :

$$LP = \prod_{i\in I_{\text{non censurees}}} \frac{\exp\left(x_i'\beta\right)}{\sum_{j\geq i}\exp\left(x_j'\beta\right)}$$

La vraisemblance partielle ne dépend pas de la fonction de risque de base $h_0(t)$, il est donc possible d'estimer l'effet du facteur explicatif sans connaitre cette fonction, par maximisation de la vraisemblance partielle. Cette dernière prenant généralement de très faibles valeurs, il est préférable de travailler avec son logarithme, la log-vraisemblance.

$$\log(LP) = \sum_{i\in I_{\text{non censurees}}} \left(\left(\exp\left(x_i'\beta\right) - \log\left(\sum_{j\geq i}\exp\left(x_j'\beta\right)\right)\right)\right)$$

Une des hypothèses fondamentales du modèle de *Cox* est la proportionnalité des risques. Plus précisément, elle signifie que le rapport des fonctions de risque à l'instant t ne dépend que des

valeurs des covariables à cet instant. En effet, pour tout t et toute paire d'individus $\{i, j\}$ de risques respectifs $h_i(t) = h(t, x_i)$ et $h_j(t) = h(t, x_j)$ on a :

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp(x_i'\beta)}{h_0(t)\exp\left(x_j'\beta\right)} = \exp\left(\left(x_i' - x_j'\right)\beta\right) = c$$

La validité de ce modèle nécessite la vérification de cette hypothèse. Par ailleurs, dans certains cas et selon certains auteurs, le modèle de *Cox* reste une bonne approximation de la réalité, même dans le cas où l'hypothsèse n'est pas entièrement respectée (Courgeau et Lelièvre, 1989).

Interprétation des résultats du modèle

L'interprétation des paramètres du modèle de *Cox* dépend du type de variable (continue ou discrète) à laquelle il est associé.

Pour une variable continue X_k, le coefficient associé β_k vérifie la condition suivante :

$$\beta_k = \frac{\partial \ln(h_0(t \mid X))}{\partial X_k}$$

Ainsi β_k représente l'élasticité du taux de hasard par rapport à la $k^{\text{ème}}$ covariable continue X_k supposée constante par rapport au temps t. β_k est l'effet de la variable X_k sur le risque instantané, toutes choses égales par ailleurs.

En effet β_k peut être vu comme le rapport de risque quand la variable associée X_k augmente d'une unité.

$$\ln\left(\frac{h_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_k(\boldsymbol{X}_{ik} + 1) + \cdots + \beta_p X_{ip})}{h_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \cdots + \beta_p X_{ip})}\right) = \beta_k$$

- Si $\beta_k > 0$, c'est-à-dire si $\exp(\beta_k) > 1$ alors le risque de décès augmente quand X_k augmente (resp. diminue quand X_k diminue).
- Si $\beta_k < 0$, c'est-à-dire si $\exp(\beta_k) < 1$ alors le risque de décès diminue quand X_k augmente (resp. augmente quand X_k diminue).
- Si $\beta_k = 0$, c'est-à-dire si $\exp(\beta_k) = 1$ alors la variable X_k n'a pas d'impact significatif sur le risque de décès.

Soit X_k une variable qualitative qui ne prend que 2 valeurs 0 ou 1. Prenons par exemple la variable "sexe". Dans ce cas, on a :

$$X_{ik} = \begin{cases} 1 & \text{si l'emprunteur est une femme} \\ 0 & \text{si l'emprunteur est un homme} \end{cases}$$

On en déduit que :

$$\ln\left(\frac{h_0(t)\exp\left(\beta_1 X_{i1}+\cdots+\beta_k\times 1+\cdots+\beta_p X_{ip}\right)}{h_0(t)\exp\left(\beta_1 X_{i1}+\cdots+\beta_k\times 0+\cdots+\beta_p X_{ip}\right)}\right)=\beta_k$$

β_k est alors le rapport de risque de décès entre un homme et une femme toutes choses égales par ailleurs.

— Si $\beta_k > 0$, c'est-à-dire si $\exp(\beta_k) > 1$ alors le risque de décès est plus élevé chez les femmes que chez les hommes.

— Si $\beta_k < 0$, c'est-à-dire si $\exp(\beta_k) < 1$ alors le risque de décès est plus élevé chez les hommes que chez les femmes.

— Si $\beta_k = 0$, c'est-à-dire si $\exp(\beta_k) = 1$ alors la variable "sexe" n'a pas d'impact significatif sur le risque de décès. Le risque de mortalité est le même chez les hommes que chez les femmes.

Hypothèses et limites du modèle de *Cox*

La validité du modèle de *Cox* requiert certaines hypothèses sur les données. La vérification de ces hypothèses repose sur les résidus de *Schoenfeld* (SCHOENFELD (1982)). Ces derniers sont calculés par observation et par variable explicative. Ils ne sont déterminés qu'aux moments observés de l'événement qui est ici le décès. Le résidu de Schoenfeld, r_{ik}, est estimé pour le $i^{ème}$ emprunteur et la $k^{ème}$ variable explicative par

$$\hat{r}_{ik} \;=\; x_{ik} \;-\; \hat{\bar{x}}_{\omega_i k}$$

où :

— x_{ik} est la valeur de la $k^{ème}$ variable explicative pour l'emprunteur i ;

— et $\hat{\bar{x}}_{\omega_i k}$ est une moyenne pondérée des valeurs des variables explicatives pour ceux qui sont soumis au risque défini au moment de l'événement donné.

— Une valeur positive de r_{ik} indique une valeur de X plus élevée que prévu au moment du décès.

La somme des résidus de *Schoenfeld* est nulle. Dans le cas d'une variable binaire, les résidus de Schoenfeld doivent être compris entre -1 et 1. On a alors :

$$\hat{r}_{ik} = \begin{cases} 0-\hat{\bar{x}}_{\omega_i k} & \text{pour } x=0 \\ 1-\hat{\bar{x}}_{\omega_i k} & \text{pour } x=1 \end{cases}$$

Le modèle de *Cox* repose principalement sur l'hypothèse que les risques sont proportionnels. Autrement dit, selon cette hypothèse, les résidus de *Schoenfeld* ne dépendent pas du temps. On peut la vérifier en effectuant un test d'indépendance entre la variable temps et toutes ou chacune des variables explicatives du modèle. Ce test est appelé le test des résidus de *Schoenfeld*. En outre, les effets non linéaires et les interactions non spécifiées ne seront pas pris en compte par ce modèle.

Cette hypothèse forte que les risques sont proportionnels peut être rejetée par certaines données en pratique. Ainsi, des alternatives de ce modèle commencent à faire leur apparition.

Les méthodes de *Machine Learning* sont de plus en plus utilisées par les actuaires comme alternative aux méthodes classiques notamment dans le cas des modèles de durée comme le modèle de *Cox*. Dans la partie ci-dessous, nous allons appliquer les techniques d'apprentissage automatiques à notre portefeuille, afin de modéliser le risque de décès.

3.1.2 *XGBoost Cox* et *XGBoost Survival Embeddings (Xgbse)*

Pour intégrer les effets non linéaires dans le cadre du modèle de *Cox*, une adaptation par *XGBoost* de Chen and Guestrin (2016) est possible. L'implémentation de *XGBoost* fournit deux méthodes d'analyse de survie : *Cox* et *Accelerated Failure Time (AFT)*. Lorsqu'il s'agit de classer les individus par risque, les deux méthodes affichent des performances compétitives (mesurées par le C-index, l'équivalent de l'AUC de la courbe de ROC pour la survie) tout en étant ultra-rapides.

Cependant, nous pouvons observer des lacunes en ce qui concerne d'autres propriétés statistiques souhaitables. Plus précisément, trois propriétés sont concernées :

— prédiction des courbes de survie plutôt que des estimations ponctuelles ;

— estimation des intervalles de confiance ;

— temps de survie attendus calibrés (non biaisés).

Bien qu'il ait besoin d'une extension pour plus de rigueur statistique, *XGBoost* reste un modèle puissant. Les résultats du *C-index* montrent que le modèle peut capturer une grande partie du signal, en étant compétitif par rapport à l'état de l'art. Nous devons simplement adapter la façon dont nous l'utilisons.

C'est dans ce cadre que Vieira et al. (2021) ont proposé une version plus adaptée du modèle *XGBoost* dans le cadre de l'analyse de survie.

Nous allons présenter, ci-dessous, le fonctionnement du modèle *XGBoost* ainsi que sa version adaptée à l'analyse de survie.

Préliminaires

Le modèle *XGBoost* appartient à la famille des modèles de Boosting. Le principe de base de cette famille consiste à construire un ensemble d'estimateurs E qui sont ensuite agrégés par une moyenne pondérée des estimations (pour la régression, $Y = \mathbb{R}$) ou un vote à la majorité (pour la classification, $Y = \mathbb{N}$). Dans notre cas, la variable d'intérêt Y correspond à la durée passée en vie dans le prêt par l'emprunteur.

Dans ce cas, l'ensemble E est créé séquentiellement. À chaque itération k, une nouvelle fonction de base f_k est sélectionnée et ajoutée à l'ensemble de sorte que la perte l de l'ensemble soit minimisée :

$$\ell(\mathrm{E}) = \sum_{k=1}^{K} \ell\left(Y, \hat{Y}^{(k-1)} + f_k(X)\right) + \Omega\left(f_k\right) \tag{3.1}$$

Ici, K représente le nombre d'éléments de E et chaque $f_k \in \mathcal{F}$ avec $\mathcal{F}$ étant l'espace des fonctions de base possibles (communément il s'agit de l'espace des arbres de régression). Le paramètre de régularisation Ω pénalise les fonctions complexes.

L'ensemble est créé à l'aide de la modélisation additive progressive, où les nouveaux arbres sont ajoutés un par un. À l'étape k, les données d'apprentissage sont évaluées sur les éléments existants de l'ensemble et les scores de prédiction correspondants $Y^{(k)}$ sont utilisés pour piloter la création de nouveaux éléments de l'ensemble. Les prédictions des fonctions de base sont combinées de manière additive :

$$\hat{Y}^{(k)} = \sum_{k=1}^{K} f_k(X) = \hat{Y}^{(k-1)} + f_k(X) \tag{3.2}$$

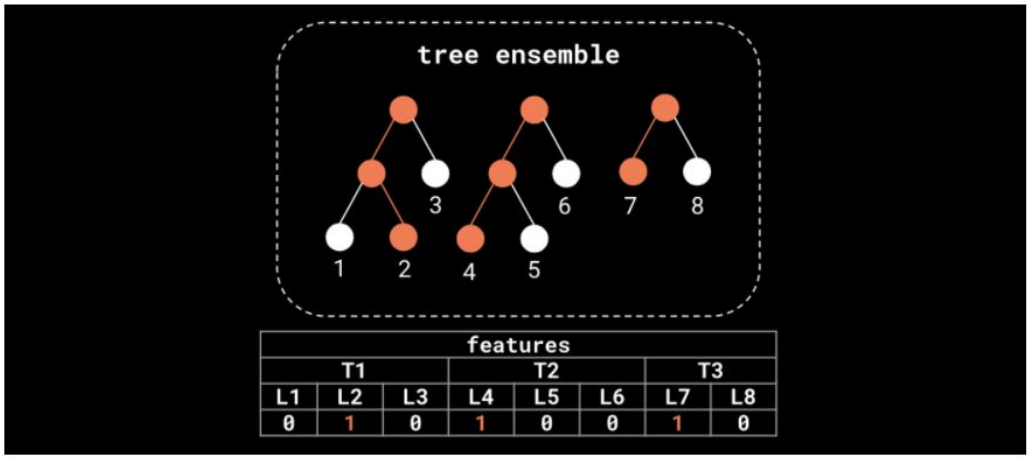
La prédiction finale pour un échantillon $\hat{y}_i$ est la somme des prédictions pour chaque arbre f_k dans l'ensemble.

$$\hat{y}_i = \sum_{k=1}^{K} f_k\left(x_i\right) \tag{3.3}$$

Utilisation de *XGBoost* comme encodeur

En plus d'être utilisées pour les tâches de prédiction, les méthodes de Gradient Boosting peuvent également être utilisées comme transformateurs de caractéristiques des données d'entrée (i.e encodeur). Les arbres qui composent l'ensemble effectuent des divisions sur les caractéristiques qui discriminent la cible, en encodant dans leur structure les informations les plus pertinentes pour la tâche à accomplir. En particulier, les nœuds terminaux (feuilles) de chaque arbre de

l'ensemble définissent une transformation de caractéristiques (embedding) des données d'entrée. La figure 3.1 illustre un cas d'usage avec trois arbres. Il ressort de cette figure que nous pouvons extraire des caractéristiques d'un modèle de forêt comme *XGBoost*, en transformant l'espace de caractéristiques original en un embedding "d'occurrence de feuille". Les nœuds orange représentent le chemin d'un seul échantillon dans l'ensemble.



GRAPHIQUE 3.1 – Utilisation de *XGBoost* comme encodeur (Vieira et al. (2021)).

Ce type d'embedding d'ensemble d'arbres possède des propriétés très pratiques :

— **sparsité et haute dimensionnalité** : les arbres traitent la non-linéarité et transforment les caractéristiques originales en un embedding sparse à haute dimension, ce qui permet aux modèles linéaires d'obtenir de bonnes performances lorsqu'ils sont entraînés sur cet embedding. Cela permet à une régression logistique entraînée sur l'embedding (en tant qu'indices de feuilles "one-hot encoded") d'avoir des performances comparables à celles de l'ensemble réel, avec l'avantage supplémentaire de la calibration des probabilités ;

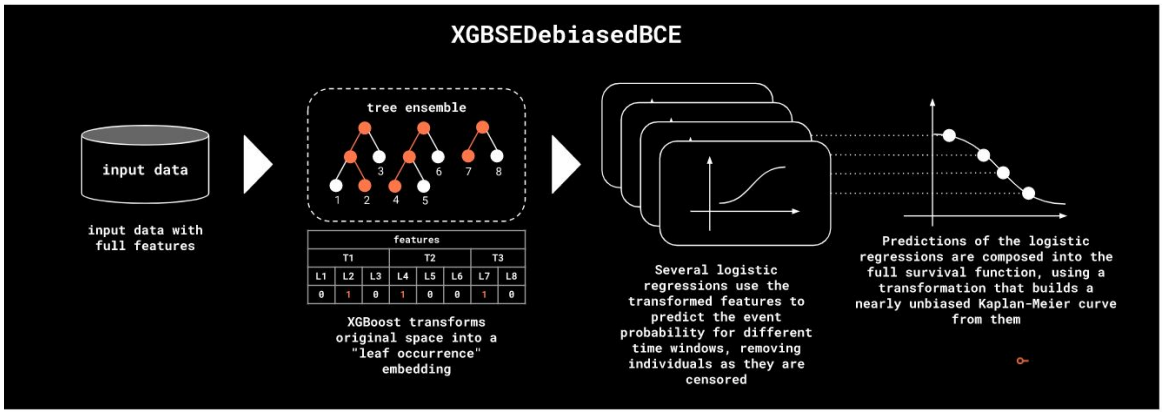
— **supervision** : les arbres fonctionnent également comme un filtre de bruit, en effectuant des divisions uniquement à travers les caractéristiques qui ont un pouvoir prédictif. Ainsi, l'embedding a en fait une dimension intrinsèque plus faible que les données d'entrée. Cela atténue la contrainte de la dimensionnalité et permet à un modèle K-Plus Proches Voisins formé sur l'embedding (en utilisant la distance de Hamming) d'avoir des performances comparables à l'ensemble réel, avec la flexibilité supplémentaire d'appliquer n'importe quelle fonction sur les ensembles de voisins pour obtenir des prédictions. Cette fonction arbitraire peut être, par exemple, un estimateur de survie sans biais tel que l'estimateur de *Kaplan-Meier*.

La méthode *Xgbse* tire parti de ces propriétés de différentes manières comme nous allons le montrer dans les parties suivantes.

***XGBSEDebiasedBCE* : régressions logistiques, fenêtres temporelles, embedding comme données d'entrée**

Cette première approche, *XGBSEDebiasedBCE*, s'inspire de la méthode de régression logistique multitâche de Yu et al. (2011), de l'approche BCE de Kvamme and Borgan (2019) et des idées de calibrage de probabilité de He et al. (2014) et Marmerola and Marmerola (2018).

Il s'agit d'entraîner un ensemble de régressions logistiques par-dessus de l'embedding produit par *XGBoost*, chacune prédisant la survie à différentes fenêtres temporelles discrètes définies par l'utilisateur. Les classificateurs éliminent les individus au fur et à mesure qu'ils sont censurés, avec des cibles qui sont des indicateurs de survie à chaque fenêtre.



GRAPHIQUE 3.2 – XGBSEKDebiasedBCE (Vieira et al. (2021)).

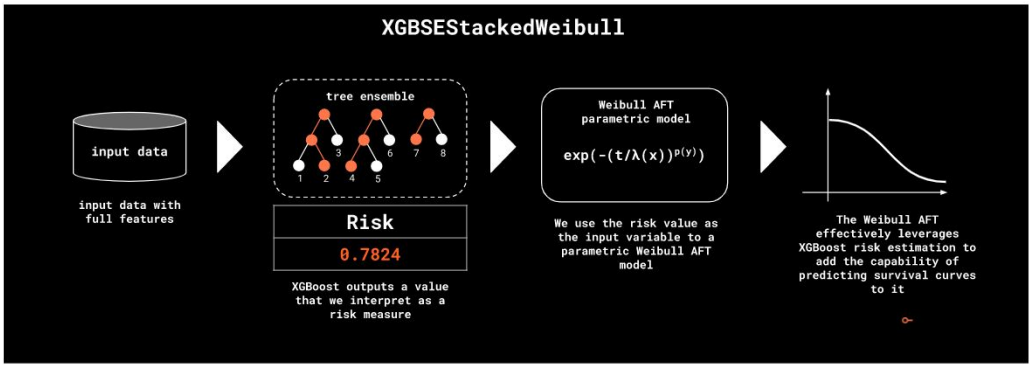
L'approche naïve tend à donner des courbes de survie biaisées, en raison de l'élimination des individus censurés. Il est donc nécessaire de procéder à quelques adaptations afin que les régressions logistiques estiment le terme $\frac{d_i}{n_i}$ (probabilités ponctuelles) dans la formule de Kaplan-Meier (KM), puis utilisent l'estimateur KM pour obtenir des courbes de survie presque sans biais.

De cette façon, nous pouvons obtenir des courbes de survie complètes à partir de *XGBoost*, et des intervalles de confiance avec des adaptations mineures (comme l'exécution de quelques tours de bootstrap).

***XGBSEStackedWeibull* : *XGBoost* comme estimateur de risque, *AFT Weibull* pour la courbe de survie**

Dans *XGBSEStackedWeibull*, nous effectuons l'empilement d'un modèle de survie *XGBoost* avec un modèle paramétrique *AFT Weibull*. Le modèle *XGBoost* s'ajuste aux données et prédit ensuite une valeur qui est interprétée comme une métrique de risque. Cette métrique de risque est introduite dans la régression de Weibull qui l'utilise comme seule variable indépendante.

Ainsi, nous pouvons bénéficier de la puissance de discrimination de *XGBoost* en même temps que de la rigueur statistique de *AFT Weibull* (par exemple, des courbes de survie calibrées).



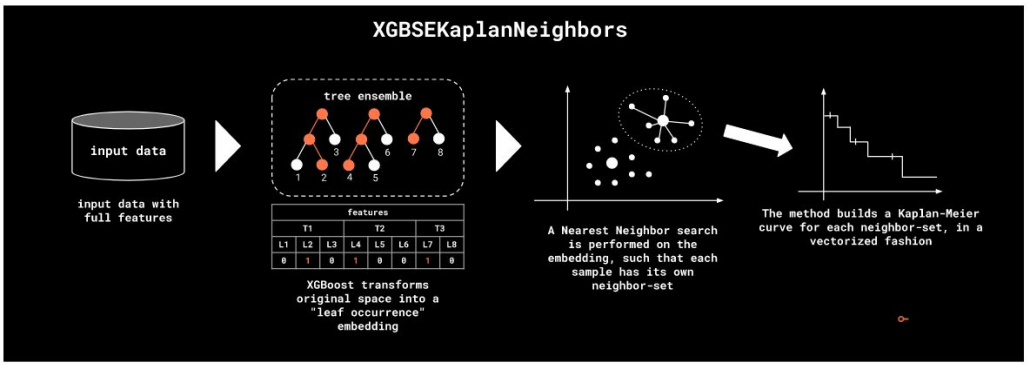
GRAPHIQUE 3.3 – XGBSEStackedWeibull (Vieira et al. (2021)).

Comme nous empilons *XGBoost* avec un modèle paramétrique à une variable (par opposition à *XGBSEDebiasedBCE*), le modèle peut être beaucoup plus rapide (notamment lors de l'apprentissage). Nous avons également de meilleures capacités d'extrapolation, en raison d'hypothèses plus fortes sur la forme de la courbe de survie.

Toutefois, ces hypothèses plus fortes peuvent ne pas convenir à certains ensembles de données aussi bien que d'autres méthodes.

***XGBSEKaplanNeighbors* : Kaplan-Meier sur les voisins les plus proches**

Comme expliqué dans la section précédente, même si l'embedding produit par *XGBoost* est sparse et de haute dimension, sa dimensionnalité intrinsèque devrait en fait être inférieure aux données d'entrée. Cela nous permet de "convertir" *XGBoost* en un modèle de plus proche voisin, où nous utilisons la distance de *Hamming* pour définir les éléments similaires comme étant ceux qui coïncident le plus aux nœuds terminaux de l'ensemble. Ensuite, à chaque ensemble de voisins, nous pouvons obtenir des estimations de survie avec des méthodes robustes telles que l'estimateur de *Kaplan-Meier*.

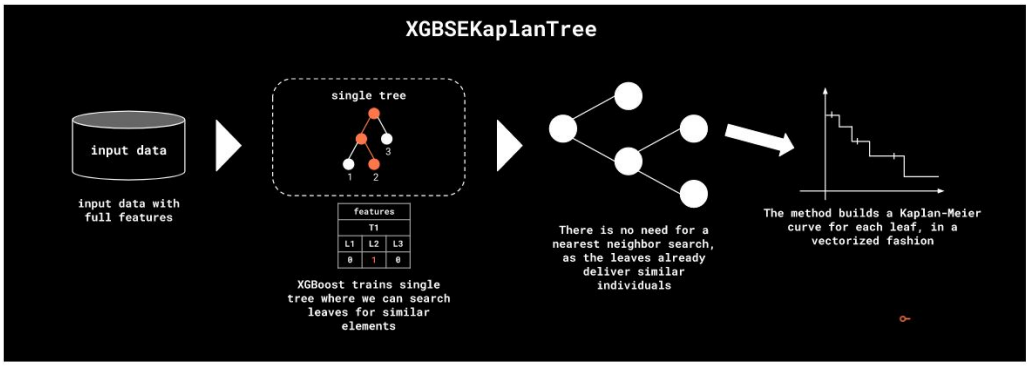


GRAPHIQUE 3.4 – XGBSEKaplanNeighbors (Vieira et al. (2021)).

Cependant, cette méthode peut s'avérer très coûteuse à l'échelle de centaines de milliers d'échantillons, en raison de la recherche du plus proche voisin, tant au niveau de l'entraînement (construction de l'index de recherche) que du scoring (recherche effective).

***XGBSEKaplanTree* : un seul arbre, et Kaplan-Meier sur ses feuilles**

Pour simplifier *XGBSEKaplanNeighbors*, nous fournissons également une implémentation à un seul arbre. Au lieu d'effectuer des recherches coûteuses de voisins les plus proches, nous ajustons un seul arbre via *XGBoost* et calculons les courbes *Kaplan-Meier* à chacune de ses feuilles.



GRAPHIQUE 3.5 – XGBSEKaplanTree (Vieira et al. (2021)).

C'est de loin l'implémentation la plus efficace, capable d'évoluer facilement vers des millions de données. Au moment de l'ajustement, l'arbre est construit et toutes les courbes *Kaplan-Meier* sont précalculées, de sorte qu'au moment du scoring, une simple requête suffira pour obtenir les estimations du modèle.

Cependant, comme nous ajustons un seul arbre, le pouvoir prédictif peut être moins bon. Cela pourrait être un compromis raisonnable, mais nous fournissons également *XGBSEBootstrapEstimator*, une abstraction bootstrap où nous pouvons adapter une forêt de *XGBSEKaplanTree* pour améliorer la précision et réduire la variance.

3.2 Critères de comparaison et outils d'interprétation

Dans la section ci-dessus, nous avons présenté plusieurs modèles (classiques et *Machine Learning*) pour l'estimation de la survie en assurance emprunteur. Ainsi, il nous faut définir des critères de comparaison pour apprécier la qualité de prédiction de ces modèles. Nous présentons ci-dessous deux indices utilisés en analyse de survie : *Concordance Index* et *Brier Score*. En outre, les algorithmes de *Machine Learning*, malgré leur succès, ressemblent à des boîtes noires à cause de leur manque d'interprétabilité. Ainsi, nous présentons aussi une méthode d'interprétation de ces modèles.

3.2.1 Concordance Index

L'indice de concordance ou $C-index$, a été introduit par Harrell et al. (1982). Cette métrique est utilisée en analyse de survie pour comparer la capacité prédictive des modèles.

Cette métrique nous permet de mesurer la capacité du modèle à ordonner les emprunteurs en fonction de leur survie. Elle est aussi très pertinente lorsque l'objectif principal du modèle est de classer les emprunteurs en fonction de leur risque de mortalité, c'est-à-dire de classer les emprunteurs de ceux qui ont la mortalité la plus faible à ceux qui ont la mortalité la plus élevée. Cependant, elle ne mesure que la capacité de classification du modèle mais pas la qualité de l'ajustement.

Le $C-index$ est défini comme une probabilité conditionnelle : les prédictions de survie du modèle $(S_i \ ; \ S_j)$ des deux emprunteurs i et j sont ordonnées de la même manière que leurs observations de survie respectives $(T_i \ ; \ T_j)$.

$$C-index = P(S_i \ < \ S_j \mid T_i \ < \ T_j)$$

S_i peut être considérée comme la durée de vie prédite par le modèle : $S_i = \mathbb{E}[T|X_i]$ ou la probabilité de survie jusqu'à la fin de la période d'étude $\hat{S}(\tau|X_i)$. Notons que, la définition classique est $C-index = P(S_i \ > \ S_j \mid T_i \ < \ T_j)$ car le score du modèle S_i est considéré. A noter que plus le score est élevé, plus la mortalité est élevée, le score et l'observation de la survie des paires doivent être classés dans des ordres opposés.

Comme pour l'AUC, $C-index = 1$ correspond à la meilleure prédiction du modèle, et $C-index = 0,5$ représente une prédiction aléatoire.

Contrairement à l'AUC, le $C-index$ prend en compte la présence de censure fréquemment rencontrée en analyse de survie. En effet, le $C-index$ ne peut être estimé que sur les paires d'observations $(i \ ; \ j)$ qui sont comparables car certaines paires ne le sont pas en présence de censure.

Si Ω désigne l'ensemble des paires comparables $(i \ ; \ j)$ où $(T_i \ < \ T_j)$, nous pouvons estimer le *C-index* de la manière suivante :

$$C-index = \frac{1}{Card(\Omega)} \sum_{(i,j)\in \ \Omega} \mathbb{1}_{M_i > M_j}$$

Une expression plus générale est proposée par Uno et al. (2011) pour permettre une comparaison précise du $C-index$ d'une étude à une autre. Ils ont proposé un estimateur libre de la

distribution de censure comme suit :

$$C - index = \frac{\sum_i \sum_i \Delta_i G(t_j)^{-2} \mathbb{1}_{t_i<t_j} \mathbb{1}_{M_i<M_j}}{\sum_i \sum_i \Delta_i G(t_j)^{-2} \mathbb{1}_{t_i<t_j}}$$

où $G(t)$ désigne la probabilité de ne pas avoir de censure jusqu'au moment t, et $j = 1$ si aucune censure, 0 sinon.

3.2.2 Brier Score

Initialement, le *Brier Score* a été introduit par Brier (1950) pour mesurer la précision des prévisions météorologiques. Ensuite, Graf et al. (1999) ont proposé d'utiliser cette métrique dans le domaine de la biostatistique pour évaluer la performance des modèles de survie. Elle est souvent utilisée pour la comparaison de plusieurs modèles.

Le *Brier Score*, noté BS, est défini comme la moyenne de la différence au carré entre les probabilités de survie prédites et observées à un instant t donné.

$$BS(t) = \frac{1}{N} \sum_i (\mathbb{1}_{T_i>t} - \hat{S}(t|X_i))^2$$

avec $\hat{S}(t|X_i)$ la probabilité de survie prédite par le modèle.

Le calcul du BS doit aussi prendre en compte la présence de censure dans les données. Ainsi, la formule ci-dessus doit être corrigée pour une estimation précise de cette métrique. En effet, si une censure a eu lieu avant le temps fixe t, nous ne pouvons pas savoir si l'emprunteur a survécu plus longtemps que t. La formule ci-dessous est proposée par les auteurs comme dans le cas du $C - index$ pour estimer la censure :

$$BS(t) = \frac{1}{N} \sum_i \frac{\hat{S}(t|X_i)^2}{G(t_i)} \mathbb{1}_{\{t_i \leq t \ ; \ d\delta_i=1\}} + \frac{(1 - \hat{S}(t|X_i))^2}{G(t)} \mathbb{1}_{\{t_i>t\}}$$

où $G(t)$ est la probabilité de ne pas observer de censure jusqu'au moment t.

3.2.3 Effets marginaux et interactions : *SHAP*

Lundberg and Lee (2017) dans leur papier "A unified approach to interpreting model predictions" ont proposé les valeurs *SHAP* (SHapley Additive exPlanations) qui offrent un haut niveau d'interprétabilité pour un modèle notamment les modèles de *Machine Learning*. Les valeurs SHAP offrent une *Interprétabilité globale*. En effet, elles permettent de montrer dans quelle mesure chaque caractéristique de l'emprunteur ou du prêt contribue, positivement ou

négativement, à l'explication de la durée passée en vie dans le prêt. C'est comme le graphique d'importance des variables, mais il est capable de montrer la relation positive ou négative de chaque variable avec le risque de mortalité. Ces effets marginaux peuvent être obtenus avec la méthode Tree SHAP qui fournit des scores en utilisant tous les ordres possibles de représentation des arbres.

Il est aussi possible d'avoir recours aux graphiques PDP (Partial Dependance Plot) de Friedman (2000) qui permettent aussi de caractériser les relations entre les variables de l'emprunteur/prêt et le risque de mortalité dans le prêt. Ces graphiques représentent la moyenne des courbes individuelles pour toutes les valeurs possibles de la variable. Ainsi, ils s'identifient à l'effet marginal moyen de la variable.

3.3 Application à notre portefeuille

Pour un contrat de prêt détenu t années, il serait intéressant d'estimer la probabilité pour un emprunteur de décéder au cours de l'année t. Cette quantité est dénommée en modèles de durée "taux de risque instantané". Il s'agit de la probabilité qu'un emprunteur durant sa durée de prêt décède au cours d'une année donnée. Une manière d'estimer ce taux est donnée dans la démarche de construction de l'estimateur de *Cox* qui lui s'intéresse à une probabilité de survie. D'un autre côté, certaines méthodes de *Machine Learning* sont adaptées pour estimer la probabilité de survie de *Cox*. Nous présentons ci-dessous les résultats des modèles appliqués sur le portefeuille des prêts terminés.

3.3.1 Méthodes classiques : Modèle de *Cox*

Dans cette partie, nous appliquons à nos données le modèle de *Cox*. Nous commençons par le choix des variables pertinentes pour l'étude de la mortalité, puis la vérification des hypothèses de *Cox*.

Pour le choix des variables explicatives, nous avons utilisé une méthode de sélection automatique avec comme critère de choix l'AIC.

Il s'en suit la vérification des hypothèses du modèle notamment l'hypothèse des risques proportionnels. On peut la vérifier en effectuant un test d'indépendance entre la variable temps et toutes ou chacune des variables explicatives du modèle. Ce test est appelé le test des résidus de *Schoenfeld*. Une p-value inférieure à 5% indique que l'hypothèse n'est pas vérifiée. L'hypothèse de risques proportionnels est donc vérifiée si pour chacune des variables explicatives ainsi que pour l'ensemble du modèle nous avons une p-value $> 5\%$.

Le tableau 3.1 donne les résultats du test de l'hypothèse de proportionnalité du modèle de *Cox*. Il ressort de l'analyse de ce tableau que la p-value est inférieure à 5% pour presque chaque variable prise individuellement et globalement. Ainsi, les données rejettent l'hypothèse selon laquelle les risques sont proportionnels.

Variables	P-value
MIP_AGE_ASSU_DT_SCRP	$0,04$
SEXE_Homme	$< 0,005$
MIP_DUR_TOTALE_PRET	$< 0,005$
MIP_CI_ASSU_GLOB	$< 0,005$
MIP_QUOTITE_GLOBALE	$0,09$
MIP_TX_EMPRUNT	$< 0,005$
MIP_PLUSIEURS_PRETS_OUI	$< 0,005$
MIP_PLUSIEURS_ASSURES_OUI	$0,92$
MIP_TYPE_GARANTIE_DC+IT	$0,44$
MIP_REMBOURSEMENT_ANTICIPE_OUI	$< 0,005$
MIP_NAT_PRET_IMMO	$< 0,005$
GLOBAL	$< 0,005$

TABLEAU 3.1 – Vérification de l'hypothèse de proportionnalité

Nous avons ensuite essayé de stratifier suivant l'âge à la souscription du prêt mais l'hypothèse reste toujours non vérifiée pour nos données.

3.3.2 Machine Learning : *XGBoost Cox* et *Xgbse*

Certaines méthodes de *Machine Learning* sont maintenant adaptées pour la modélisation de durée notamment le *XGBoost*. Le package *Xgbse* (Xgboost survival embedding) de python est une alternative du modèle de régression des risques proportionnels de *Cox*. Il combine la méthode *XGBoost* avec un ensemble de méthodes classiques telles que Kaplan Meier, les k plus proches voisins et la régression logistique pour estimer la probabilité de survie.

La calibration de ces modèles commence par la recherche des valeurs optimales des hyperparamètres. Ensuite, nous avons estimé les probabilités de survie pour chacun de ces modèles sur l'échantillon d'apprentissage.

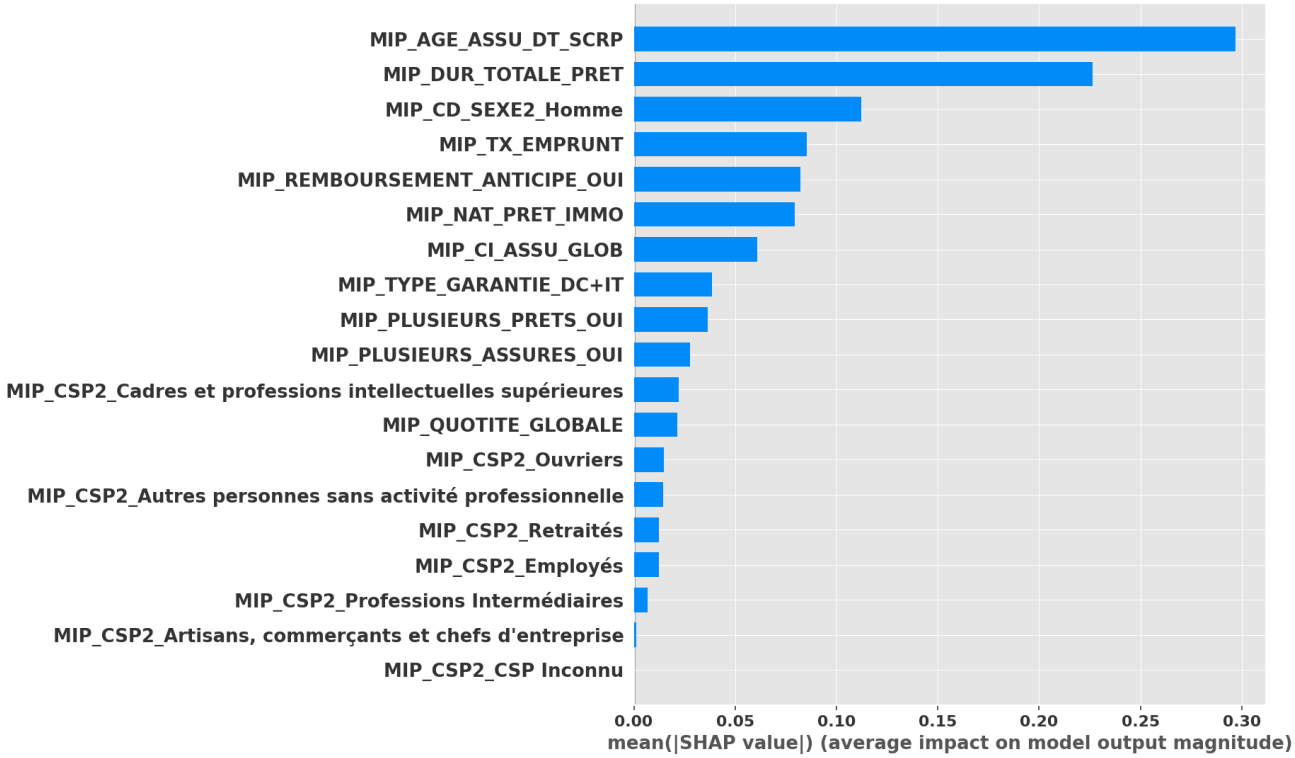
Afin de choisir le modèle qui calibre le mieux nos données de survie, nous avons évalué la performance de chacun d'entre eux en utilisant notre échantillon de validation original. Les critères de choix utilisés sont le *C-index* et le *Brier Score*. Les résultats de ces indicateurs sont renseignés dans le tableau 3.2 ci-dessous. Il ressort de l'analyse de ces deux indicateurs que le modèle *XGBoost* classique performe plus que ses versions adaptées (XGBSEDebiasedBCE, XGBSEKaplanNeighbors et XGBSEKaplanTree). Cependant, les différences de performance

ne sont pas très significatives. Ces versions adaptées de *XGBoost* ont l'avantage de fournir une estimation non biaisée des probabilités de survie. Elles permettent aussi d'avoir une courbe de survie à la place d'une estimation ponctuelle comme le ferait le *XGBoost*.

Modèles	C-index	Brier Score
XGBoost	0,766	–
XGBSEDebiasedBCE	0,754	0,127
XGBSEKaplanNeighbors	0,740	0,136
XGBSEKaplanTree	0,732	0,154

TABLEAU 3.2 – Métriques de performance pour les mêmes variables explicatives

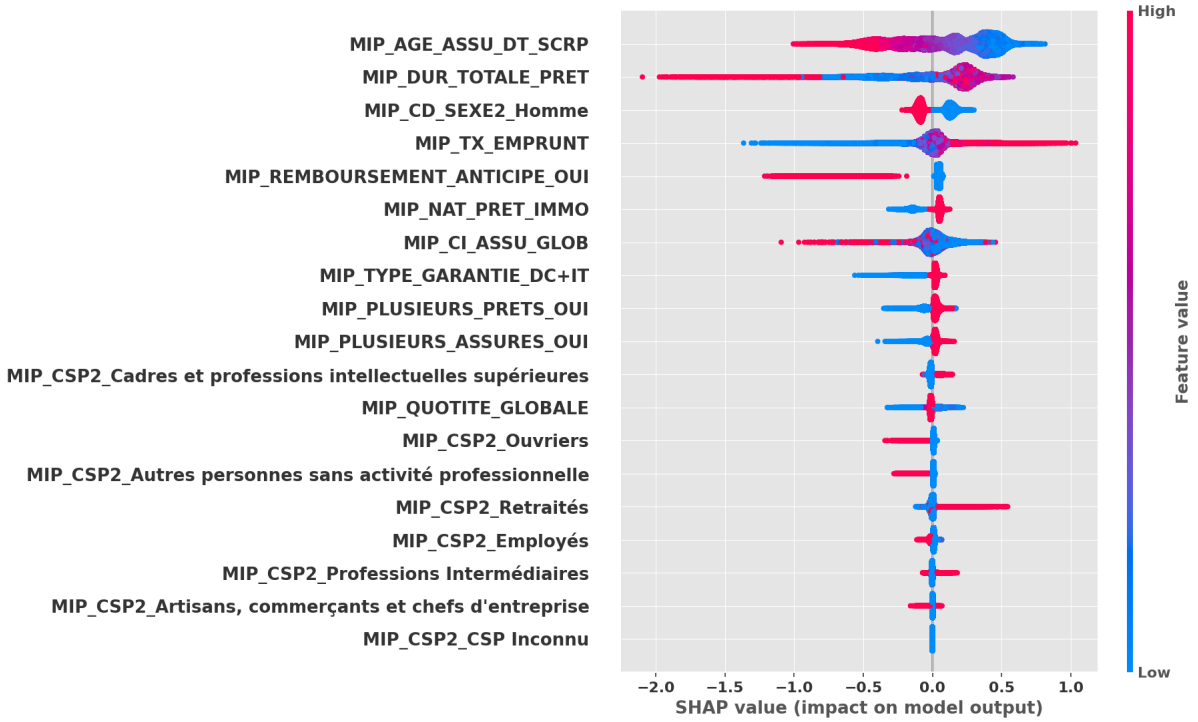
Au regard de la figure 3.6, les variables les plus importantes pour le modèle *XGBoost* en ce qui concerne l'explication de la durée passée en vie dans le prêt sont : l'âge de l'assuré à la souscription du prêt, la durée du prêt et le sexe.



GRAPHIQUE 3.6 – Shap-Importance des variables dans le modèle *XGBoost*

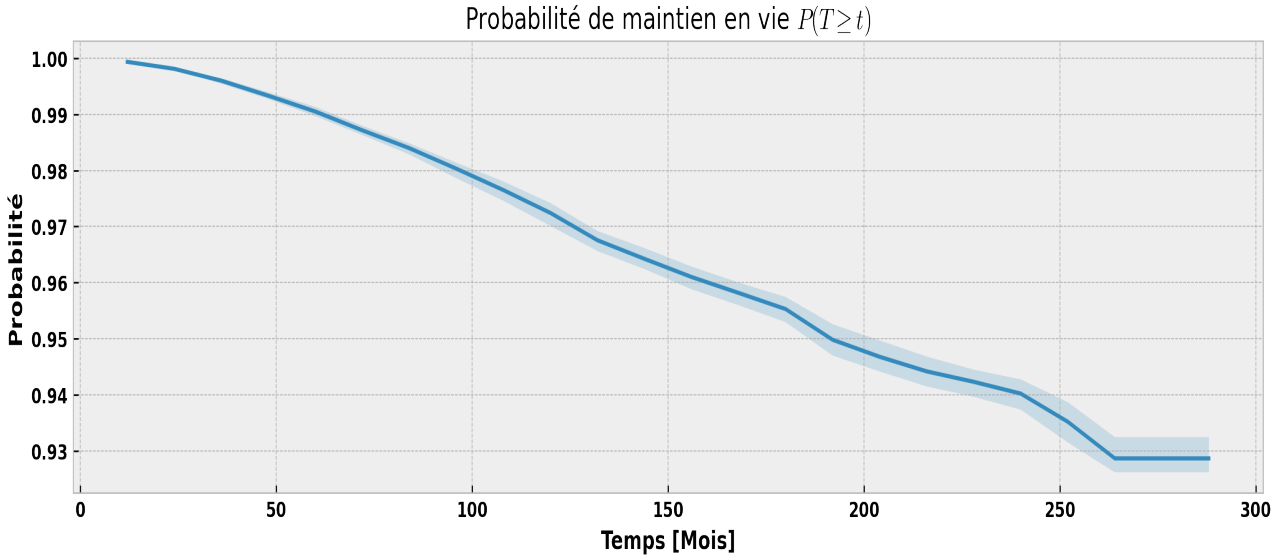
D'un autre côté, l'analyse des effets marginaux de la figure 3.7 laisse entrevoir que plus l'âge à la souscription du prêt est élevé moins la durée passée en vie dans le prêt est importante. En outre, l'effet marginal sur la variable sexe informe que le risque de mortalité est plus élevé chez les hommes que chez les femmes (modalité de référence). Aussi, les prêts portant sur des taux d'emprunts élevés enregistrent une durée passée en vie dans le prêt plus faible que ceux dont le taux d'emprunt est bas. Enfin, l'effet marginal de la nature du prêt laisse entrevoir que

la durée passée en vie est plus accentuée chez les prêts destinés à des achats immobiliers par rapport à ceux destinés à des achats de consommation (référence).



GRAPHIQUE 3.7 – Effets marginaux des variables explicatives sur la durée passée en vie dans le prêt

Pour terminer, comme le modèle *XGBoost* classique ne permet pas d'avoir directement une courbe de survie non biaisée, nous avons utilisé le meilleur modèle de la famille des *Xgbse* (*XGBSEDebiasedBCE*) pour construire notre table de maintien en vie dans le prêt.



GRAPHIQUE 3.8 – Table de maintien en vie dans le prêt via *XGBSEDebiasedBCE*

4) Modélisation de la garantie IT

Le chapitre précédent a permis de déterminer le coût individuel pour la couverture contre le risque DC en mettant en place une table d'expérience de maintien en vie dans le prêt.

Ce chapitre a pour but de déterminer le coût individuel pour la garantie IT en modélisant les prestations versées sur la durée du prêt. De la même manière pour le DC, nous allons d'abord présenter les méthodes classiquement utilisées et leurs limites. Puis dans un second temps, nous décrirons l'apport que pourrait avoir des méthodes alternatives d'apprentissage dites *Machine Learning*.

4.1 Estimation des prestations en IT

Dans cette partie, nous proposons une méthodologie d'estimation des prestations versées dans le cadre de la garantie IT pour les emprunteurs qui ont pris la couverture "DC + IT". La littérature actuarielle est à la base du choix de ces méthodes. Les méthodes classiques généralement utilisées à cet effet sont basées sur les *modèles linéaires généralisés*. En ce qui concerne les méthodes *Machine Learning*, nous avons le *Gradient Boosting*, *XGBoost* ou les *Réseaux de neurones*.

4.1.1 Les méthodes classiques : *Tweedie*

La méthode standard utilisée pour modéliser le coût des sinistres est de calibrer une loi Gamma et pour la fréquence des sinistres on utilise souvent une loi de Poisson ou une loi Binomiale Négative. Cependant, le calibrage de la loi Gamma nécessite de ne retenir que les montants de prestation strictement positifs car son support ne contient pas 0. En d'autres termes, les emprunteurs n'ayant pas subi l'IT doivent être exclus de la modélisation des coûts. Pour éviter la modélisation en deux étapes (fréquence et coût), nous choisissons parmi les méthodes classiques la distribution de *Tweedie* pour modéliser les prestations en IT.

Notons par B_k le montant des prestations pour le $k^{ième}$ arrêt travail et N le nombre de sinistres IT durant la durée du prêt.

La distribution de *Tweedie* correspond à la somme composée suivante

$$X = \begin{cases} \sum_{k=1}^{N} B_k & \text{si } N > 0 \\ 0 & \text{si } N = 0 \end{cases}$$

où N suit une loi de Poisson $\mathcal{P}(\lambda)$ et B_k suit une loi gamma $\Gamma(\alpha, \beta)$. La masse de probabilité en 0 vaut $e^{-\lambda} = P(N = 0)$. De plus, on connait son espérance et sa variance

$$E[X] = \frac{\lambda\alpha}{\beta}, \text{Var}[X] = \frac{\lambda\alpha}{\beta^2}(1+\alpha).$$

Cette distribution est paramétrisée comme suit : $\mathcal{T}w(\mu, \phi, p)$ où μ est la moyenne de la distribution, ϕ est le paramètre de dispersion, $\phi > 0$, et p est l'indice de la distribution.
Sous cette forme, nous avons $E[X] = \mu, \text{Var}[X] = \phi\mu^p \Rightarrow$ la déviation $V(\mu) = \mu^p$.
La loi de *Tweedie* appartient à la famille exponentielle. En effet pour $1 < p < 2$, l'identification est obtenue en choisissant

$$\theta = -\frac{1}{(p-1)\mu^{p-1}}, a(x) = x, b(x) = -\frac{1}{x^{\frac{2-p}{p-1}}(p-2)(p-1)^{\frac{2-p}{p-1}}}$$

et

$$c(x, \phi) = \mathbb{1}_{x=0} + \mathbb{1}_{x>0} \ln \left(\frac{\tilde{e}\left(\frac{x^{\frac{2-p}{p-1}}}{\phi(2-p)(p-1)^{\frac{2-p}{p-1}}} \right) - 1}{x} \right).$$

Ainsi, la loi de *Tweedie* a une double caractéristique : faire partie de la famille exponentielle, être une loi mixte avec une masse en zéro et une partie continue au-delà. Donc un modèle de régression *Tweedie* permet de modéliser conjointement la fréquence et la sévérité des sinistres.

Notons y_i le coût total de le couverture IT pour un emprunteur i, i.e. y_i vaut zéro pour un emprunteur n'ayant pas subit l'IT durant son prêt et sinon la somme des prestations pour un emprunteur ayant au moins un sinistre IT. On note par x_i le vecteur des variables explicatives de l'emprunteur i.
On suppose $Y_i \sim \mathcal{T}w(\mu_i, \phi, p)$ pour des variables indépendantes et une forme paramétrique des variables de régression

$$\mu_i = h(< \boldsymbol{x}_i, \beta >) = h(\eta_i)$$

La log-vraisemblance s'écrit donc pour $1 < p < 2$

$$\ln \mathcal{L}(\beta) = \sum_{i=1}^{n} \frac{y_i \theta(\mu_i, p) - b(\theta(\mu_i, p))}{\phi} + \sum_{i=1}^{n} c(y_i, \phi, p) = \sum_{i=1}^{n} \frac{y_i \frac{\mu_i^{1-p}}{1-p} - \frac{\mu_i^{2-p}}{2-p}}{\phi} + \sum_{i=1}^{n} c(y_i, \phi, p)$$

où c est le terme indépendant de μ_i. La dérivée partielle vaut

$$\frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta_k} = \frac{1}{\phi}\sum_{i=1}^{n}\left(y_i x_i^{(k)} h'(\eta_i)\, h(\eta_i)^{-p} - x_i^{(k)} h'(\eta_i)\, h(\eta_i)^{1-p}\right) = \frac{1}{\phi}\sum_{i=1}^{n}\frac{x_i^{(k)}}{\mu_i^p} h'(\eta_i)\,(y_i - \mu_i)$$

Les équations du score qui en découlent sont utilisées par les algorithmes d'optimisation numériques pour l'estimation des paramètres.
Les fonctions liens usuelles pour la régression *Tweedie* sont des fonctions de $\mathbb{R}_+$vers $\mathbb{R}$. Le choix le plus simple est le lien logarithme, on peut choisir un lien puissance.

Enfin, le choix du paramètre p de la loi de *Tweedie* est important. En effet, ce paramètre permet de déterminer les sous-familles de cette distribution. Ainsi, une recherche optimale de cette valeur pour nos données sera effectuée à partir d'un ensemble de valeurs définis a priori.

4.1.2 La régression *XGBoost*

Les développements récents du modèle *XGBoost* permettent de faire un embedding de ce modèle avec la distribution de *Tweedie*. Ces avancés rendent possible l'utilisation de *XGBoost* pour modéliser les prestations en assurance. Nous allons explorer ce modèle pour la prédiction des prestations en arrêt de travail dans les contrats d'assurance emprunteur.

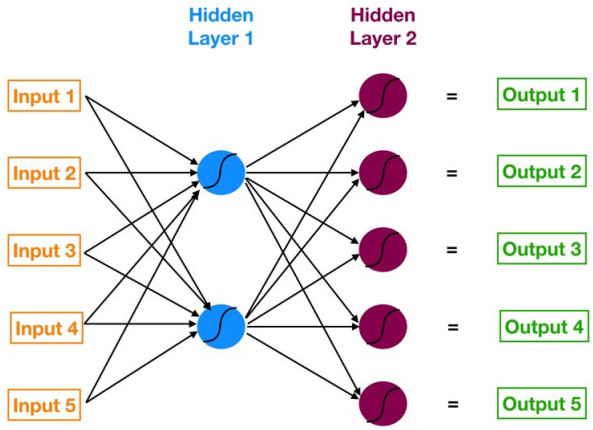
4.1.3 Les *réseaux de neurones*

L'apprentissage profond est de plus en plus évoqué pour des sujets traitant des phénomènes complexes. Les récentes victoires de l'apprentissage profond en reconnaissance d'images et autres ouvrent la voie à son utilisation en actuariat. Les *réseaux de neurones* sont les moteurs de l'apprentissage profond. Et même s'ils ressemblent à des boîtes noires, au fond, ils essaient d'accomplir la même chose que n'importe quel autre modèle : faire de bonnes prédictions.

Dans cette partie, nous allons présenter le principe de fonctionnement des *réseaux de neurones*.

On distingue les réseaux monocouches et les réseaux multicouches (Multi Layer Perceptron). Dans le premier cas, nous avons deux couches : une en entrée et une en sortie qui sont directement liées entre elles. En ce qui concerne le deuxième cas, des couches cachées sont ajoutées au premier cas pour augmenter la précision des estimations. Nous représentons ci-dessous le diagramme d'un *réseau de neurone* simple avec cinq entrées, cinq sorties et deux couches cachées de neurones.

GRAPHIQUE 4.1 – Principe de fonctionnement du Multi Layer Perceptron



En commençant par la gauche, nous avons : la couche d'entrée de notre modèle en orange, notre première couche cachée de neurones en bleu, notre deuxième couche cachée de neurones en magenta et la couche de sortie (la prédiction) de notre modèle en vert.
Les flèches qui relient les nœuds montrent comment tous les neurones sont interconnectés et comment les données passent de la couche d'entrée à la couche de sortie.
Ce passage se fait à l'aide d'une fonction dite d'activation. Si on note par y la valeur de sortie, g la fonction d'activation, x_1,...,x_n les valeurs d'entrée et ω_0,...,ω_n les poids associés à chaque neurone et dont les valeurs sont estimées dans la phase d'apprentissage. On a :

$$y = h(x_1, ..., x_p) = g(w_0 + \sum_{j=1}^{p} w_j x_j) = g(\alpha_0 + \alpha' x)$$

Les différents types de neurones se distinguent par la nature g de leur fonction d'activation. Les principaux types sont :

- linéaire g est la fonction identité,
- Elu (Exponential Linear Unit) qui avec $\alpha > 0$ donne $g(x) = x \ si \ x > 0$ et $g(x) = \alpha \times (\exp(x) - 1) \ si \ x < 0$,
- sigmoïde $g(x) = \frac{1}{1+\exp(x)}$,
- ReLU $g(x) = max(0, x)$ (rectified linear unit),
- softmax $g(x)_j = \frac{\exp(x_j)}{\sum_{k=1}^{K} \exp(x_k)}$ pour tout $k \in \{1, ..., K\}$,
- ...

Les modèles linéaires, sigmoïdaux, ReLU, Elu, softmax sont bien adaptés aux algorithmes d'apprentissage car leur fonction d'activation est différentiable ; ce sont les plus utilisés.

Pour l'estimation des coûts des prestations IT, nous devons imposer la condition de positivité dans la couche de sortie. La fonction d'activation *Elu* est utilisée à cet effet où la valeur 1 est affectée dans le cas où les entrées aux couches de sorties sont négatives. Cette fonction a aussi pour avantage de réduire fortement la volatilité au niveau de la descent de gradient.

Une fois la fonction d'activation choisie, il s'en suit le choix de la fonction de perte. En effet, pour avoir une prédiction optimale, on minimise une fonction dite de perte lors de l'apprentissage. Le choix de cette fonction dépend de la nature de la variable d'intérêt à modéliser. Pour des problèmes de régression, qui est notre cas, la fonction de perte la plus utilisée est l'erreur quadratique moyenne (Mean Squared Error) en supposant une distribution normale des données.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Enfin, il est reconnu de manière unanime dans la littérature sur le sujet que les résultats des réseaux de neurones dépendent (très) fortement du choix des hyperparamètres.
Dans les logiciels comme Python où ces algorithmes sont implémentés, il est donné une option *GridSearchCV* pour le choix optimal des hyperparamètres.
Les paramètres les plus déterminants lors de l'apprentissage des réseaux de neurones sont :

- Le choix de l'algorithme d'optimisation et le taux d'apprentissage pour la convergence. Le *stochastic gradient descent* et le *RMSprop* sont généralement utilisés à cet effet ;
- Choix du nombre de cycles d'apprentissage (*epochs*) ;
- Le nombre d'observations utilisé à chaque étape d'apprentissage (*batch*) ;
- Le "*droptout*" c'est à dire le paramètre de décrochage pour éviter le "sur-apprentissage" ;
- Les taux de pénalisations : *Ridge* ou *Lasso*.

4.2 Critère de comparaison avec le *Gini-Index*

Dans la section précédente, trois modèles ont été retenus pour l'estimation des prestations en IT. Ainsi, il nous faut définir un critère de comparaison pour jauger de la qualité de prédiction de ces modèles et choisir le "meilleur" d'entre eux. Pour ce faire, l'indice de Gini a été utilisé comme critère de choix. Il est souvent utilisé en actuariat pour la mesure de la qualité de prédiction de la prime pure. L'approche a été proposée par Frees et al. (2014) et est basée sur la courbe de *Lorenz*. Cette dernière a été développée par Max O. Lorenz en vue d'une représentation graphique des inégalités de revenu. L'indice de Gini développé par le statisticien

italien Corrado Gini en 1912 est une mesure statistique permettant de résumer l'information contenue dans cette courbe. Nous retrouvons en abscisse de cette courbe la part cumulée de la population et en ordonnée la part cumulée des revenus. La répartition parfaite des ressources se trouve à la diagonale de la courbe et donc tout écart à cette diagonale représente une situation d'inégalité. Cette approche consiste à définir le coefficient de Gini comme le double de l'aire comprise entre la courbe et la diagonale.

Dans le domaine de l'assurance, Frees et al. (2014) ont proposé une version adaptée de cet indice en intégrant le concept de relativité car les montants modélisés sont généralement asymétriques. Ainsi, la métrique obtenue permet de comparer les prédictions et de choisir la plus adéquate.

Notons par y_i le coût des prestations de l'emprunteur i, $i = 1, ..., N$ et $X_1, ..., X_K$ les caractéristiques de l'emprunteur et du contrat de prêt. Nous cherchons à comparer les prédictions de deux méthodes d'apprentissage P_i et S_i. Nous avons dans ce cas le coût relatif défini comme suit :

$$R(x_i) = \frac{S(x_i)}{x_i}$$

Cet indice de relativité permet de construire la courbe de Lorenz ordonnée à partir des fonctions de répartition suivantes :

$$\hat{F}_y(s) = \frac{\sum_{i=1}^{N} y_i \mathbb{1}_{R_i \leqslant s}}{\sum_{i=1}^{N} y_i} \text{ et } \hat{F}_P(s) = \frac{\sum_{i=1}^{N} P_i \mathbb{1}_{R_i \leq s}}{\sum_{i=1}^{N} P_i}$$

L'indice de Gini, représentant l'aire sous la courbe de Lorenz, est estimé par :

$$\widehat{\text{Gin}\, i} \approx \frac{1}{N} \sum_{i=1}^{N} \left(\hat{F}_P\left(R_i\right) - \hat{F}_L\left(R_i\right) \right)$$

Lorsque $F_P(s) - F_L(s) \geq 0$ alors cette tarification est jugée profitable à l'entreprise car les primes sont supérieurs aux pertes. Dans une situation d'équilibre, cela signifie qu'il existe une sous population sur-tarifiée et une autre sous-tarifiée. L'indice permet ainsi de voir à quel point l'assureur parvient à discriminer efficacement sa population en séparant les populations hautement risquées et celles à risque faible. En termes de choix de modèle il sera choisi celui qui parvient le mieux à réaliser cette distinction. Le coefficient de Gini peut également être approximé en terme de covariance et dans ce cas on a Gini $\approx \frac{2}{N} \widehat{\text{Cov}}(PP, \text{rang}(R))$ où PP représente la perte normalisée (c'est à dire divisée par la moyenne) associée à une prime (y/P). Les détails du calcul aboutissant à cette approximation peuvent être retrouvés dans Frees et al.

(2014).

4.3 Application à notre portefeuille

Pour la modélisation du montant des prestations en IT, nous avons utilisé les prêts terminés afin d'éviter des phénomènes de censures sur les montants versés. Nous avons ensuite découpé le portefeuille en apprentissage (75%) et validation (25%). Nous avons testé trois catégories de modèles : les modèles GLM (Tweedie), les modèles d'agrégation (*XGBoost*) et l'apprentissage profond (Réseaux de neurones). Le choix de ces trois catégories de modèles nous permet de capter l'apport des méthodes de *Machine Learning* dans l'analyse multivariée de la sinistralité en IT des contrats emprunteurs. En effet, même si les modèles GLM offrent une meilleure explicabilité des résultats, les méthodes de *Machine Learning* nous permettent d'avoir une meilleure capacité de prédiction. De plus, certains outils d'aide à l'interprétation des méthodes de *Machine Learning* commencent à faire leur apparition. Il s'agit notamment du module SHAP qui permet de représenter l'importance, les effets marginaux et les interactions des variables.

Nous avons utilisé l'échantillon de validation pour choisir le "meilleur" modèle. Pour ce faire, nous avons effectué des prédictions du montant des prestations de l'échantillon de validation avec les modèles calibrés sur l'échantillon d'apprentissage. Ainsi, nous avons pu calculer deux métriques de performances le Root Mean Square Error (RMSE) et l'Indice de Gini. Le tableau 4.1 présente les deux indicateurs pour les différents modèles.

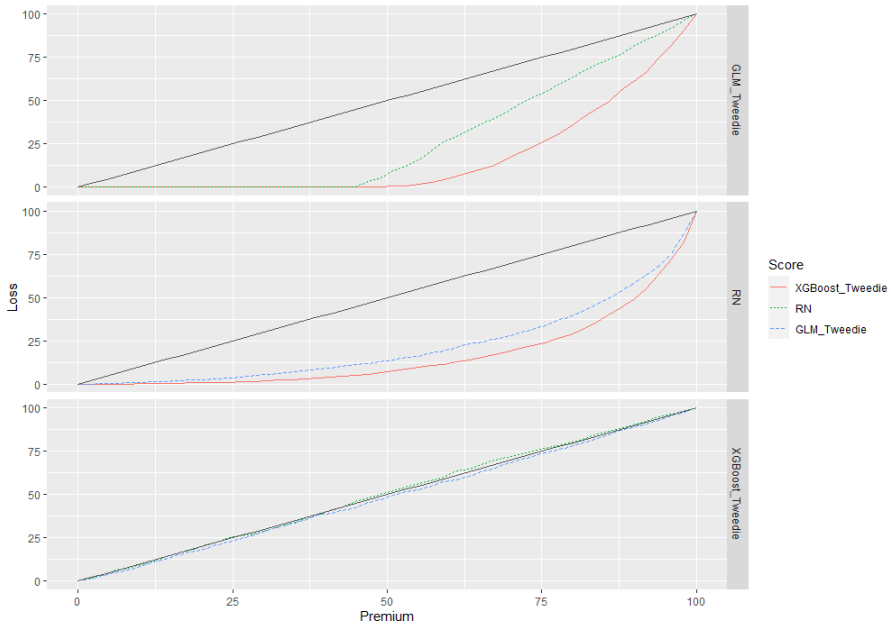
	GLM Tweedie	XGBoost	Réseaux de neurones
Indice de Gini	0,671	0,7495	0,578
RMSE	20627,271	1045,703	1052,454

TABLEAU 4.1 – Métriques de performance pour les mêmes variables explicatives

En se référant à l'indice de Gini et le RMSE, nous obtenons de meilleure performance avec le modèle *XGBoost*. Nous notons que les réseaux de neurones ont un indice de Gini moins importants par rapport aux autres modèles. Une explication possible de ce résultat est la difficulté de paramétrisation de ce dernier même si nous avons effectué une optimisation de ses paramètres. Nous choisissons donc le modèle *XGBoost* pour le calcul des provisions pour sinistres à payer des prestations en IT.

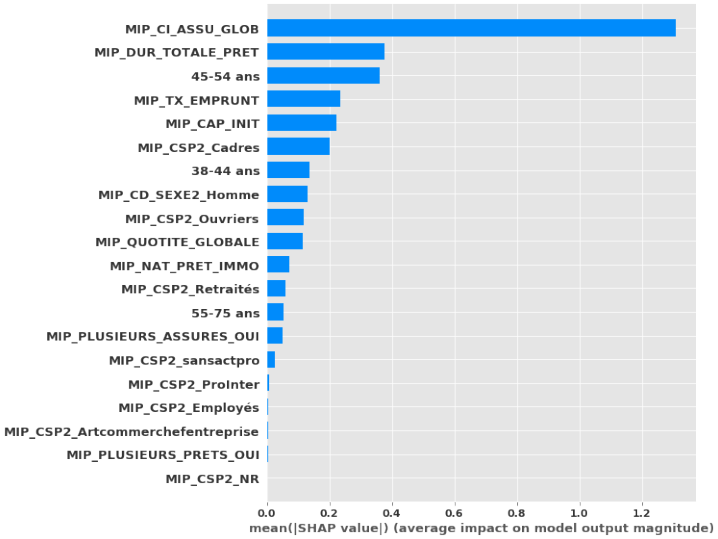
D'un autre côté, nous avons représenté les courbes de *Lorenz* ordonnées (voir 4.2 ci-dessous) pour chacun de nos trois modèles. Ces graphiques donnent entre autres une représentation visuelle du pouvoir de discrimination du portefeuille par ces modèles. Comme indiqué ci-dessus,

le modèle *XGBoost* donne la plus grande aire sous la diagonale et donc discrimine le plus le portefeuille.



GRAPHIQUE 4.2 – Courbes de Lorenz ordonnées par modèles sur les prestations en IT

En outre, nous avons aussi analysé l'importance des variables dans la modélisation ainsi que les effets d'interactions entre les variables explicatives du modèle *XGBoost*. Ainsi, nous avons représenté dans la figure 4.3 ci-dessous le Shap-Importance des variables dans le modèle *XGBoost*. Il en ressort de l'analyse de la Shap-importance que le capital assuré "MIP_CI_ASSU_GLOB" est la variable la plus importante dans la prédiction des prestations en IT du modèle *XGBoost*. Il s'en suit la durée totale du prêt, la tranche d'âge à la souscription [45 ; 55 ans), le taux d'emprunt, la somme initiale, les cadres, la tranche d'âge [38 ; 45 ans) et le sexe.



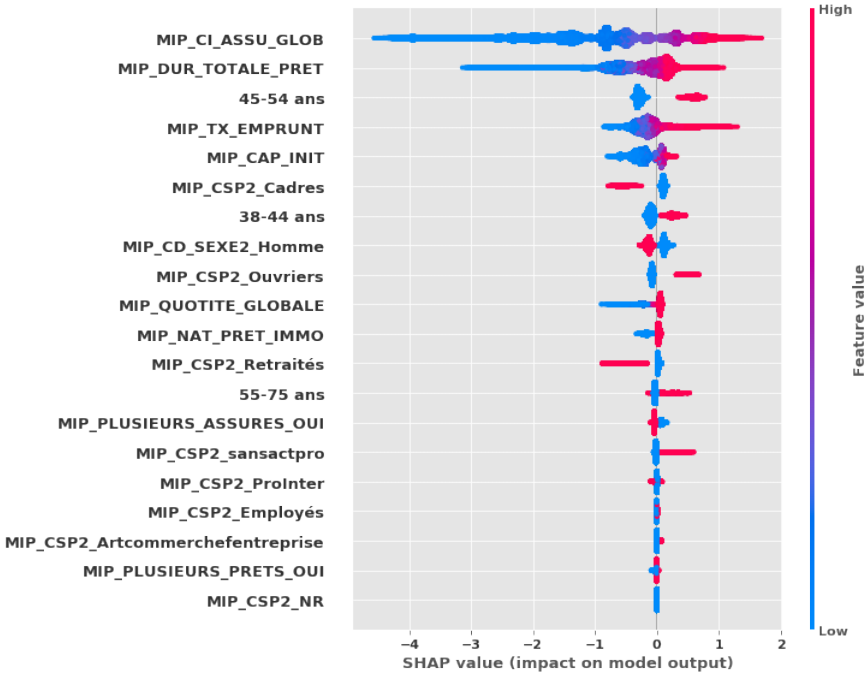
GRAPHIQUE 4.3 – Shap-Importance des variables dans le modèle *XGBoost*

Le graphique 4.4 illustre les effets marginaux des variables explicatives sur le montant des prestations en IT des emprunteurs. Sur ce graphique, on comprend donc que plus le capital assuré est élevé, plus le montant des prestations en IT l'est aussi. Cette même remarque est observée pour le capital initial emprunté. D'un autre côté, les emprunteurs qui choisissent une quotité globale faible ont des montants de prestations moins importants.

En ce qui concerne les effets de l'âge à la souscription du prêt sur le montant des prestations en IT, nous constatons que les emprunteurs appartenant à la tranche d'âge à la souscription [45 ; 55 ans) ou [38 ; 45 ans) ont des montants de prestations en IT plus importants que ceux de la tranche [18 ; 38 ans) (modalité de référence). Cependant les emprunteurs de plus de 55 ans ne présentent pas un effet interprétable sur la prestation. Nous remarquons aussi que les hommes ont des montants de prestations en IT plus faibles que les femmes.

En ce qui concerne la catégorie socio-professionnelle, nous observons que les emprunteurs cadres et les retraités ont des montants de prestations plus faibles par rapport à la modalité de référence (emprunteurs agriculteurs). D'un autre côté, les emprunteurs ouvriers ont des montants de prestations plus élevés que les agriculteurs.

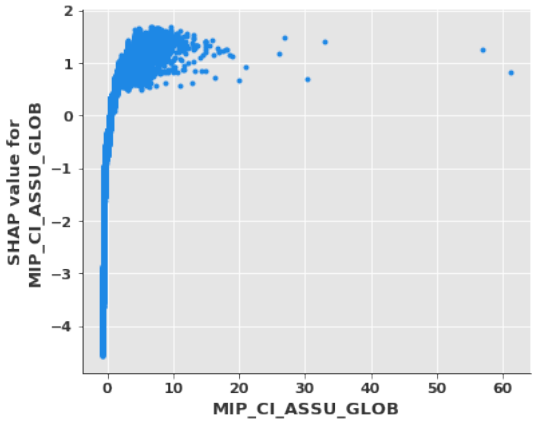
Pour ce qui est de la durée totale du prêt, nous observons qu'elle impacte positivement le montant des prestations : plus elle est importante, plus le montant des prestations l'est aussi.



GRAPHIQUE 4.4 – Effets marginaux des variables explicatives sur le montant des prestations

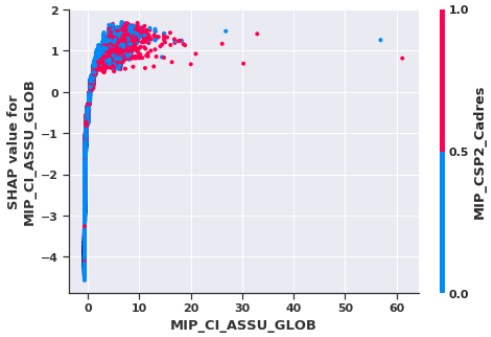
En plus de l'analyse groupée des effets marginaux, nous pouvons aussi avoir une analyse indivi-

duelle de ces derniers à l'aide des "shape values" et des graphiques pdp. Nous avons représenté dans le graphique 4.5 l'effet marginal de la somme assurée sur le montant des prestations. L'ordonnée de ce graphe peut être interprété comme l'effet marginal moyen de cette variable. En d'autres termes, le graphique fournit la variation qu'on pourrait observer sur le montant des prestations à la suite d'une augmentation unitaire du capital assuré. Nous retrouvons la même remarque que précédemment selon laquelle plus la somme assurée est importante plus le montant des prestations en IT est élevé.

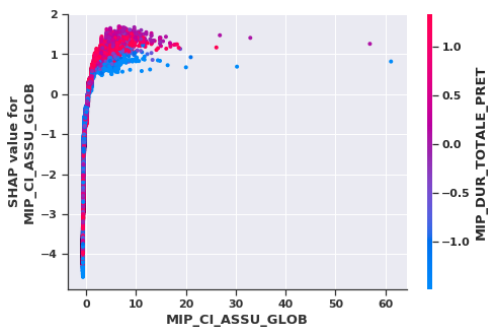


GRAPHIQUE 4.5 – Effet marginal du capital assuré sur le montant des prestations

Pour analyser les possibles interactions avec la somme assurée, nous avons représenté les variables qui interagissent plus avec elle (voir graphique 4.6). L'analyse de ce dernier laisse entrevoir que les prêts qui mobilisent des capitaux faibles ne contiennent presque pas de cadres. D'un autre côté, pour les prêts avec des sommes assurées importantes, les prestations en IT les plus faibles sont enregistrées chez les cadres. Également, lorsque nous croisons avec la durée totale du prêt, nous remarquons que chez les prêts mobilisant des capitaux élevés ce sont ceux dont la durée est importante qui font appel à des prestations en IT élevées.



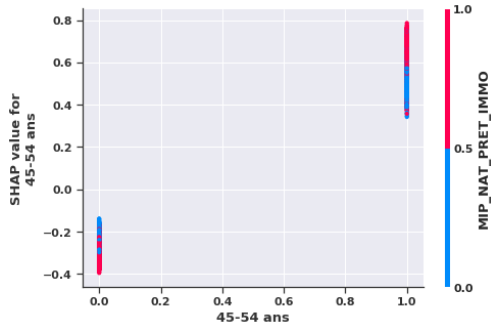
GRAPHIQUE 4.6 – Effet marginal de la somme assurée et interaction avec les cadres.



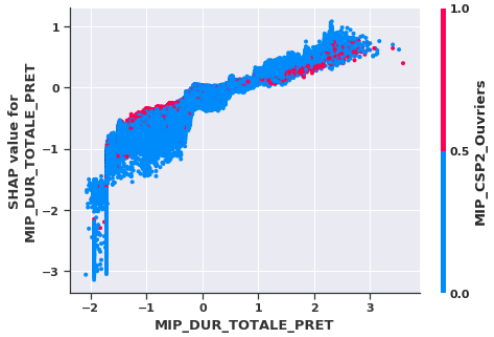
GRAPHIQUE 4.7 – Effet marginal de la somme assurée et interaction avec la durée totale du prêt.

Pour terminer, les graphiques ci-dessous représentent d'une part l'interaction entre l'âge et la

nature du prêt. Et d'autre part, l'interaction entre la durée totale du prêt et les ouvriers. Nous observons sur ces derniers que chez les emprunteurs dont la tranche d'âge à la souscription est [45 ; 55 ans) ce sont les prêts destinés à la consommation qui mobilisent des montants de prestations plus importants. Chez les autres tranches, ce sont les prêts immobiliers. Aussi, chez les prêts de courte durée, les emprunteurs ouvriers sont ceux qui demandent plus de prestations en IT. L'effet inverse est observé chez les prêts de durée longue.



GRAPHIQUE 4.8 – Effet marginal de l'âge et interaction avec la nature du prêt.



GRAPHIQUE 4.9 – Effet marginal de la durée du prêt et interaction avec les ouvriers.

5) Impacts sur les provisions et la rentabilité

DANS les deux chapitres ci-dessus, nous avons modélisé le risque de décès ou d'arrêt de travail de l'emprunteur séparément.

Dans ce dernier chapitre, nous cherchons à proposer des méthodes alternatives de provisionnement et de calcul de rentabilité en utilisant les résultats obtenus pour ces deux risques. Nous présentons en premier lieu le calcul de la Valeur Actuelle Probable pour les deux risques. Ensuite, nous présentons les impacts de nos résultats sur le calcul des PSAP. Enfin, nous exposons les impacts des résultats sur le calcul de la rentabilité du portefeuille.

5.1 Calcul de la sinistralité globale

Dans la section 1.1.4 du chapitre 1, nous avons présenté le principe général de calcul des engagements de l'assureur en cas de décès ou d'arrêt de travail de l'emprunteur. Dans les deux parties ci-dessous, nous illustrons comment ces derniers sont calculés grâce à nos résultats.

5.1.1 Calcul de la VAP Décès

Dans le cas du décès, nous avons calibré une fonction de survie qui représente la probabilité de survie ou de maintien en vie dans le prêt. Plus particulièrement, pour un emprunteur, nous disposons de la probabilité qu'il décède à un mois, deux mois, ..., n mois du prêt. Le capital restant dû du $k^{ème}$ mois du prêt ou bien le montant de la prestation à la charge de l'assureur est donné par :

$$CRD_k = CI \times \frac{(1+i)^n - (1+i)^k}{(1+i)^n - 1}$$

où i est le taux d'emprunt et n est le dernier mois d'échéance.

La formule ci-dessus est obtenue en faisant l'hypothèse de remboursement par mensualité constante où les versements périodiques (amortissement + intérêt i) sont constants pendant toute la durée de vie de l'emprunt.

Il en découle la formule de l'engagement de l'assureur en cas de décès de l'emprunteur :

$$VAP\_\text{Assureur}\_\text{DC} = \sum_{k=1}^{\text{Dass}} CRD_k \times \frac{l_{x(k)}}{l_{x(0)}} \times \frac{l_{x(k)} - l_{x(k+1)}}{l_{x(k)}} \times (1 + i_{\text{mens}})^{-(k+0,5)}$$

où Dass représente la durée de la garantie exprimée en mois.
En utilisant nos résultats sur la modélisation du risque décès, $\frac{l_{x(k)}}{l_{x(0)}}$ correspond à la probabilité de maintien en vie dans le prêt après k mois ($P(T \geq k)$) et $\frac{l_{x(k)} - l_{x(k+1)}}{l_{x(k)}}$ correspond à la probabilité que la durée de maintien en vie dans le prêt soit comprise entre k et $k+1$ sachant qu'elle est déjà supérieure à k ($P(k \leq T \leq k+1 | T \geq k)$). La nouvelle formule obtenue en remplaçant les quantités de la table de mortalité par celles de la table de maintien en vie est la suivante :

$$VAP\_\text{Assureur\_DC} = \sum_{k=1}^{\text{Dass}} CRD_k \times P(T \geq k) \times P(k \leq T \leq k+1 | T \geq k) \times (1 + i_{\text{mens}})^{-(k+0,5)}$$

En remplaçant la probabilité conditionnelle par sa valeur, nous obtenons que :

$$VAP\_\text{Assureur\_DC} = \sum_{k=1}^{\text{Dass}} CRD_k \times P(T \geq k) \times \frac{P(T \geq k) - P(T \geq k+1)}{P(T \geq k)} \times (1 + i_{\text{mens}})^{-(k+0,5)}$$

Une fois cette quantité obtenue, nous allons faire de même pour l'arrêt de travail.

5.1.2 Calcul de la VAP IT

Cependant, pour l'arrêt travail, il est possible de calibrer deux modèles qui expriment en réalité la même chose. En effet, nous avons cherché à prédire le montant des prestations en IT pour un emprunteur donné en se basant sur l'historique de notre portefeuille d'étude. La régression *XGBoost* a été choisie comme "meilleur" modèle de prédiction. Dans ce cas, cette prédiction constitue directement la valeur de l'engagement de l'assureur pour la couverture de l'arrêt de travail.

$$VAP\_\text{Assureur\_IT} = \text{Prestation prédite par XGBoost}$$

D'un autre côté, comme pour le décès, il est possible de calibrer une probabilité de maintien en IT. C'est la méthode la plus utilisée pour le calcul des engagements de l'assureur en cas d'arrêt de travail.
Dans ce cas, on obtient le calcul des prestations comme suit :

$$\text{prestation}_k = \sum_{j=1}^{Dass-k} REMB_{j+k} \times (1 + i_{mens})^{-(j+0,5)} \times \frac{l_{x(k)}^{maintien\_incapacite_{j+k}}}{l_{x(k)}^{maintien\_incapacite_k}}$$

Il s'en suit que la valeur de l'engagement de l'assureur est donnée par :

$$VAP\_\text{Assureur\_IT} = \sum_{k=1}^{\text{Dass}} \text{prestation}_k \times \text{proba entrée en IT en k.}$$

Enfin, l'engagement global de l'assureur est donné par :

$$VAP\_\text{Assureur} = VAP\_\text{Assureur\_DC} + VAP\_\text{Assureur\_IT}.$$

5.1.3 Application à notre portefeuille

L'objectif de cette partie est de valider l'utilisation de nos modèles de DC et IT pour le calcul des provisions pour sinistre à payer. Pour ce faire, nous appliquons les modèles sur les contrats de prêts clôturés pour lesquels les prestations sont totalement observées. L'idée est de vérifier à quel point ces modèles s'approchent des prestations réelles. Le critère de performance utilisé dans ce cas est l'écart relatif par rapport à la prestation réelle. La formule est donnée comme suit :

$$Ecart\ en\ \% = \frac{\text{valeur prédite - valeur réelle}}{\text{valeur réelle}} \times 100$$

Les résultats agrégés par classe de risque des coûts de la garantie DC sont représentés dans le tableau 5.1. L'analyse de ces résultats laisse entrevoir que les écarts avec les valeurs réelles sont globalement assez élevés. Plus particulièrement, une analyse suivant la classe de risque montre que la prédiction dans la classe 6 s'éloigne trop du coût observé dans cette classe. D'autre part, la classe 5 est celle qui a un écart avec la valeur réelle la plus faible (soit $-3,46\%$). Enfin, notre table d'expérience sous-estime généralement le coût du DC observé dans la classe de risque.

Ces écarts peuvent être expliqués, d'une part, par notre méthode d'amortissement des prêts sur la durée. En effet, nous avons fait l'hypothèse d'un remboursement annuel pour tous les prêts car nous ne pouvons pas tenir en compte des particularités de chacun de ces prêts. D'autre part, nous avons observé une faible réalisation du risque de DC dans notre portefeuille (environ 1%). Ainsi, nos modèles de *Machine Learning* ne disposent pas suffisamment de réalisation du risque DC pour leur apprentissage : le *C-index* s'élève à 0,766.

Segmentation	Effectif (%)	Prestation réelle	Avec la table de maintien	Ecart en %
Classe 1	17,31%	8 849 699,23	11 606 796,40	31,15%
Classe 2	15,76%	24 737 580,88	17 261 711,72	-30,22%
Classe 3	16,82%	48 641 521,78	39 382 918,94	-19,03%
Classe 4	17,35%	26 092 425,77	23 138 470,89	-11,32%
Classe 5	17,29%	47 629 217,31	45 981 715,97	-3,46%
Classe 6	15,47%	258 943 456,40	162 757 463,95	-37,15%
Total	100,00%	414 893 901,37	300 129 077,87	-27,66%

TABLEAU 5.1 – Validation de la table de maintien en vie pour le calcul de la PSAP DC

Les résultats agrégés par classe de risque des prestations en IT sont représentés dans le tableau 5.2 ci-dessous. Il ressort de l'analyse des résultats de ce tableau que le modèle *XGBoost* s'approche globalement des prestations réelles par classe de risque. En outre, une analyse par classe de risque montre que le modèle s'ajuste presque parfaitement pour la classe 4 (niveau de risque moyen). D'un autre côté, nous notons un écart un peu important de $-10,12\%$ dans la classe 6 qui est celle la plus risquée. Enfin, nous remarquons aussi que le modèle *XGBoost* sous-estime globalement le montant des PSAP. Ainsi, il est intéressant de sur-ajuster ses prédictions lors du calcul des PSAP pour les prêts en cours ou pour les futurs souscripteurs.

Segmentation	Effectif (%)	Prestation réelle	Prédiction *XGBoost*	Ecart en %
Classe 1	17,31%	970 568,10	945 000,6	-2,63%
Classe 2	15,76%	1 562 945,22	1 646 113	5,32%
Classe 3	16,82%	4 237 758,03	3 983 598	-6,00%
Classe 4	17,35%	**7 561 741,65**	**7 566 726**	**0,07%**
Classe 5	17,29%	11 378 649,96	10 858 630	-4,57%
Classe 6	15,47%	32 843 353,44	29 519 940	-10,12%
Total	100,00%	58 555 016,40	54 520 007,60	-6,89%

TABLEAU 5.2 – Validation de *XGBoost* pour le calcul de la PSAP en IT

5.2 Impacts sur le calcul des PSAP

Les résultats de la section précédente ont moins bien validé l'utilisation de notre table de maintien en vie dans le prêt pour le calcul des PSAP pour le risque DC pour les raisons citées ci-dessous. Nous avons néanmoins réalisé des projections avec cette table afin d'avoir un aperçu sur le coût attendu du risque DC de notre portefeuille. Cependant, ils ont permis de valider l'utilisation du modèle de régression *XGBoost* pour le calcul du coût attendu dans le cas de la couverture d'un emprunteur contre le risque IT.

Cette section a pour but d'utiliser ces modèles afin de réaliser des projections sur les prêts en cours de notre portefeuille. Cette projection permet entre autres de calculer les provisions à l'ultime de ces prêts.

Dans le tableau 5.3 suivant, nous avons les résultats des PSAP DC pour les prêts en cours dans le portefeuille. Il en ressort de l'analyse de ce dernier que le coût de la garantie DC sur la durée résiduelle des prêts de notre portefeuille s'élève environ à 2 milliards d'euros. Une analyse suivant la classe de risque montre que la classe 6 mobilise plus de provision avec 778 millions d'euros. D'autre part la classe 1, au contraire, ne mobilise que 100 millions. Comme nous l'avons

signalé dans la section sur le backtesting, ces PSAP doivent faire l'objet d'un surajustement pour tenir en compte la sous-estimation de notre loi de survie d'expérience.

Segmentation	Effectif (%)	Engagement futur DC
Classe 1	8,61%	102 076 701,14
Classe 2	17,15%	115 060 258,39
Classe 3	18,29%	224 913 289,52
Classe 4	18,47%	208 354 164,18
Classe 5	16,28%	487 116 680,14
Classe 6	21,20%	778 628 914,50
Total	100,00%	1 916 150 007,87

TABLEAU 5.3 – PSAP à l'ultime du risque DC pour les prêts en cours

Les valeurs des PSAP IT renseignées dans le tableau A.8 sont obtenues par prédiction du modèle *XGBoost* sur la durée résiduelle du prêt.

Segmentation	Effectif (%)	Prestation versée	Engagement futur IT
Classe 1	8,61%	753 765,93	514 056,40
Classe 2	17,15%	4 107 939,44	1 398 944,00
Classe 3	18,29%	7 150 201,35	3 085 452,00
Classe 4	18,47%	15 711 631,43	5 754 540,00
Classe 5	16,28%	12 161 075,69	6 858 002,00
Classe 6	21,20%	12 630 094,54	14 546 080,00
Total	100,00%	52 514 708,38	32 157 074,40

TABLEAU 5.4 – PSAP à l'ultime du risque IT pour les prêts en cours

Il ressort de l'analyse de ce tableau que les provisions à l'ultime de la garantie IT de notre portefeuille s'élève à 32 millions d'euros. Plus particulièrement, nous remarquons que la classe 1 nécessite que 500 milles de PSAP alors que la classe 6 en nécessite 14 millions. Enfin comme le montre le backtesting, le modèle *XGBoost* a tendance à un peu sous-estimer les coûts de la garantie IT. Ainsi ces PSAP peuvent être légèrement inférieures à la sinistralité future.

5.3 Impacts sur le calcul de la rentabilité du portefeuille

Les objectifs de cette section sont d'une part de faire l'état des lieux de la rentabilité du portefeuille en utilisant les prêts terminés. Et d'autre part, elle propose aussi une méthode de calcul de la rentabilité à l'ultime pour les prêts en cours en utilisant nos résultats de modélisation. L'indicateur de rentabilité utilisé à cet effet est le *Loss Ratio*.

5.3.1 Rentabilité constatée sur les prêts terminés

Pour chaque prêt terminé, nous disposons du montant total des prestations versées toute garantie confondue. D'un autre côté, nous avons dans notre portefeuille la prime globale versée par chaque emprunteur ayant clôturé son prêt.
En rapportant aux prestations totales aux primes totales encaissées, nous avons obtenu le *Loss Ratio* par classe de risque :

$$\text{Loss Ratio (en \%)} = \frac{\text{Prestation}}{\text{Prime}} \times 100$$

Le tableau 5.5 ci-dessous récapitule les prestations totales versées (toute garantie confondue) par CNP Assurances sur chaque classe de risque. Nous avons aussi récapitulé sur chacune de ces classes, les primes totales reçues de ces prêts.

Segmentation	Effectif (%)	Prestation	Prime	Loss Ratio
Classe 1	17,31%	9 820 267,33	39 515 229,75	24,85%
Classe 2	15,76%	26 596 420,57	42 309 909,84	62,86%
Classe 3	16,82%	52 879 279,81	85 276 711,52	62,01%
Classe 4	17,35%	33 654 167,42	71 372 561,21	47,15%
Classe 5	17,29%	59 007 867,27	109 767 681,54	53,76%
Classe 6	15,47%	291 786 809,80	192 608 575,85	151,49%
Total	100,00%	473 744 812,20	540 850 669,71	87,59%

TABLEAU 5.5 – Rentabilité constatée sur les prêts terminés, Loss Ratio par classe de risque

Il en ressort de l'analyse de ce tableau que le *Loss Ratio* est globalement à l'ordre de 87,59% (inférieur à 100%). Ainsi, la tarification existante dans le portefeuille est donc assez performante. Une analyse suivant la classe de risque montre que le *Loss Ratio* est supérieur à 100% pour la classe 6 mais on récupère cette perte à travers les autres classes notamment la classe 1.

5.3.2 Rentabilité à l'ultime sur les prêts en cours

Pour chaque prêt en cours, nous pouvons calculer grâce à nos modèles les provisions pour sinistre à payer sur les risques DC et IT. Aussi, grâce à notre table d'expérience de maintien en vie, nous pouvons calculer le montant de prime attendu toute garantie confondue pour un emprunteur.

Plus explicitement, la formule de calcul des primes est donnée comme suit :

$$\text{Prime globale attendue} = T_a \times \sum_{j=1}^{\text{Dass}} P(T \geq j) \times CI \text{ ou } CRD_j \times \left(1 + i_f\right)^{-j}$$

Avec :

— CI : le Capital Initial ;

— CRD_j : le Capital Restant Dû à la $j^{ème}$ année du prêt ;

— T_a : tarif appliqué global CI ou CRD

— $Dass$: le nombre de versements de primes (durée du prêt) ;

— $P(T \geq j)$: probabilité que l'emprunteur survive à la $j^{ème}$ année de son prêt obtenue avec notre table de maintien ;

— i_f : le taux d'intérêt technique.

Les résultats du calcul de la rentabilité à l'ultime sont renseignés dans le tableau 5.6. Il ressort de l'analyse de ce tableau que le *Loss Ratio* est globalement à l'ordre de 71,05%. En outre, une analyse suivant les classes de risques laisse entrevoir que la classe 6 a un *Loss Ratio* supérieur à 100% mais ceux des autres classes de risques restent inférieurs à 100%.

Segmentation	Effectif (%)	Prestation attendue	Prime attendue	Loss Ratio à l'ultime
Classe 1	8,61%	103 184 394,90	292 512 723,12	35,28%
Classe 2	17,15%	120 989 902,71	286 389 235,88	42,25%
Classe 3	18,29%	235 736 104,39	448 858 996,75	52,52%
Classe 4	18,47%	230 471 749,54	448 858 996,75	51,35%
Classe 5	16,28%	506 373 044,70	522 573 284,61	96,90%
Classe 6	21,20%	811 208 957,76	764 243 459,78	106,15%
Total	100,00%	2 007 964 154,00	2 825 391 159,73	71,07%

TABLEAU 5.6 – Loss Ratio à l'ultime des prêts en cours

Conclusion

En somme, il était question dans ce mémoire d'analyser et de quantifier en maille fine les deux plus grands risques en assurance emprunteur : le Décès (DC) et l'Incapacité de Travail (IT). Pour ce faire, nous avons confronté les approches classiques et *Machine Learning* dans la résolution de cette problématique.

A cet effet, nous avons commencé par segmenter notre portefeuille par scoring en utilisant la régression logistique. Les variables binaires de DC et d'entrée en IT sont utilisées lors de cette segmentation. Les résultats de cette dernière nous ont permis de construire six classes de risques homogènes de notre portefeuille d'étude. Ensuite, nous avons mis en place une loi d'expérience de maintien en vie dans le prêt en testant le modèle classique de *Cox*, le *XGBoost Cox* et le *XGBoost survival embedding (Xgbse)*. Pour l'arrêt de travail, l'estimation des provisions pour sinistres à payer a été obtenue en testant trois modèles : *Tweedie*, *XGBoost* et les *Réseaux de neurones*. En combinant les résultats issus des risques DC et IT, nous avons pu challenger les méthodes de provisionnement et de tarification existantes dans le portefeuille.

Les résultats ont montré que l'hypothèse du modèle de *Cox* selon laquelle les risques sont proportionnels ne sont pas vérifiés par nos données. Par conséquent, nous avons utilisé ses alternatives *Machine Learning* pour construire notre table de maintien en vie dans le prêt. D'après le modèle *XGBoost Cox*, les variables les plus importantes en ce qui concerne l'explication de la durée passée en vie dans le prêt sont : l'âge de l'assuré à la souscription du prêt, la durée du prêt et le sexe. Aussi selon ce modèle, plus l'âge à la souscription du prêt est élevé moins la durée passée en vie dans le prêt est importante. Et le risque de mortalité est plus élevé chez les hommes que chez les femmes. Cependant, le *XGBoost Cox* offre des estimations ponctuelles plutôt que la prédiction des courbes de survie. En outre, il ne permet pas l'estimation d'intervalles de confiance des probabilités de survie. Par conséquent, le module *Xgbse* est construit en ce sens pour rendre *XGBoost* beaucoup plus compatible avec les données de durées. Nous avons ainsi utilisé ce module pour construire notre courbe de survie. Néanmoins, elle affiche une légère sous performance par rapport au *XGBoost* classique en se référant à l'indice de concordance.

Les résultats ont également montré que dans la modélisation des PSAP en IT, nous obtenons des meilleures performances avec le modèle *XGBoost*. En effet, l'indice de Gini ainsi que le RMSE conduisent au choix de ce modèle. Il en ressort de l'analyse des résultats que les variables les plus importantes dans la détermination des PSAP en IT sont : le capital assuré, la durée

totale du prêt et la tranche d'âge [45 ; 55 ans). Plus particulièrement, nous constatons que les emprunteurs appartenant à la tranche d'âge à la souscription [45 ; 55 ans) ou [38 ; 45 ans) ont des montants de prestations en IT plus importants que ceux de la tranche [18 ; 38 ans) (modalité de référence).

Les résultats de l'estimation des risques DC et IT nous ont permis de proposer une nouvelle méthode de provisionnement de ces deux risques et d'apprécier la robustesse de la tarification existante dans le portefeuille. En effet, d'une part, le backtesting du modèle de régression *XGBoost* affiche des écarts aux prédictions réelles assez satisfaisants. Ce qui valide son utilisation pour le calcul des provisions pour sinistres à payer des prêts en cours ainsi que des nouveaux prêts. D'autre part, à l'aide de notre table de maintien, nous avons pu calculer la prime globale attendue de la part de chaque emprunteur. Cette prime est utilisée par la suite pour déterminer l'indicateur de rentabilité en assurance : le *loss ratio*. L'analyse de ce dernier a validé globalement la robustesse de la tarification existante dans le portefeuille. Du point de vue opérationnel, ces résultats vont permettre de visualiser rapidement la déformation de notre portefeuille d'étude en termes de sinistralités. Ainsi, la CNP Assurance peut en quelque sorte utiliser ces différents résultats comme un premier indicateur de suivi de risque.

Enfin, la limite principale de l'étude demeure la non prise en compte explicite de l'interdépendance entre les risques DC et IT pour un emprunteur qui est en IT. En effet, nous avons agrégé les deux risques par sommation ce qui suppose une indépendance entre les deux. Alors qu'en réalité, un emprunteur qui entre dans l'état Incapacité de Travail a une probabilité de passer à l'état DC beaucoup plus forte que quelqu'un qui se trouve dans l'état valide. L'agrégation des deux risques devrait plutôt se faire en utilisant par exemple un coefficient de passage de l'état IT à l'état DC. Néanmoins, cette approche reste valide pour un nouveau souscripteur de contrat d'assurance de prêt. Dans le cadre de la littérature actuarielle, cette étude est utile dans la mesure où elle illustre la contribution des méthodes de *Machine Learning* pour analyser la sinistralité en assurance emprunteur. Une autre piste d'extension possible est la proposition d'intervalles de confiance aux estimations des prestations versées en IT avec les réseaux de neurones. Celle-ci se fait grâce à la méthode de *Monte Carlo*.

Bibliographie

BABIN, S. (2016). Création de tables de mortalité d'expérience après segmentation d'un portefeuille de prêts personnels par scoring. Mémoire Institut des Actuaires.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1983). Classification and regression trees.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1) :1–3. Publisher : American Meteorological Society Section : Monthly Weather Review.

Chen, T. and Guestrin, C. (2016). Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the royal statistical society series b-methodological*, 34 :187–220.

Frees, E. W. J., Meyers, G., and Cummings, A. D. (2014). Insurance Ratemaking and a Gini Index. *The Journal of Risk and Insurance*, 81(2) :335–366. Publisher : [American Risk and Insurance Association, Wiley].

Friedman, J. (2000). Greedy function approximation : A gradient boosting machine. *The Annals of Statistics*, 29.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18) :2529–2545.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, 247 18 :2543–6.

He, X., Pan, J., Jin, O., Xu, T., Liu, B., Xu, T., Shi, Y., Atallah, A., Herbrich, R., Bowers, S., and Candela, J. (2014). Practical lessons from predicting clicks on ads at facebook. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

KALFA, S. (2017). Analyse de la rentabilité d'un contrat d'assurance emprunteur solvabilité 2. Mémoire Institut des Actuaires.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53 :457–481.

KOYE, G. K. (2019). Comparaison des methodes classiques et alternatives avec le machine learning pour la construction d'une table de mortalite d'expérience best estimate. Mémoire Institut des Actuaires.

Kvamme, H. and Borgan, (2019). The Brier Score under Administrative Censoring : Problems and Solutions. *arXiv :1912.08581 [cs, stat]*. arXiv : 1912.08581.

Lundberg, S. (2021). slundberg/shap. original-date : 2016-11-22T19 :17 :08Z.

Lundberg, S. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *arXiv :1705.07874 [cs, stat]*. arXiv : 1705.07874.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2019). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv :1802.03888 [cs, stat]*. arXiv : 1802.03888.

Marmerola, G. D. and Marmerola, G. D. (2018). Calibration of probabilities for tree-based models.

Ndiaye, A. (2020). Estimation des Prestations, PSAP et Intervalles de confiance en assurance santé. Mémoire Institut des Actuaires.

SCHOENFELD, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1) :239–241.

Uno, H., Cai, T., Pencina, M., D'Agostino, R., and Wei, L. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30 :1105–17.

Vieira, D., Gimenez, G., Marmerola, G., and Estima, V. (2021). Xgboost survival embeddings : improving statistical properties of xgboost survival analysis implementation.

Xacur, O. and Garrido, J. (2015). Generalised linear models for aggregate claims : to Tweedie or not ? *European Actuarial Journal*, 5.

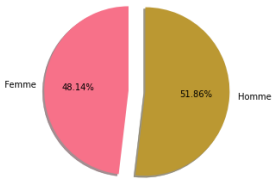
Yu, C.-N., Greiner, R., Lin, H.-C., and Baracos, V. (2011). Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. page 9.

Variables	NB OBS	MIN	MAX	Q1	MOYENNE	Q3	Ecart-type
MIP_AGE_ASSU_DT_SCRP	2 597 550	18	70	33	42	51	11,28
MIP_GENERATION	2 597 550	1986	2019	2001	2005	2008	4,44
MIP_CAP_INIT	2 597 550	100,63	2 014 647	3 000	24 910,27	30 000	40 379
MIP_QUOTITE_GLOBALE	2 597 550	0,01	1	0,7	0,86	1	0,23
MIP_CI_ASSU_GLOB	2 597 550	1,50	1 800 000	2 500	20 922	24 000	34 280
MIP_DUR_TOTALE_PRET	2 597 550	2	384	62	113,58	160	57,07
MIP_DUREE_REELLE_PRET	2 597 550	0	329	61	109	150	55,54
MIP_MNT_PRIM_AN_GLOB	2 597 550	0	13 205	7,88	66,88	80	113,13
MIP_TARIF_APP_GLOBAL	2 597 550	0	0,013	0,0032	0,0032	0,0032	0,001
MIP_TX_EMPRUNT	2 597 550	0	0,093	0,034	0,04	0,047	0,01

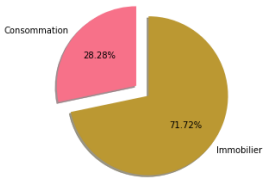
TABLEAU A.1 – Description des variables quantitatives.

Variables	NB OBS	MIN	MAX	Q1	MOYENNE	Q3	Ecart-type
MIP_AGE_ASSU_DT_SCRP	16 026	18	70	40	48,6	57	11,15
MIP_AGE_ASSU_DT_SINIS	16 026	20	75	48	55	64	10,7
MIP_GENERATION	16 026	1994	2019	2003	2007	2010	4,74
MIP_CAP_INIT	16 026	150	2 014 647	6 500	43 419	65 000	53 500
MIP_QUOTITE_GLOBALE	16 026	0,1	1	1	1	1	0,2
MIP_CI_ASSU_GLOB	16 026	24,1	56 248	38 565	20 922	56 248	46 887
MIP_DUR_TOTALE_PRET	16 026	12	384	110	159	192	69,9
MIP_DUREE_REELLE_PRET	16 026	0	265	40	78,96	111	48,1
MIP_MNT_PRIM_AN_GLOB	16 026	0	5 681,4	20,52	140,43	198	189,93
MIP_TARIF_APP_GLOBAL	16 026	0,0009	0,0132	0,0032	0,0038	0,0039	0,001
MIP_TX_EMPRUNT	16 026	0	0,0920	0,0333	0,0387	0,0455	0,01

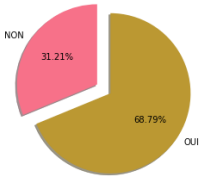
TABLEAU A.2 – Description des variables quantitatives des emprunteurs sinistrés.



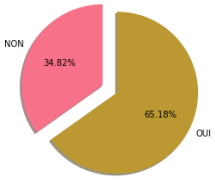
GRAPHIQUE A.1 – Répartition des prêts selon le sexe de l'emprunteur.



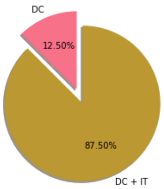
GRAPHIQUE A.2 – Répartition des prêts selon la nature.



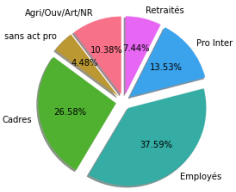
GRAPHIQUE A.3 – Répartition des prêts selon la présence de plusieurs emprunteurs.



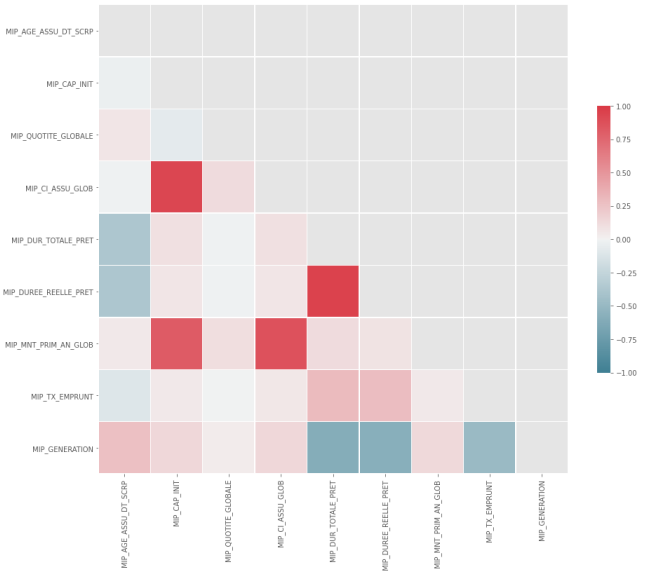
GRAPHIQUE A.4 – Répartition des prêts selon la détention de plusieurs prêts.



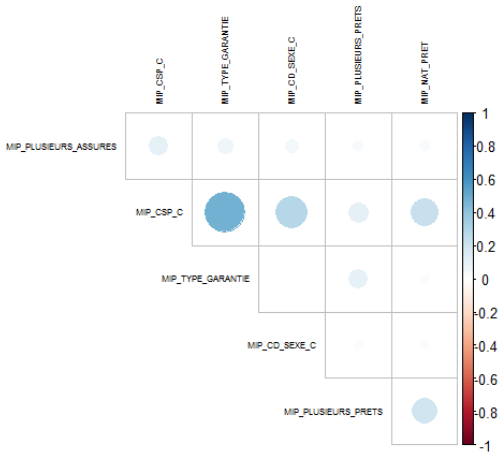
GRAPHIQUE A.5 – Répartition des prêts selon le type de couverture.



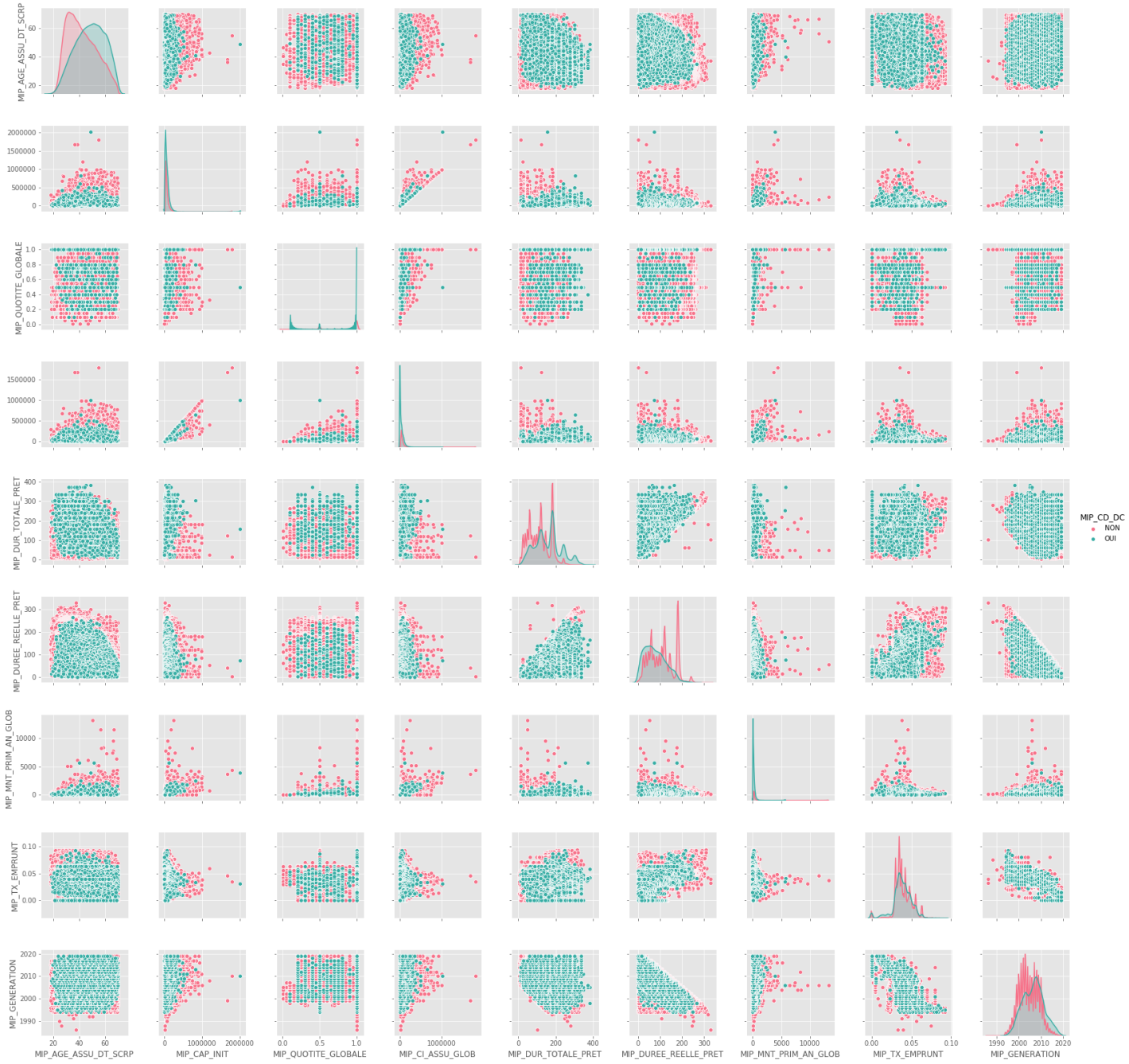
GRAPHIQUE A.6 – Répartition des prêts selon la catégorie socio-professionnelle.



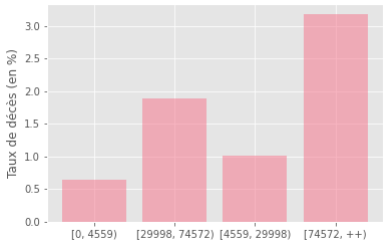
GRAPHIQUE A.7 – Matrice de corrélation des variables quantitatives.



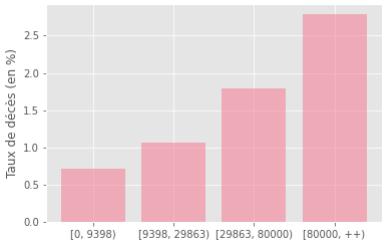
GRAPHIQUE A.8 – Matrice de corrélation des variables qualitatives.



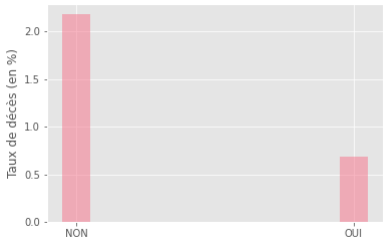
GRAPHIQUE A.9 – Matrice de distribution des variables quantitatives suivant le décès.



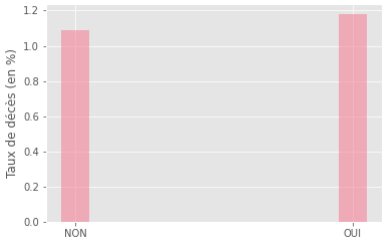
GRAPHIQUE A.10 – Taux de décès en fonction du capital assuré.



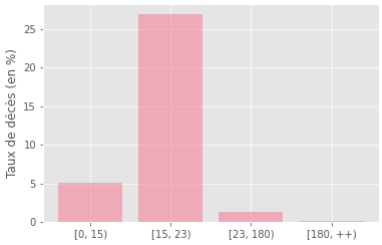
GRAPHIQUE A.11 – Taux de décès en fonction du capital initial.



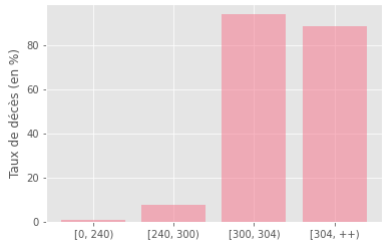
GRAPHIQUE A.12 – Taux de décès en fonction de la présence de plusieurs emprunteurs.



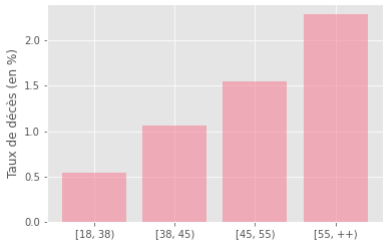
GRAPHIQUE A.13 – Taux de décès en fonction de la détention de plusieurs prêts.



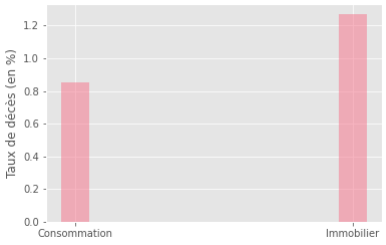
GRAPHIQUE A.14 – Taux de décès en fonction de la durée réelle du prêt.



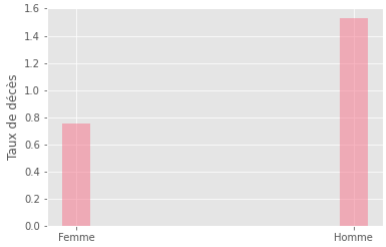
GRAPHIQUE A.15 – Taux de décès en fonction de la durée totale du prêt.



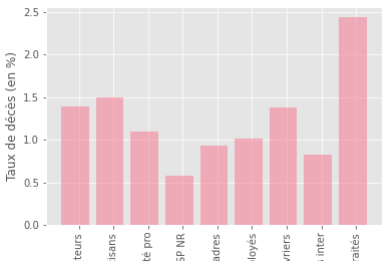
GRAPHIQUE A.16 – Taux de décès en fonction de l'âge à la souscription.



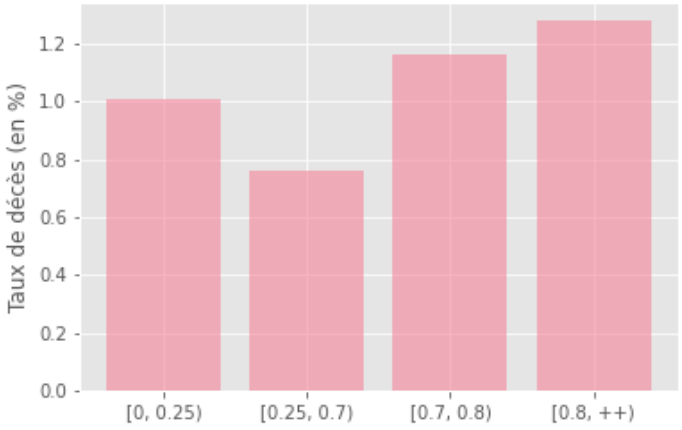
GRAPHIQUE A.17 – Taux de décès en fonction de la nature du prêt.



GRAPHIQUE A.18 – Taux de décès en fonction du sexe.

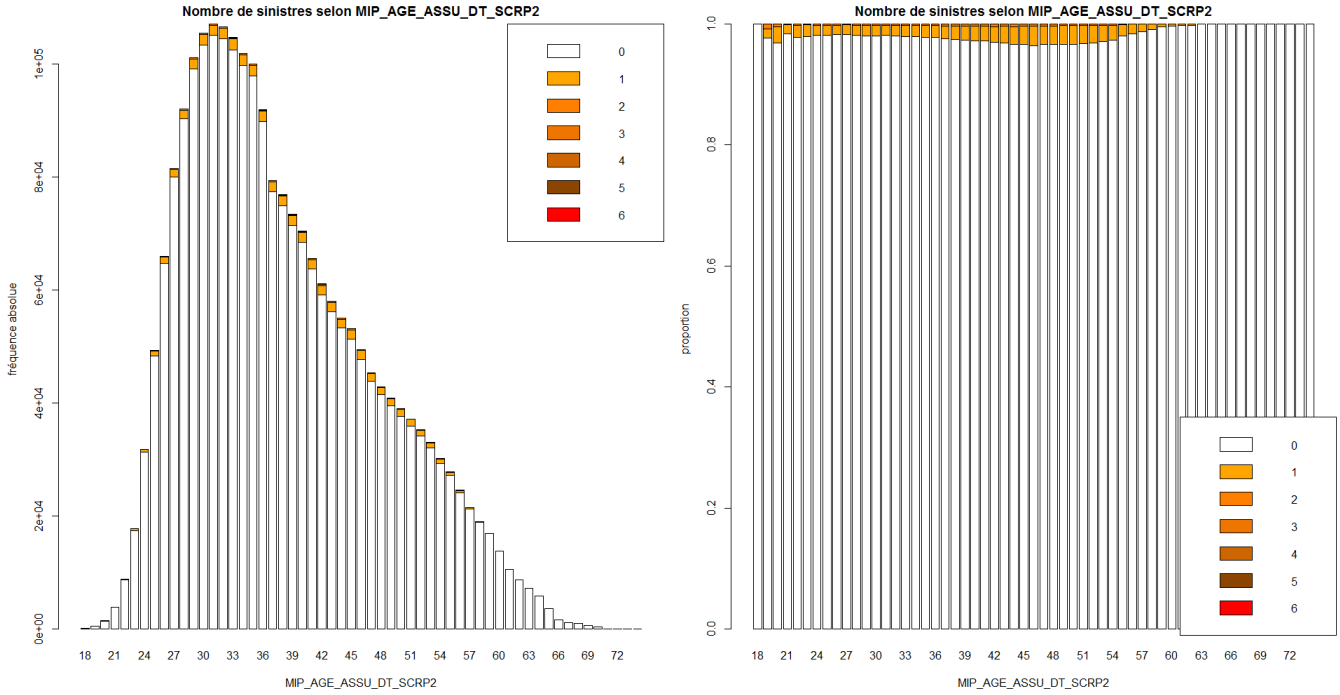


GRAPHIQUE A.19 – Taux de décès en fonction du CSP.

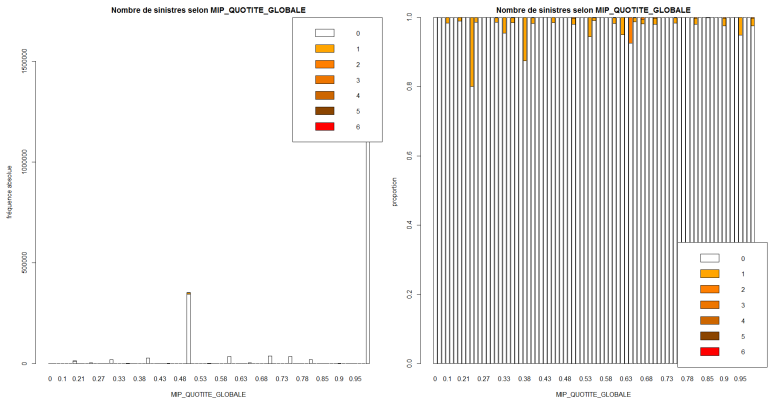


GRAPHIQUE A.20 – Taux de décès en fonction de la quotité.

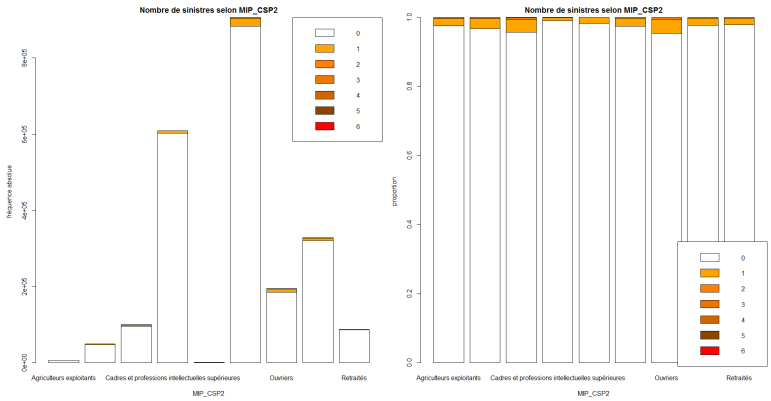
GRAPHIQUE A.21 – Nombre de sinistres selon l'âge à la souscription



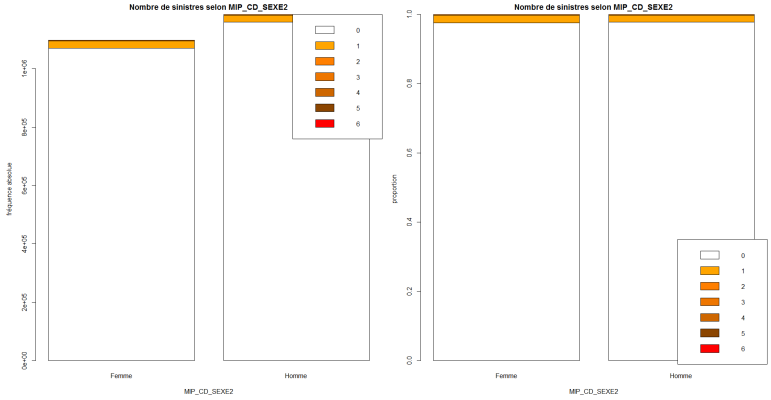
GRAPHIQUE A.22 – Nombre de sinistres selon la quotité assurée



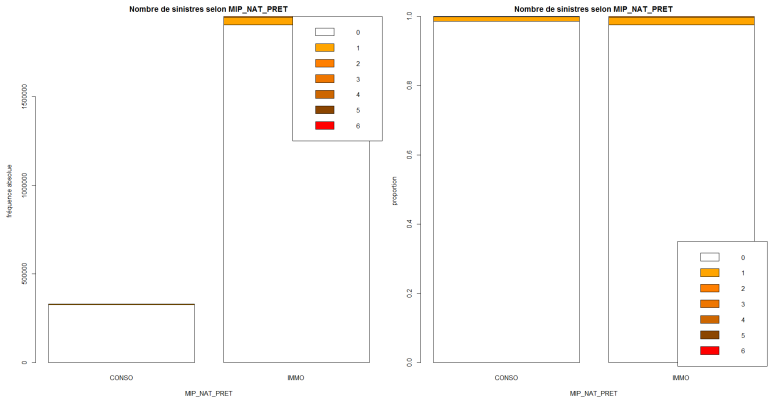
GRAPHIQUE A.23 – Nombre de sinistres selon la catégorie socio-professionnelle



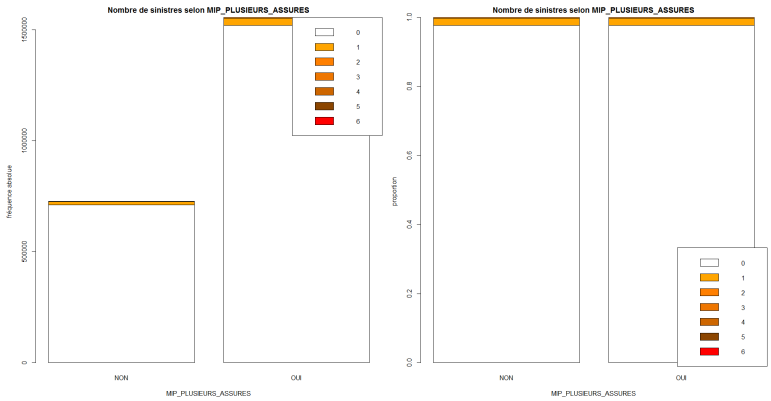
GRAPHIQUE A.24 – Nombre de sinistres selon le sexe



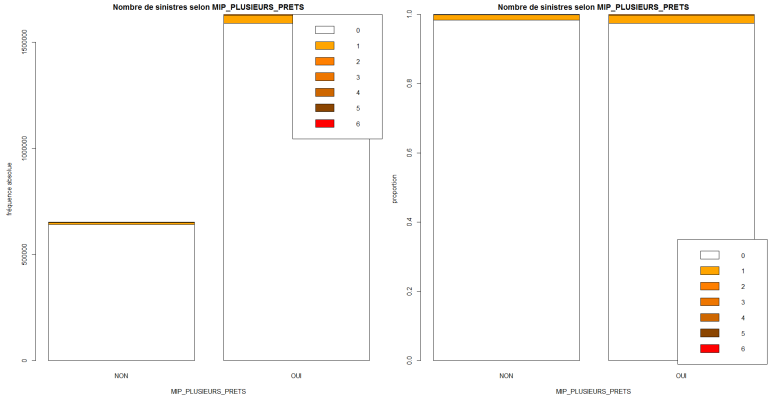
GRAPHIQUE A.25 – Nombre de sinistres selon la nature du prêt



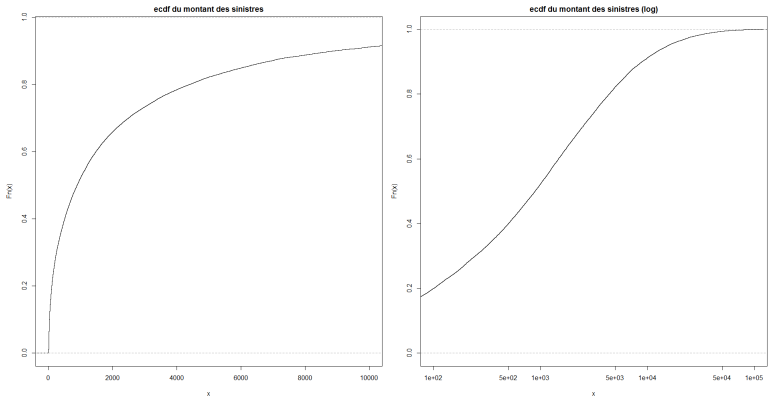
GRAPHIQUE A.26 – Nombre de sinistres selon la présence de plusieurs emprunteurs



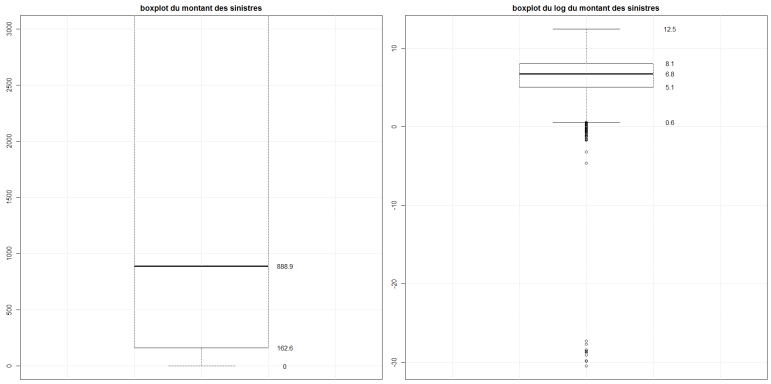
GRAPHIQUE A.27 – Nombre de sinistres selon la détention de plusieurs prêts



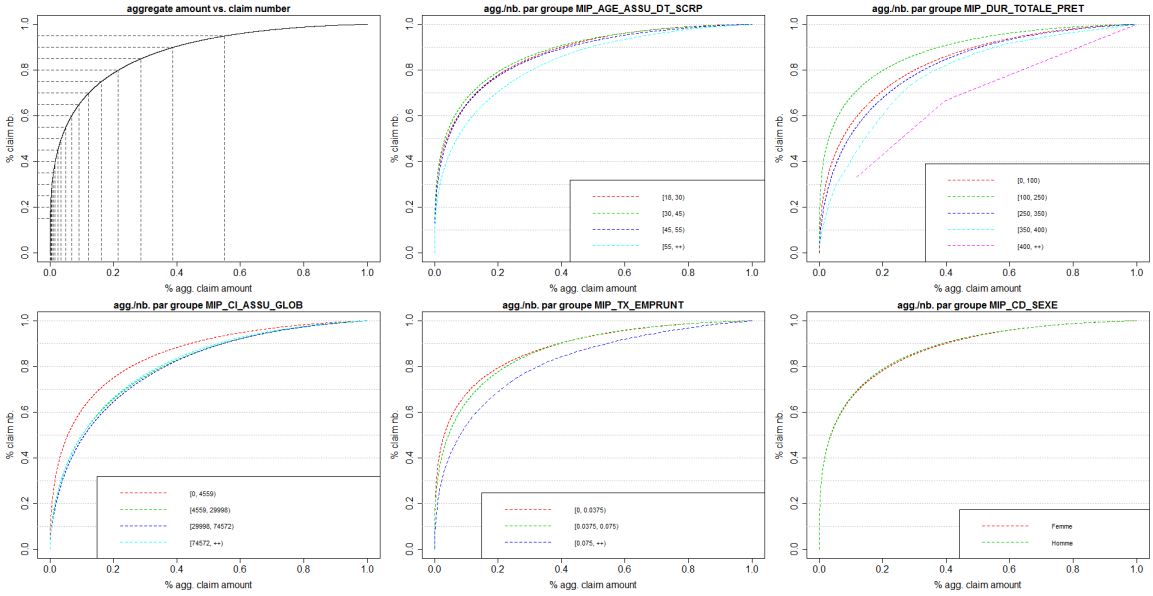
GRAPHIQUE A.28 – Fonction de répartition empirique

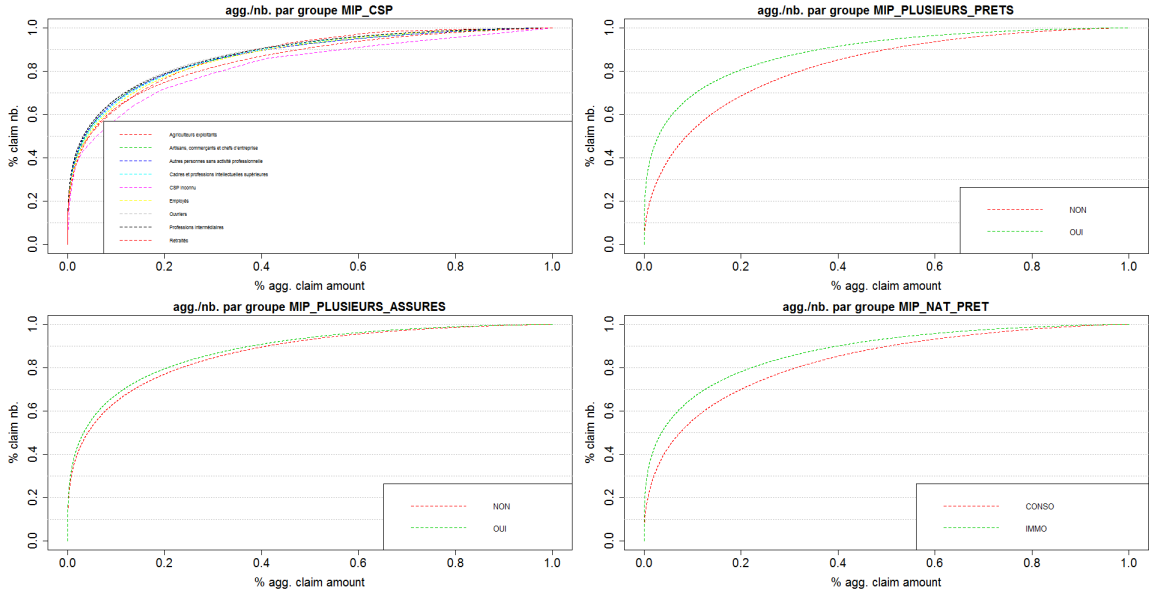


GRAPHIQUE A.29 – Boxplot des montants et des log-montants

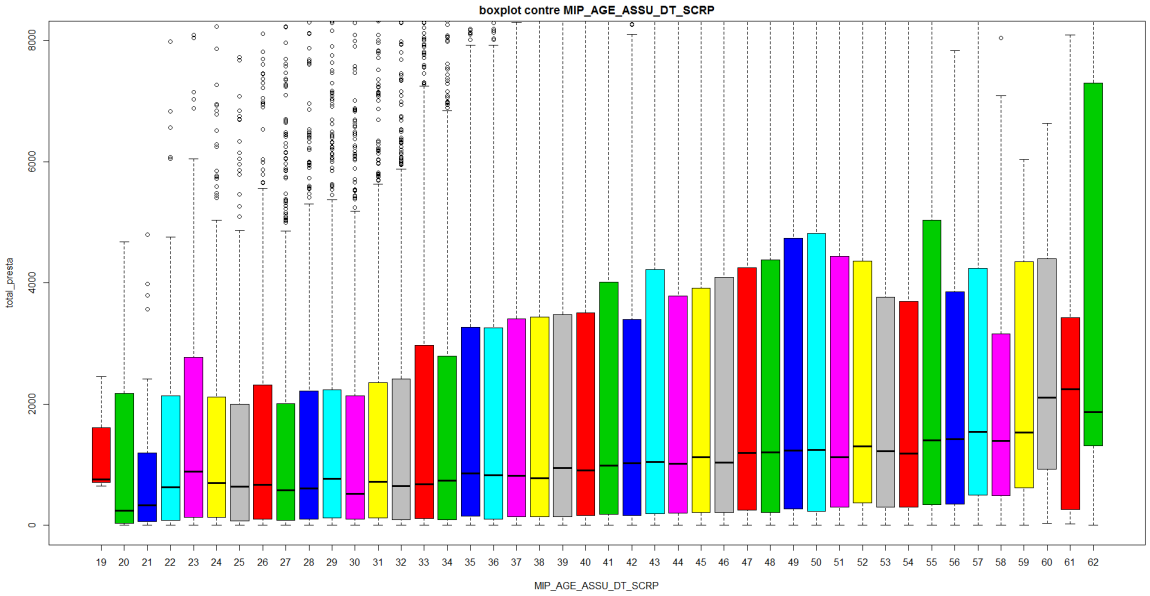


GRAPHIQUE A.30 – Montants de sinistre sur les caractéristiques

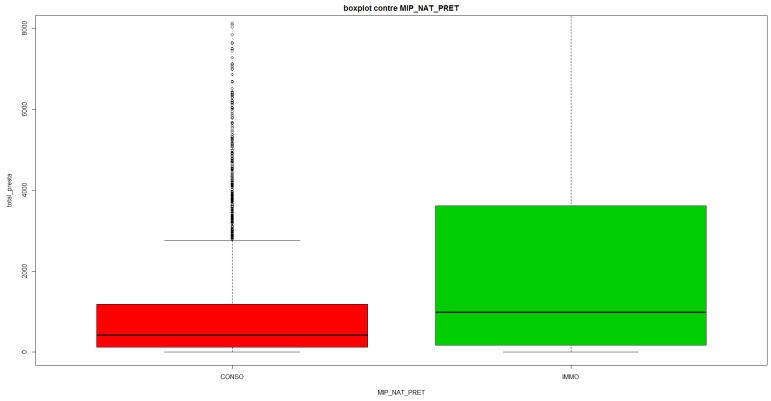




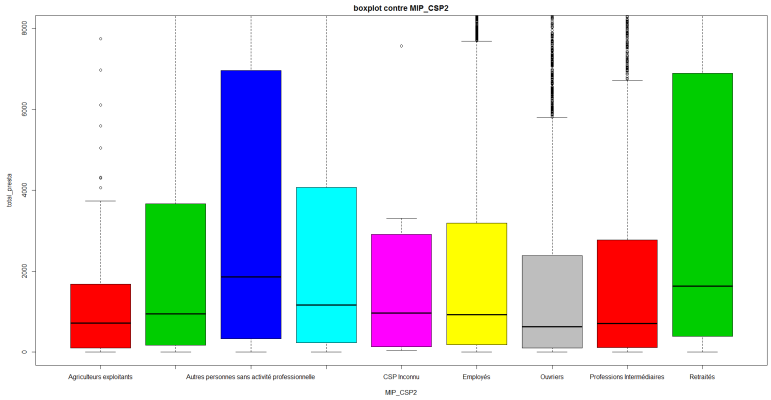
GRAPHIQUE A.31 – Boxplot des montants de sinistre selon l'âge à la souscription



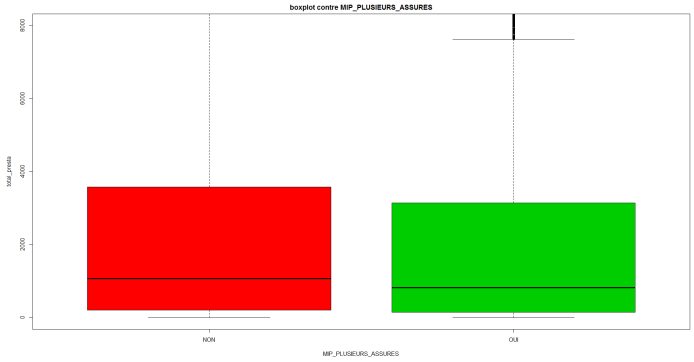
GRAPHIQUE A.32 – Boxplot des montants de sinistre selon la nature du prêt



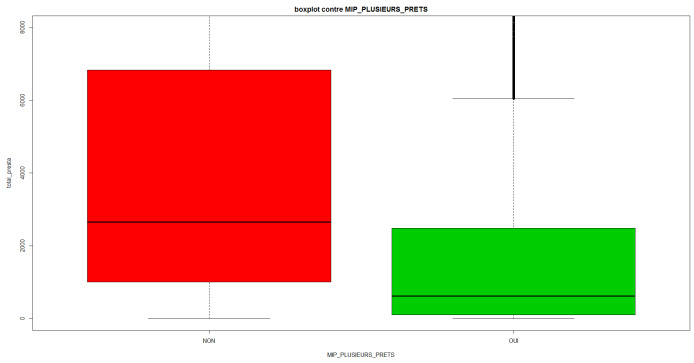
GRAPHIQUE A.33 – Boxplot des montants de sinistre selon la CSP de l'emprunteur



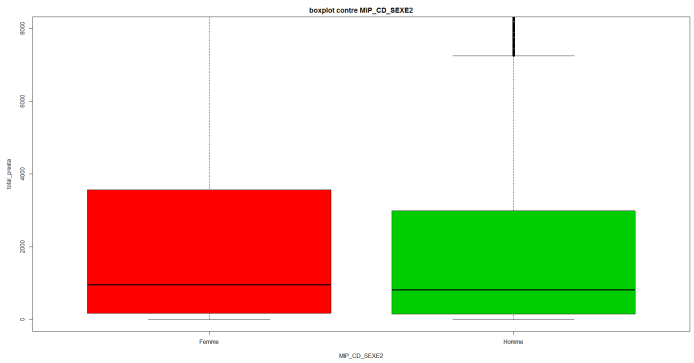
GRAPHIQUE A.34 – Boxplot des montants de sinistre selon la présence de plusieurs emprunteurs



GRAPHIQUE A.35 – Boxplot des montants de sinistre selon la détention de plusieurs prêts



GRAPHIQUE A.36 – Boxplot des montants de sinistre selon le sexe de l'emprunteur



A.1 Test de Hosmer Lemeshow (H-L)

Pour évaluer si le modèle logistique est capable de prédire le décès, on utilisera la statistique de qualité d'ajustement connue sous le nom de statistique de Hosmer-Lesmeshow. On la calcule en répartissant les probabilités estimées en déciles.

Nous considérons les notations suivantes :

— n_j, le nombre d'emprunteurs dans le groupe j, $j \in \{1, ..., 10\}$;

— d_j, le nombre de décès dans le groupe j ;

— q_j, la probabilité de décès moyenne observée pour le groupe j ;

— $\hat{q_j}$, la probabilité de décès moyenne estimée pour le groupe j.

L'hypothèse de base de ce test stipule que tous les décès sont indépendants aussi bien dans chaque groupe, que dans tous les groupes.

H_0 se définit alors comme suit : "Les probabilités de décès observées correspondent aux probabilités de décès attendues" (bonne adéquation du modèle aux données).

En outre, la statistique de test T_{H-L} est donnée par :

$$T_{H-L} = \sum_{j=1}^{10} \frac{(d_j - n_j \times \hat{q_j})^2}{n_j \times \hat{q_j} \times (1 - \hat{q_j})}$$

Sous H_0, cette statistique suit une loi du Chi-deux à 8 degrés de liberté. Ainsi, si la statistique est supérieure au quantile du Chi-deux à 8 degrés de liberté, H_0 est rejetée.

A.2 Découpage des variables avec *CART*

La discrétisation des variables quantitatives nous a permis pour certains de nos modèles (scoring et prestations IT) de mieux expliciter nos résultats.

Pour le scoring, la discrétisation est réalisée sous forme supervisée avec CART en utilisant comme variable d'intérêt le décès. L'algorithme découpe une variable quantitative V_i en k_i intervalles discrets et disjoints, comme indiqué ci-dessous :

$$D_i = \{[d_0, d_1), ..., [d_{k_{i-1}}, d_{k_i})\}$$

où d_0 et d_{k_i} sont les valeurs minimale et maximale de V_i, respectivement. Enfin, $Pi = \{d_1, d_2, ..., d_{k_{i-1}}\}$ désigne l'ensemble complet de points de coupe pour chaque attribut continu i. L'objectif de l'algorithme CART est de trouver le meilleur P_i pour l'attribut cible i.

A.3 Scoring du risque IT

Variables	Modalités	Estimation	Contributions	Notes	Niveau risque
MIP_AGE_ASSU_DT_SCRP_D	[18, 38)	0	17	4	Intermédiaire
MIP_AGE_ASSU_DT_SCRP_D	[38, 45)	0,4463	17	11	risqué
MIP_AGE_ASSU_DT_SCRP_D	[45, 55)	0,7435	17	17	très risqué
MIP_AGE_ASSU_DT_SCRP_D	[55, ++)	-0,2066	17	0	très peu risqué
MIP_NAT_PRET	IMMO	0	10	10	risqué
MIP_NAT_PRET	CONSO	-0,5686	10	0	très peu risqué
MIP_DUR_TOTALE_PRET_D	[0, 118)	0	49	0	très peu risqué
MIP_DUR_TOTALE_PRET_D	[118, 194)	0,8639	49	15	Intermédiare
MIP_DUR_TOTALE_PRET_D	[194, 258)	1,5926	49	28	risqué
MIP_DUR_TOTALE_PRET_D	[258, ++)	2,8104	49	49	très risqué
MIP_CSP_D	Agriculteurs	0,7335	25	13	risqué
MIP_CSP_D	Artisans	1,0053	25	18	risqué
MIP_CSP_D	Ouvriers	1,4351	25	25	très risqué
MIP_CSP_D	Retraités	0,2965	25	5	peu risqué
MIP_CSP_D	Cadres	0	25	0	très peu risqué
MIP_CSP_D	Employés	0,7827	25	14	risqué
MIP_CSP_D	Sans act Pro/NR	0,9987	25	18	très risqué
MIP_CSP_D	Pro Inter	0,7687	25	14	risqué

TABLEAU A.3 – Grille de score pour le risque IT.

Note de Synthèse

NIAKH Fallou

Apport du Machine Learning dans l'analyse multivariée de la sinistralité des contrats emprunteur

Préambule

Dans un des portefeuilles d'assurance emprunteur collectif détenus par CNP Assurances en collaboration avec ses banques partenaires couvrant les risques Décès (DC) et Incapacité de travail (IT), la tarification existante est du type tranche d'âge à la souscription × tranche de capital emprunté. Cependant, outres ces deux variables de tarification, nous avons aussi à notre disposition des variables comme la catégorie socioprofessionnelle, la durée totale du prêt, le capital assuré, la quotité d'assurance, etc. De ce fait, il convient de se demander si ces variables non tarifaires contribuent à l'explication de la sinistralité du portefeuille ?

Depuis déjà plusieurs décennies, les actuaires ont développé des méthodes statistiques traditionnelles pour estimer ces risques. Récemment, face à la grande quantité des données en entreprise, l'apport considérable du *Machine Learning* dans la résolution des problèmes d'apprentissage supervisé est aujourd'hui un fait. De plus, les algorithmes de *Machine Learning* permettent de capturer la structure de l'information sans recourir à des hypothèses fortes sur les distributions des variables, contrairement aux méthodes statistiques traditionnelles. Par conséquent, les algorithmes de *Machine Learning* peuvent être plus efficaces pour capturer et modéliser des phénomènes complexes.

Dans l'optique d'optimiser son processus de suivi de risque en assurance emprunteur, CNP Assurances souhaite éprouver les méthodes *Machine Learning*. Cette étude naît ainsi de cette préoccupation et se focalise sur deux garanties en contrats emprunteur : le décès et l'arrêt de travail. Elle analyse et quantifie ces deux risques en maille fine. Nous avons comparé à cet effet la robustesse et la performance des méthodes classiques et *Machine Learning*. Pour le risque

de DC, nous avons testé trois méthodes : le modèle classique de *Cox*, le *XGBoost Cox* et le *XGBoost survival embedding* dit *Xgbse*. Pour l'arrêt de travail, nous avons modélisé les PSAP en utilisant trois familles de modèles : les modèles linéaires généralisés (*GLM Tweedie*), les méthodes d'agrégation (*XGBoost Tweedie*) et les méthodes d'apprentissage profond (*Réseaux de neurones*). Les résultats obtenus sur ces deux risques ont permis de proposer des méthodes alternatives de provisionnement et de calcul de la rentabilité dans le portefeuille.

Présentation des données

Pour la réalisation de cette étude, nous disposons d'un portefeuille de contrats collectifs d'assurance emprunteur d'un établissement de crédit. Le portefeuille est ainsi constitué de contrats liés à des prêts pour financer des achats immobiliers ou des prêts de consommation couvrant les risques suivants : DC et IT. La base de données contient 2,8 millions de couples prêt/assuré observés à partir de 2007. Nous disposons des caractéristiques de l'assuré (âge à la souscription, le sexe, la catégorie socioprofessionnelle) et du prêt (capital initial, durée, quotité, dates de début et de fin,...). Le sexe est utilisé seulement à titre académique. Nos variables d'intérêts dans cette étude sont : les variables indicatrices indiquant si l'emprunteur est décédé ou pas ou s'il est une fois entré en IT durant sa durée de prêt, la durée passée en vie dans le prêt et les prestations totales en IT versées durant un contrat d'assurance emprunteur.

Segmentation des emprunteurs par scoring

Avant de se lancer sur la modélisation des risques DC et IT, nous avons cherché tout d'abord à segmenter notre portefeuille avec la régression logistique. Cette segmentation est importante dans la mesure où elle permet de déceler rapidement les profils de risques du portefeuille. Les variables binaires de DC et d'entrée en IT sont utilisées à cet effet. Grâce aux estimations des coefficients de ces deux modèles, nous avons établi une grille de score qui, à chaque modalité de chaque variable utilisée, associe un nombre de points compris entre 0 et 100. Le risque global d'un emprunteur est obtenu en faisant la somme de ses scores sur les risques DC et IT. Ainsi, à chaque emprunteur est attribuée une note de risque comprise entre 0 et 200.

Une fois la grille de score établie, nous avons découpé notre score global en différentes classes de risque afin d'avoir des taux de décès et d'entrée en IT observés différents pour chaque classe.

Segmentation	Effectif (%)	Taux de décès	Taux d'entrée en IT
Classe 1	17,31%	0,45%	0,79%
Classe 2	15,76%	1,06%	1,12%
Classe 3	16,82%	1,04%	1,44%
Classe 4	17,35%	0,65%	2,81%
Classe 5	17,29%	1,11%	3,63%
Classe 6	15,47%	3,09%	6,31%

TABLEAU A.4 – Segmentation du portefeuille suivant les risques DC et IT

Modélisation du risque DC

Pour un contrat de prêt détenu t années, il serait intéressant d'estimer la probabilité pour un emprunteur de décéder au cours de l'année t. Il s'agit de la probabilité qu'un emprunteur durant sa durée de prêt décède au cours d'une année donnée. Une manière d'estimer ce taux est donnée dans la démarche de construction de l'estimateur de *Cox* qui lui s'intéresse à une probabilité de survie. D'un autre côté, certaines méthodes de *Machine Learning* sont adaptées pour estimer la probabilité de survie de *Cox*. Nous présentons ci-dessous les résultats de la construction de la table de maintien en vie dans le prêt obtenus en testant les modèles de *Cox*, *XGBoost Cox* et *XGBoost survival embeding* (*Xgbse*).

Le modèle de *Cox*

Le modèle de risques proportionnels de *Cox* est une technique statistique qui permet d'étudier la relation entre la survie d'un individu et plusieurs variables explicatives. Dans notre cas, nous avons régressé la durée passée en vie dans le prêt en fonction des caractéristiques de l'emprunteur et de celles du prêt.
La validité de ce modèle nécessite la vérification de l'hypothèse fondamentale selon laquelle les risques sont proportionnels. Nos données ne vérifient pas cette hypothèse même après stratification suivant l'âge à la souscription du prêt.

Par conséquent, nous n'avons pas utilisé le modèle de *Cox* pour la construction de notre table de maintien en vie dans le prêt.

Les modèles *XGBoost* et *Xgbse*

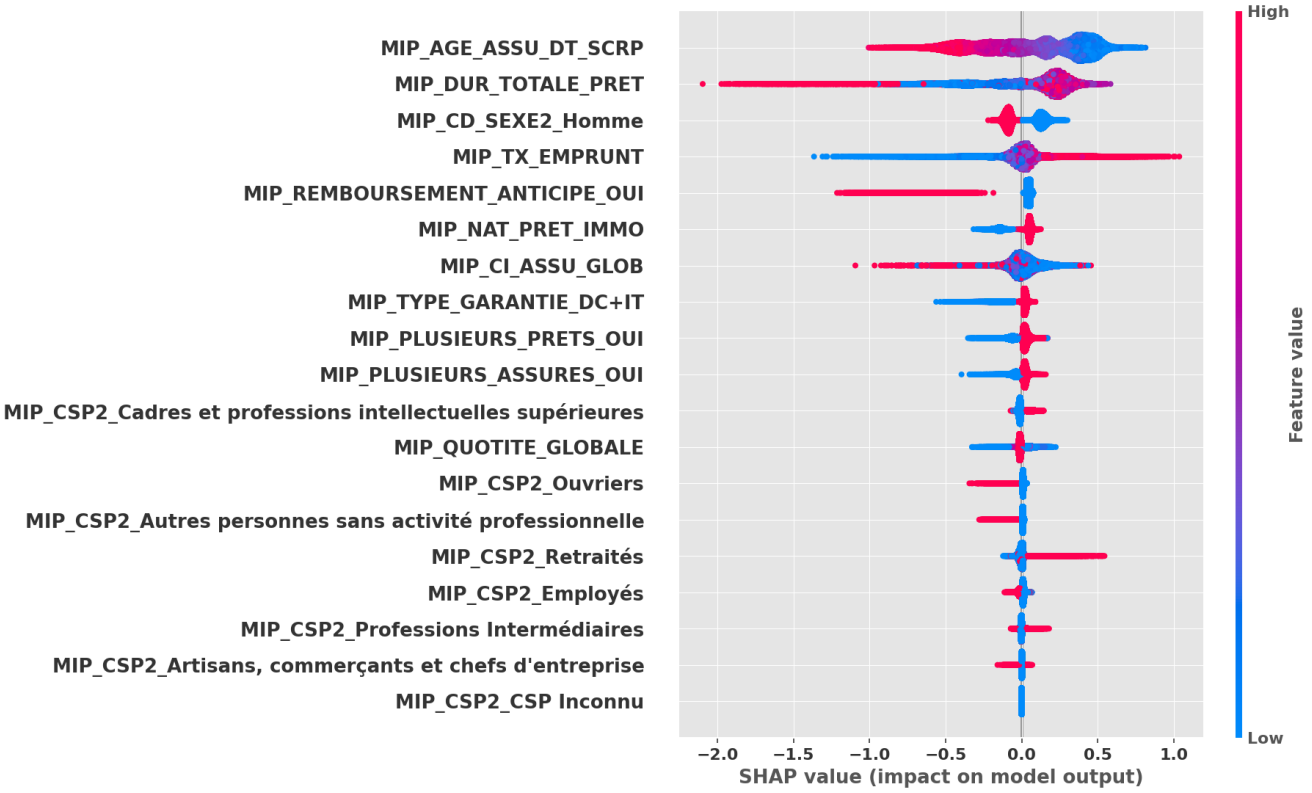
Le module *Xgbse* (*XGBoost survival embedding*) de python est une alternative du modèle de régression des risques proportionnels de *Cox*. Il combine la méthode *XGBoost* avec un ensemble de méthodes classiques telles que Kaplan Meier, les k plus proches voisins et la régression

logistique pour estimer la probabilité de survie. L'avantage de ce module par rapport au modèle *XGBoost* classique est qu'elle permet d'obtenir des prédictions des courbes de survie plutôt que des estimations ponctuelles. Elle permet aussi d'avoir une estimation des intervalles de confiance et des temps de survie attendus calibrés non biaisés. A l'issue de cette expérimentation, nous avons constaté que le modèle de *XGBoost* classique donne une légère meilleure performance par rapport à *Xgbse* en se référant à l'indice de concordance (voir tableau A.5).

Modèles	C-index	Brier Score
XGBoost	0,766	–
XGBSEDebiasedBCE	0,754	0,127
XGBSEKaplanNeighbors	0,740	0,136
XGBSEKaplanTree	0,732	0,154

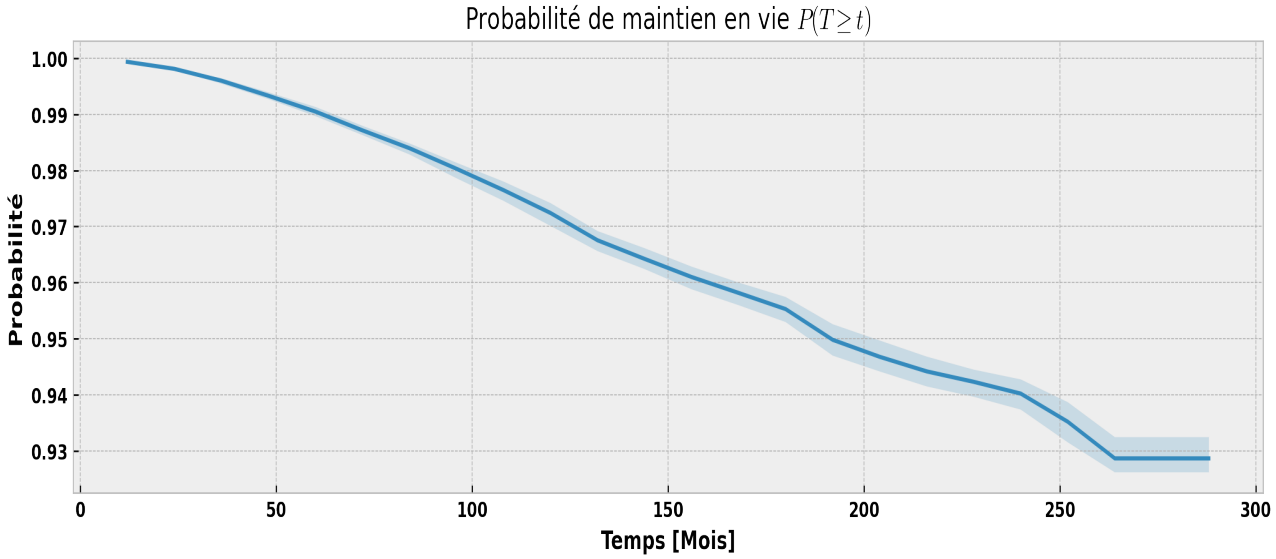
TABLEAU A.5 – Métriques de performance pour les mêmes variables explicatives

Au regard de la figure A.37, les variables les plus importantes pour le modèle *XGBoost* en ce qui concerne l'explication de la durée passée en vie dans le prêt sont : l'âge de l'assuré à la souscription du prêt, la durée du prêt et le sexe.



GRAPHIQUE A.37 – Effets marginaux des variables explicatives sur le risque de mortalité

Enfin, comme le modèle *XGBoost* classique ne permet pas d'avoir directement une courbe de survie non biaisée, nous avons utilisé le meilleur modèle de la famille des *Xgbse* (*XGBSEDebiasedBCE*) pour construire notre table de maintien en survie dans le prêt.



GRAPHIQUE A.38 – Table de maintien en vie dans le prêt via *XGBSEDebiasedBCE*

Modélisation du risque IT

La modélisation du coût individuel de la garantie arrêt de travail est réalisée dans ce mémoire en testant trois familles de méthodes : les modèles linéaires généralisés (*GLM Tweedie*), les méthodes d'agrégation (*XGBoost*) et les méthodes d'apprentissage profond (*Réseaux de neurones*). Le choix de ces trois catégories de modèles nous permet de capter l'apport des méthodes de *Machine Learning* dans l'analyse multivariée de la sinistralité en IT des contrats emprunteurs. La comparaison de ces modèles est effectuée en utilisant deux critères de performances : l'indice de Gini et le Root Mean Square Error (*RMSE*). L'indice de Gini permet de choisir le modèle qui s'adapte le plus à nos données mais aussi de mesurer le pouvoir de discrimination des hauts et des bas risques. Nous avons aussi eu recours au module *SHAP* de Lundeberg et al 2017 pour l'interprétation de nos résultats.

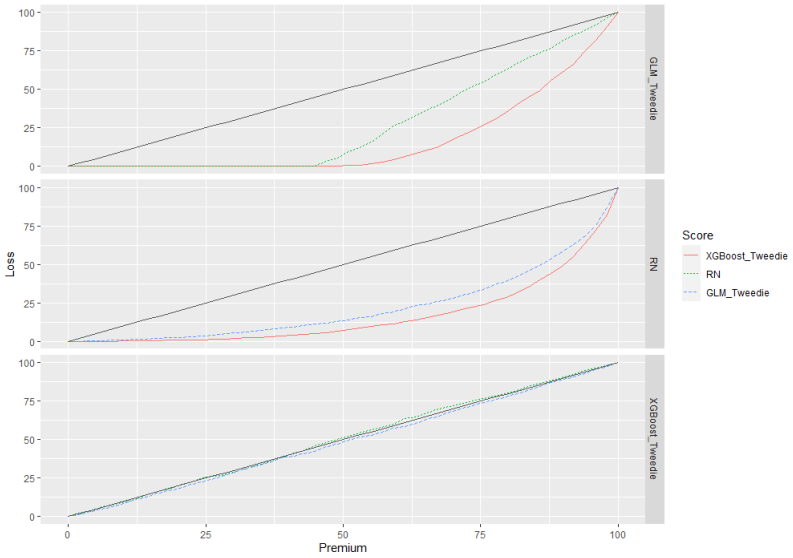
A l'issue de nos expérimentations, nous observons sur le tableau A.6 de meilleure performance avec le modèle *XGBoost*.

	GLM Tweedie	*XGBoost*	*Réseaux de neurones*
Indice de Gini	0,671	0,7495	0,578
RMSE	20627,271	1045,703	1052,454

TABLEAU A.6 – Métriques de performance pour les mêmes variables explicatives

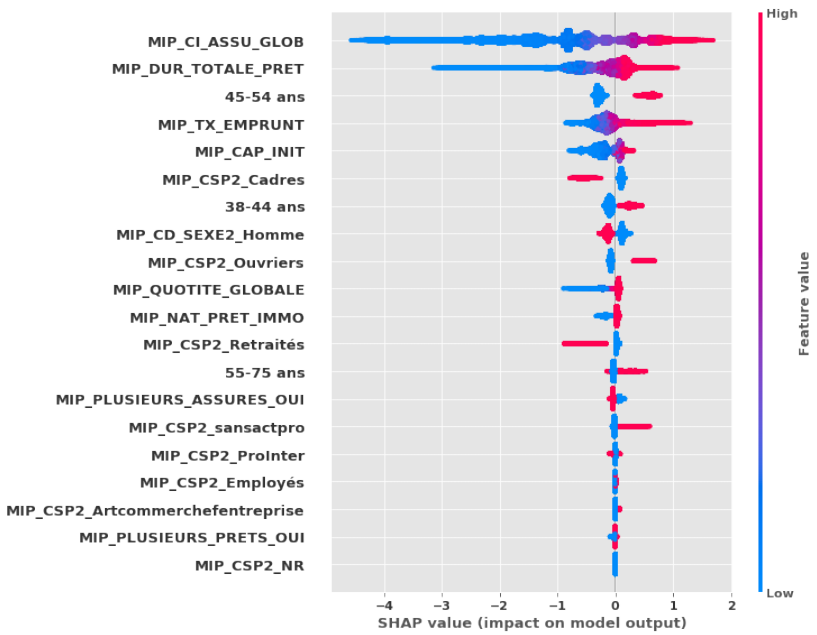
D'un autre côté, nous avons représenté les courbes de *Lorenz* ordonnées (voir A.39 ci-dessous) pour chacun de nos trois modèles. Ces graphiques donnent entre autres une représentation visuelle du pouvoir de discrimination du portefeuille par ces modèles. Comme indiqué ci-dessus,

le modèle *XGBoost* donne la plus grande aire sous la diagonale et donc discrimine le plus le portefeuille.



GRAPHIQUE A.39 – Courbes de Lorenz ordonnées par modèles sur les prestations en IT

En outre, nous avons analysé l'importance des variables dans la modélisation ainsi que les effets d'interactions entre les variables explicatives du modèle *XGBoost*. Le graphique A.40 illustre à la fois l'importance et les effets marginaux des variables explicatives sur le montant des prestations en IT des emprunteurs. Il en ressort de l'analyse de la Shap-importance que le capital assuré "MIP_CI_ASSU_GLOB" est la variable la plus importante dans la prédiction des prestations en IT du modèle *XGBoost*. Il s'en suit la durée totale du prêt, la tranche d'âge à la souscription [45 ; 55 ans), le taux d'emprunt, la somme initiale, les cadres, la tranche d'âge [38 ; 45 ans) et le sexe. L'analyse des effets marginaux montre que les emprunteurs appartenant à la tranche d'âge à la souscription [45 ; 55 ans) ou [38 ; 45 ans) ont des montants de prestations en IT plus importants que ceux de la tranche [18 ; 38 ans) (modalité de référence). Cependant les emprunteurs de plus de 55 ans ne présentent pas un effet interprétable sur la prestation. Nous remarquons aussi que les hommes ont des montants de prestations en IT plus faibles que les femmes. En ce qui concerne la catégorie socio-professionnelle, nous observons que les emprunteurs cadres et les retraités ont des montants de prestations plus faibles par rapport à la modalité de référence (emprunteurs agriculteurs). D'un autre côté, les emprunteurs ouvriers ont des montants de prestations plus élevés que les agriculteurs. Enfin, pour ce qui est de la durée totale du prêt, nous observons qu'elle impacte positivement le montant des prestations : plus elle est importante, plus le montant des prestations l'est aussi.



GRAPHIQUE A.40 – Effets marginaux des variables explicatives sur le montant des prestations

Impacts sur les provisions et la rentabilité

Les résultats de la modélisation des risques DC et IT nous ont permis de réaliser des projections en termes de provisionnement et de rentabilité de notre portefeuille d'étude.

Les tableaux ci-dessous représentent les projections de l'engagement de CNP Assurances pour les garanties DC et IT des prêts en cours dans notre portefeuille. Il ressort de l'analyse que le coût de la garantie DC pour les prêts en cours s'élève environ à 2 milliards d'euros. De plus, une analyse suivant les classes de risques montre que la classe 6 mobilise plus de provision avec 778 millions d'euros. La classe 1, au contraire, ne mobilise que 100 millions. Pour le risque IT, le coût global attendu sur la durée résiduelle des prêts est de 32 millions d'euros. Plus particulièrement, nous remarquons la classe 1 ne nécessite que 500 milles d'euros de provision alors que la classe 6 en nécessite 14 millions.

Segmentation	Effectif (%)	Engagement futur DC
Classe 1	8,61%	102 076 701,14
Classe 2	17,15%	115 060 258,39
Classe 3	18,29%	224 913 289,52
Classe 4	18,47%	208 354 164,18
Classe 5	16,28%	487 116 680,14
Classe 6	21,20%	778 628 914,50
Total	100,00%	1 916 150 007,87

TABLEAU A.7 – PSAP à l'ultime du risque DC pour les prêts en cours

Segmentation	Effectif (%)	Prestation versée	Engagement futur IT
Classe 1	8,61%	753 765,93	514 056,40
Classe 2	17,15%	4 107 939,44	1 398 944,00
Classe 3	18,29%	7 150 201,35	3 085 452,00
Classe 4	18,47%	15 711 631,43	5 754 540,00
Classe 5	16,28%	12 161 075,69	6 858 002,00
Classe 6	21,20%	12 630 094,54	14 546 080,00
Total	100,00%	52 514 708,38	32 157 074,40

TABLEAU A.8 – PSAP à l'ultime du risque IT pour les prêts en cours

D'un autre côté, nous avons calculé le *Loss Ratio* à l'ultime pour chacune de nos classes de risques et globalement.

Les résultats du calcul de la rentabilité à l'ultime sont renseignés dans le tableau ci-dessous. Il ressort de l'analyse de ce tableau que le *Loss Ratio* est globalement de 71,05%. En outre, une analyse suivant les classes de risques laisse entrevoir que la classe 6 a un *Loss Ratio* supérieur à 100% mais ceux des autres classes de risques restent inférieurs à 100%.

Segmentation	Effectif (%)	Prestation attendue	Prime attendue	Loss Ratio à l'ultime
Classe 1	8,61%	103 184 394,90	292 512 723,12	35,28%
Classe 2	17,15%	120 989 902,71	286 389 235,88	42,25%
Classe 3	18,29%	235 736 104,39	448 858 996,75	52,52%
Classe 4	18,47%	230 471 749,54	448 858 996,75	51,35%
Classe 5	16,28%	506 373 044,70	522 573 284,61	96,90%
Classe 6	21,20%	811 208 957,76	764 243 459,78	106,15%
Total	100,00%	2 007 964 154,00	2 825 391 159,73	71,07%

TABLEAU A.9 – Loss Ratio à l'ultime des prêts en cours

Executive summary

NIAKH Fallou

Apport du Machine Learning dans l'analyse multivariée de la sinistralité des contrats emprunteur

Preamble

In one of the group loan insurance portfolios held by CNP Assurances in collaboration with a partner bank covering the risks of death and disability, the existing pricing is of age range × loan capital range type. However, in addition to these two pricing variables, we also have at our disposal other variables such as the socio-professional category, the total duration of the loan, the insured capital, the insurance quota, etc. Therefore, it is worth asking whether these non-pricing variables contribute to the explanation of the sinistrality of the portfolio ?

For several decades, actuaries have developed traditional statistical methods to estimate these risks. Recently, faced with the high volume of data in companies, the considerable contribution of *Machine Learning* to the resolution of supervised learning problems is now a fact. Moreover, *Machine Learning* algorithms allow capturing the structure of information without making strong assumptions on the distributions of variables, contrary to traditional statistical methods. Therefore, *Machine Learning* algorithms can be more effective in capturing and modeling complex phenomena.

Aiming to optimize its risk monitoring process in loan insurance, CNP Assurances wishes to experiment with *Machine Learning* methods. This study originates from this preoccupation and focuses on two guarantees in loan contracts : death and work disability. It compares the robustness and performance of classical and *Machine Learning* methods applied to these two risks. To this end, two methods are used to model mortality in the loan contract : the classical model of *Cox*, *XGBoost Cox* and *XGBoost survival embedding* known as *Xgbse*. For work disability, we modeled the costs using three families of models : generalized linear models (*GLM Tweedie*),

aggregation methods (*XGBoost Tweedie*) and deep learning methods (*neural networks*). The results obtained on these two risks allowed us to propose alternative methods for reserving and calculating profitability in the portfolio.

Data presentation

For this study, we have a portfolio of group loan insurance contracts from a credit institution. The portfolio is made up of contracts linked to loans to finance real estate purchases or consumer loans covering the following risks : death and disability. The database contains 2.8 million loan/insured couples observed from 2007. We have the characteristics of the insured (age at subscription, gender, socio-professional category) and of the loan (initial capital, duration, insurance quota, start and end dates,...). Gender is used for academic purposes only. Our variables of interest in this study are : the indicator variables that specifies whether or not the borrower died or entered into disability during his or her loan duration, the duration spent alive in the loan, and the total disability costs paid out during a loan insurance contract.

Borrower segmentation by scoring

Before starting the mortality and disability risks modeling, we first tried to segment our portfolio using logistic regression. This segmentation is important insofar as it allows us to quickly detect the risk profiles of the portfolio. The binary variables of death and disability entry are used for this purpose. Using the estimates of the coefficients of these two models, we established a score grid that associates a number of points between 0 and 100 to each modality of each variable used. A borrower's overall risk is obtained by adding up his or her scores on death and disability risks. Thus, to each borrower is assigned a risk score between 0 and 200.

Once the score grid is established, we split our overall score into different risk classes to have different observed death and disability entry rates for each class (see figure below).

Segmentation	Proportion (%)	Death rate	Work disability rate
Class 1	17,31%	0,45%	0,79%
Class 2	15,76%	1,06%	1,12%
Class 3	16,82%	1,04%	1,44%
Class 4	17,35%	0,65%	2,81%
Class 5	17,29%	1,11%	3,63%
Class 6	15,47%	3,09%	6,31%

TABLEAU A.10 – Portfolio segmentation according to death and disability risks.

Mortality risk modeling

For a loan contract held for t years, it would be interesting to estimate the probability of a borrower dying in year t. This is the probability that a borrower during the duration of the loan dies in a given year. One way of estimating this probability is given in the approach to constructing the estimator of *Cox*, which is interested in the probability of survival. On the other hand, some methods of *Machine Learning* are adapted to estimate the probability of survival of *Cox*. We present below the results of the construction of the survival law of experience in the loan obtained with the model of *Cox* and *XGBoost survival embeding* (*Xgbse*).

Cox model

The proportional hazards model of *Cox* is a statistical technique that allows us to study the relationship between the survival of an inidvidual and several explanatory variables. In our case, we regressed the time spent alive in the loan as a function of borrower and loan characteristics. The validity of this model requires the verification of the fundamental hypothesis that risks are proportional. Our data do not verify this hypothesis even after stratification by age at loan subscription.

Therefore, we did not use the *Cox* model for the construction of our survival law of experience in the loan.

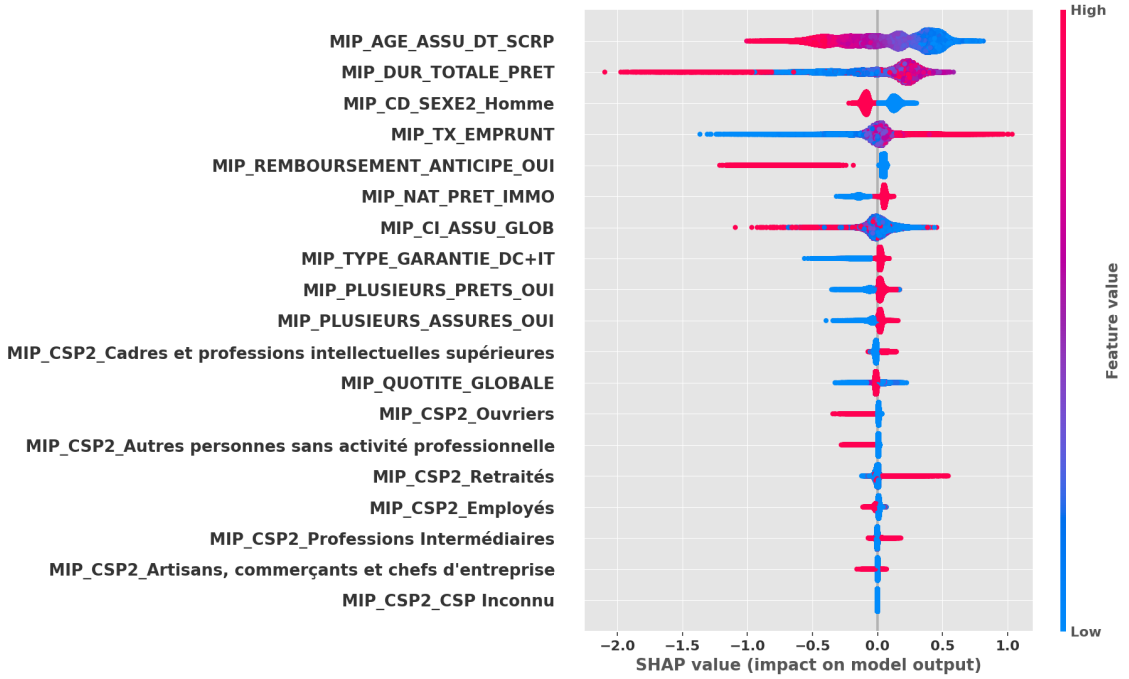
XGBoost Cox and *XGBoost survival embeding* models

The python module *Xgbse* (*XGBoost survival embedding*) is an alternative to the proportional hazards regression model of *Cox*. It combines the *XGBoost* method with a set of classical methods such as Kaplan Meier, k-nearest neighbors and logistic regression to estimate the survival probability. The advantage of this module over the classical *XGBoost* model is that it provides predictions of survival curves rather than point estimates. It also allows us to have an estimate of the confidence intervals and the calibrated expected survival times without bias. At the end of this experimentation, we found that the classical *XGBoost* model gives a slightly better performance compared to *Xgbse* when referring to the concordance index (see A.11).

Models	C-index	Brier Score
XGBoost	0.766	–
XGBSEDebiasedBCE	0.754	0.127
XGBSEKaplanNeighbors	0.740	0.136
XGBSEKaplanTree	0.732	0.154

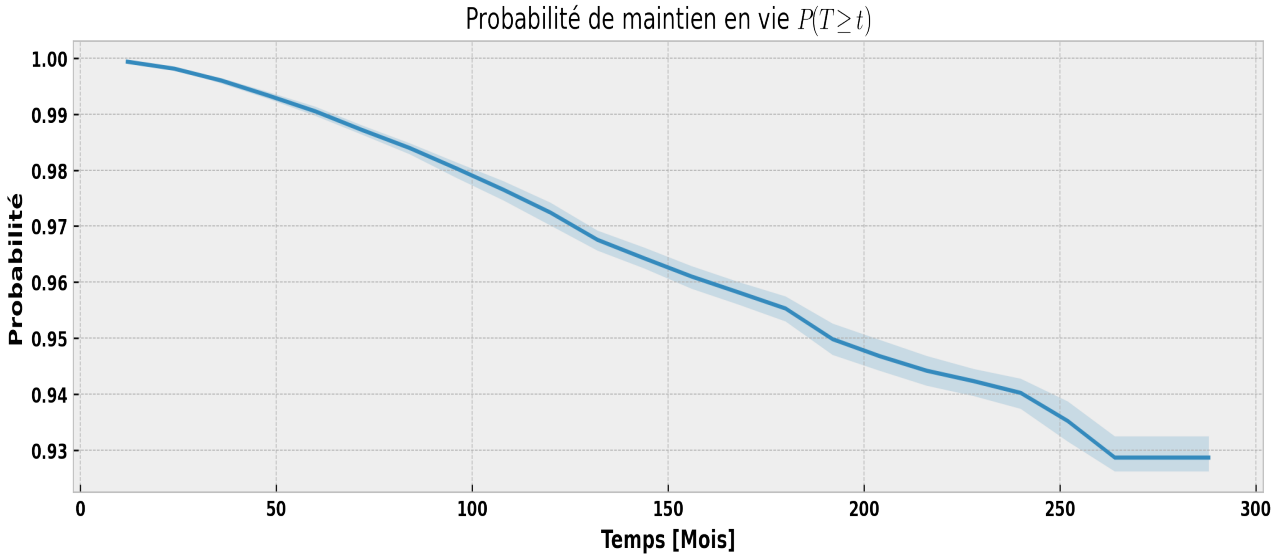
TABLEAU A.11 – Performance metrics for the same explanatory variables

From Figure A.41, the most important variables for the model *XGBoost* in explaining the time spent alive in the loan are : the age of the insured at the subscription of the loan, the duration of the loan, and gender.



GRAPHIQUE A.41 – Marginal effects of explanatory variables on mortality risk

Finally, since the classical *XGBoost* model does not directly yield an unbiased survival curve, we used the best model of the *Xgbse* family (*XGBSEDebiasedBCE*) to construct our survival law of experience in the loan.



GRAPHIQUE A.42 – Survival law of experience in the loan table using *XGBSEDebiasedBCE*

Work disability risk modeling

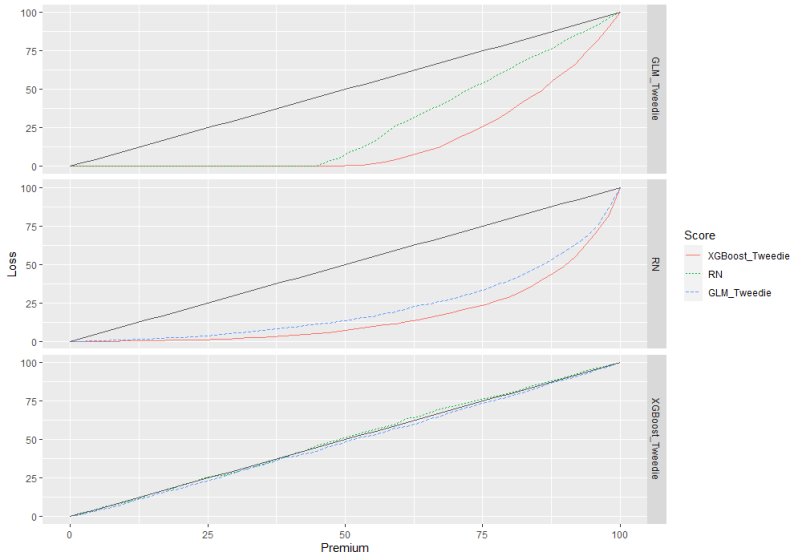
The modeling of the individual cost of the work disability guarantee is carried out in this master thesis by testing three families of methods : generalized linear models (*GLM Tweedie*), aggregation methods (*XGBoost*) and deep learning methods (*Neural Networks*). The choice of these three categories of models allows us to capture the contribution of *Machine Learning* methods in the multivariate analysis of claims in work disability of loan contracts. The comparison of these models is done using two performance criteria : *Gini index* and *Root Mean Square Error (RMSE)*. *Gini index* allows us to choose the model that best fits our data but also to measure the power of discrimination of high and low risks. We also use the module *SHAP* of Lundberg and Lee (2017) for the interpretation of our results.

At the end of our experiments, we observe on the table A.12 that the model *XGBoost* performs better.

	GLM Tweedie	*XGBoost*	*Neural Networks*
Gini index	0.671	0.7495	0.578
RMSE	20627.271	1045.703	1052.454

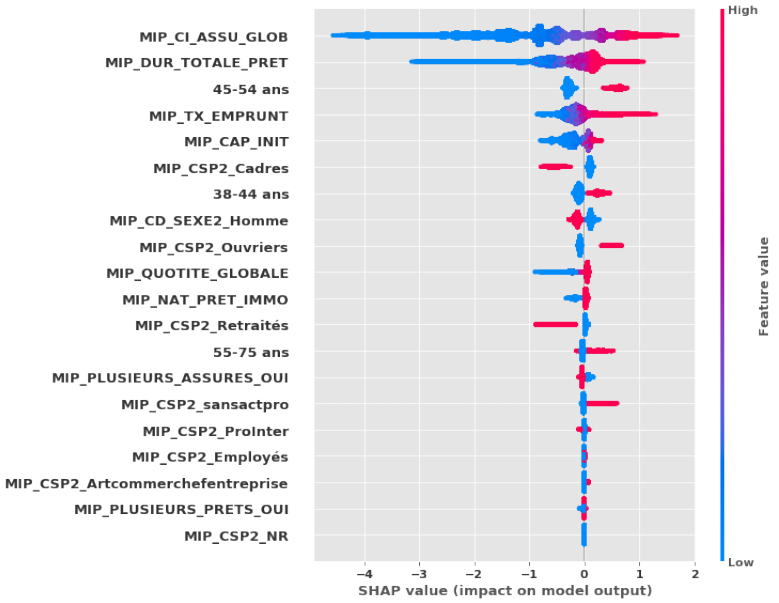
TABLEAU A.12 – Performance metrics for the same explanatory variables

Alternatively, we have represented the ordered *Lorenz* curves (see A.43 below) for each of our three models. These graphs provide, among other things, a visual representation of the portfolio discrimination power of these models. As shown above, the *XGBoost* model gives the largest area under the diagonal and thus most discriminates the portfolio.



GRAPHIQUE A.43 – Lorenz curves ordered by models on disability costs

Furthermore, we analyzed the importance of the variables in the model as well as the interaction effects between the explanatory variables in *XGBoost* model. The graph A.44 illustrates both the importance and the marginal effects of the explanatory variables on borrowers' disability costs. The Shap-importance analysis shows that the sum insured "MIP_CI_ASSU_GLOB" is the most important variable in predicting *XGBoost* model's disability costs. This is followed by total loan duration, age at subscription [45 ; 55), loan rate, borrowed sum, executives, age range [38 ; 45) and gender. Marginal effects analysis shows that borrowers in the [45 ; 55) or [38 ; 45)] age at subscription have larger amounts of disability costs than those in the [18 ; 38) (baseline condition). However, borrowers over age 55 do not have an interpretable effect on the disability costs. We also note that men have lower amounts of work disability costs than women. With respect to socio-professional category, we observe that executive and retired borrowers have lower costs compared to the reference modality (farmer borrowers). On the other hand, manual workers have higher costs than farmers. Finally, we observe that the total duration of the loan has a positive impact on the costs : the longer it is, the higher is the costs.



GRAPHIQUE A.44 – Marginal effects of explanatory variables on disability costs

Impacts on reserving and profitability

The results of death and disability risk modeling allowed us to make projections in terms of reserving and profitability of our study portfolio.

The tables below represent CNP Assurances' projected commitment for death and disability coverage for loans outstanding in our portfolio. According to the analysis, the cost of death cover for outstanding loans amounts to around 2 billion euros. Moreover, an analysis by risk

class shows that class 6 requires more provisions with 778 million euros. Class 1, on the other hand, mobilizes only 100 million. For disability risk, the expected overall cost over the remaining maturity of the loans is 32 million euros. More specifically, we note that class 1 requires a reserve of 500 thousand euros, while class 6 requires 14 million euros.

Segmentation	Proportion (%)	Future commitment Death
Class 1	8.61%	102 076 701.14
Class 2	17.15%	115 060 258.39
Class 3	18.29%	224 913 289.52
Class 4	18.47%	208 354 164.18
Class 5	16.28%	487 116 680.14
Class 6	21.20%	778 628 914.50
Total	100.00%	1 916 150 007.87

TABLEAU A.13 – PSAP at the ultimate risk of Death for outstanding loans.

Segmentation	Proportion (%)	Amount paid out	Future commitment Disability
Class 1	8.61%	753 765.93	514 056.40
Class 2	17.15%	4 107 939.44	1 398 944.00
Class 3	18.29%	7 150 201.35	3 085 452.00
Class 4	18.47%	15 711 631.43	5 754 540.00
Class 5	16.28%	12 161 075.69	6 858 002.00
Class 6	21.20%	12 630 094.54	14 546 080.00
Total	100.00%	52 045 689.53	32 157 074.40

TABLEAU A.14 – PSAP at the ultimate risk of Disability for outstanding loans.

Furthermore, we have calculated the ultimate loss ratio for each of our risk classes and globally.

Segmentation	Proportion (%)	Expected cost	Expected Premium	Ultimate Loss Ratio
Class 1	8.61%	103 184 394.90	292 512 723.12	35.28%
Class 2	17.15%	120 989 902.71	286 389 235.88	42.25%
Class 3	18.29%	235 736 104.39	448 858 996.75	52.52%
Class 4	18.47%	230 471 749.54	448 858 996.75	51.35%
Class 5	16.28%	506 373 044.70	522 573 284.61	96.90%
Class 6	21.20%	811 208 957.76	764 243 459.78	106.15%
Total	100.00%	2 007 495 135.14	2 825 391 159.73	71.05%

TABLEAU A.15 – Ultimate Loss Ratio for outstanding loans.

The results of the profitability calculation at the ultimate are shown in the table below. An analysis of this table shows that the overall loss ratio is 71.05%. Moreover, an analysis according to the risk classes shows that class 6 has a *Loss Ratio* higher than 100% but those of the other risk classes remain lower than 100%.