duas
Diplôme Universitaire des **Actuaires de Strasbourg**

**Université de Strasbourg**

*Mémoire d'actuariat - 2023*

Marie Le Pennec

---

# Détection et modélisation de la résiliation en Prévoyance Individuelle

---

BPCE ASSURANCES

*Tuteur académique*
Jean Bérard

*Tuteur entreprise*
Patrick Gomis

# Avant-propos

Pour des raisons de confidentialité, certains résultats chiffrés ont été modifiés sans perte de généralités.

Les calculs dans ce mémoire ont été réalisés à l'aide des langages R et Python.

# Résumé

*Mots clés : attrition, résiliation précoce, machine learning, modèles de survie, Kaplan-Meier, Hoem, Best Estimate, projection à moyen terme, Produit Net Bancaire*

L'objectif de ce mémoire est d'étudier le comportement de résiliation des contrats de type temporaire décès issus de la nouvelle offre Famille, ainsi que d'améliorer la modélisation de la résiliation dans le cadre des projections des provisions techniques *Best Estimate* et des projections à moyen terme.

Une analyse préliminaire de nos données montre une résiliation précoce importante des contrats Famille, avec une forte hétérogénéité par apporteur. Compte tenu du volume important de données à notre disposition, nous testons différents modèles de *machine learning* afin de trouver les facteurs explicatifs. Nous utilisons pour cela des données issues de l'entrepôt de données Prévoyance Individuelle, permettant une étude à la maille cohorte année-mois, et ayant trait aux caractéristiques des clients, de leurs contrats et de leurs conditions de souscription. Cela nous permet d'identifier de potentiels leviers d'action afin d'augmenter le taux de rétention des contrats Famille.

Après étude de la résiliation précoce, nous modélisons la résiliation à horizon durée de vie du contrat à l'aide de modèles classiques d'analyse de survie (Kaplan-Meier et Hoem). Les lois de résiliation sont construites sur les données des contrats Famille, par âge et par ancienneté. Bien que l'impact de la nouvelle loi de résiliation sur le *Best Estimate* sous Solvabilité II soit moindre, la loi de résiliation par ancienneté s'avère pertinente pour les projections à moyen terme, notamment pour modéliser l'écoulement des contrats en stock, le chiffre d'affaires et le Produit Net Bancaire à trois ans. Ainsi, des lois spécifiques par cohorte et apporteur pourraient être très intéressantes à ajouter en paramètres d'entrée des scénarios des plans à moyen terme.

# Abstract

*Keywords : attrition, early churn, machine learning, survival analysis, Kaplan-Meier, Hoem, Best Estimate, medium-term plans, Net Banking Income*

The aim of this thesis is to study the attrition behavior of term life insurance contracts from the new offer, and to improve the modeling of attrition in the context of technical reserve projections under Solvency II and medium-term projections.

A preliminary study of our data shows a high level of early termination of contracts, with considerable heterogeneity by bank. Given the large volume of data at our disposal, we test different machine learning models to find explanatory factors and detect early customer churn. To do so, we are using data from our data warehouse, which enables us to carry out a study at a year-month cohort level. The data we use for this study covers customer, contract and underwriting information. This enables us to identify potential levers of action to increase the retention rate of our contracts.

After studying early churn, we then analyze attrition over the lifetime of the contract, using classic survival analysis models (Kaplan-Meier and Hoem). Attrition rates are constructed by age and seniority.

Although the impact of the new attrition rates on Best Estimate reserves is not significant, the attrition rates by seniority are relevant for medium-term projections, notably for modeling the run-off of in-stock contracts, sales and Net Banking Income over three years. Thus, specific laws by cohort and bank could be very interesting for setting commercial and financial targets for medium-term plans.

# Remerciements

Je tiens tout d'abord à remercier chaleureusement Patrick GOMIS pour son excellent encadrement, sa très forte implication et ses conseils avisés. Son expertise et sa disponibilité continuelle ont été de véritables atouts pendant mon alternance et je lui en suis très reconnaissante.

J'adresse également mes sincères remerciements à tous les membres du Département Actuariat Produits Prévoyance et du Département Actuariat Produits ADE pour leur bienveillance et leurs encouragements précieux. Je remercie plus généralement tous les collaborateurs du Groupe BPCE avec qui j'ai pu échanger et qui ont pu m'aider pendant la réalisation de ce mémoire.

Je tiens à remercier Adrien MOYAUX, avec qui j'ai eu le plaisir de travailler pendant cette année d'alternance, pour son immense soutien autant sur le plan professionnel que personnel.

Un grand merci à mon tuteur pédagogique Jean BERARD pour son accompagnement tout au long de ce mémoire, ainsi qu'à tout le corps professoral du DUAS pour l'enseignement dispensé.

Je remercie également mes amies Pauline MILLE et Tiphaine HERSEMEULE pour leurs conseils et leurs relectures.

Enfin, je remercie mes parents pour leur soutien infaillible durant toutes mes études.

# Table des matières

# Introduction

L'attrition est le fait, pour un client (individu, entreprise...) de quitter un fournisseur de biens ou de services, une marque ou un produit [1]. Comme pour toutes les sociétés de services, la résiliation chez un assureur est au coeur de la stratégie marketing et est une mesure d'adéquation de l'offre produit à la demande. En effet, un taux d'attrition élevé constaté dès la première année de souscription constitue une anomalie du point de vue de l'assureur. Ce phénomène doit être détecté et les mesures correctrices recherchées.

L'un des objectifs du plan stratégique 2024 de BPCE Assurances est d'augmenter le taux d'équipement en assurance prévoyance des clients des Banques Populaires et Caisses d'Epargne. Un des leviers majeurs pour atteindre cet objectif est de limiter l'attrition des contrats prévoyance. Nous nous focaliserons sur les contrats de la nouvelle offre Famille. Ce sont des contrats de type temporaire décès qui représentent près de la moitié du portefeuille de prévoyance individuelle, dont le lancement est relativement récent (à partir de 2016), mais avec une production soutenue et une volumétrie significative.

Le risque comportemental de résiliation est également suivi de près par les actuaires qui modélisent les flux de trésorerie futurs à partir de modèles à états intégrant des hypothèses sur les résiliations probables, pour des besoins de tarification, de calcul de rentabilité et de production d'états réglementaires. Afin de répondre aux exigences des réglementations Solvabilité II et IFRS 17, les hypothèses de projection des flux de trésorerie doivent être *best estimate*, c'est-à-dire qu'elles doivent être les plus proches possible du risque intrinsèque du portefeuille.

En pratique, les lois d'expérience permettent d'obtenir cette meilleure estimation des hypothèses. Toutefois, la loi de résiliation actuellement utilisée dans le calcul des meilleures

1. Définition : Wikipédia

estimations des provisions techniques est unique pour tous les produits et tous les risques (décès toutes causes, décès accidentel, arrêt de travail...). Elle renvoie des résultats cohérents au global mais ne reflète pas l'hétérogénéité observée sur les produits présents dans le portefeuille prévoyance individuelle.

La résiliation est aussi un facteur important dans le cadre des projections à moyen terme (souvent trois ans). Ces projections présentent un intérêt dans la gestion des risques et la planification stratégique des compagnies d'assurance à un horizon faible. Elles prennent notamment en compte les évolutions du nombre de contrats actifs, des primes, ainsi que du Produit Net Assureur (PNA) et du Produit Net Bancaire (PNB). Le PNA et le PNB représentent la part des primes qui reviennent respectivement à l'assureur et à la banque ayant vendu le contrat. Ainsi, les projections à moyen terme permettent entre autres de communiquer aux Banques Populaires et Caisses d'Epargne une estimation de leur PNB futur sur des produits particuliers. Généralement, un taux de résiliation global est appliqué, mais nous aimerions mettre en place une modélisation plus fine .

Le but de ce mémoire est d'étudier le comportement de résiliation des contrats Famille et d'améliorer les lois de résiliation actuellement utilisées dans le modèle de projection des flux de trésorerie et dans les projections à moyen terme. Pour cela, nous implémenterons deux approches distinctes, l'une à base de techniques d'apprentissage automatique (*machine learning*), l'autre à partir de modèles de survie dans un cadre non paramétrique.

Tout d'abord, nous présenterons BPCE Assurances ainsi que les produits Famille sur lesquels nous nous concentrons dans ce mémoire. Puis, dans une première partie, nous détaillerons le traitement des données réalisé et les premiers résultats de la phase exploratoire par suivi de cohortes années-mois. En particulier, la détection d'une forte hétérogénéité dans la résiliation selon les établissements durant la première année de souscription sera analysée. Pour cela, nous mettrons en œuvre des méthodes d'apprentissage automatique supervisé pour essayer de dégager des facteurs explicatifs de cette attrition infra-annuelle. Ensuite, nous modéliserons la résiliation sur tout le cycle de vie du contrat grâce à des modèles de survie classiques (Hoem et Kaplan-Meier). Enfin, nous mesurerons l'impact des lois de résiliation construites sur les données des contrats Famille, notamment dans le calcul des provisions techniques sous Solvabilité II et dans le cadre de projection à moyen terme.

# Chapitre 1

# Contexte général

## 1.1 BPCE Assurances

BPCE Assurances, anciennement Natixis Assurances, appartient au Groupe BPCE issu de la fusion des réseaux Banques Populaires et Caisses d'Epargne en 2009. Le Groupe BPCE est actuellement le deuxième réseau bancaire en France[COR22] avec 36 millions de clients fin 2021. BPCE Assurances est en charge de la gestion des sujets d'assurance pour le Groupe BPCE. Les produits d'assurance commercialisés par le réseau BPCE couvrent les domaines de l'assurance individuelle, de l'assurance non-vie, ainsi que de la gestion privée. Ces offres s'adressent à diverses clientèles, incluant les particuliers, les professionnels, les professions libérales, les agriculteurs, les entreprises, ainsi que les associations. Fin 2021, BPCE Assurances faisait un chiffre d'affaires de 11 milliards d'euros et gérait 11,5 millions de contrats.

BPCE Assurances se distingue principalement en deux pôles :

- L'assurance non-vie (complémentaire santé, assurance automobile, multirisque habitation, garantie des accidents de la vie, assurance des équipements multimédias, protection juridique, assurance parabancaire, télésurveillance et assurance des professionnels)
- L'assurance de personnes (assurance vie, épargne, retraite, assurance décès, assurance dépendance et assurance des emprunteurs)

BPCE Assurances possède une filiale, BPCE Vie, qui traite entre autres des activités de prévoyance individuelle.

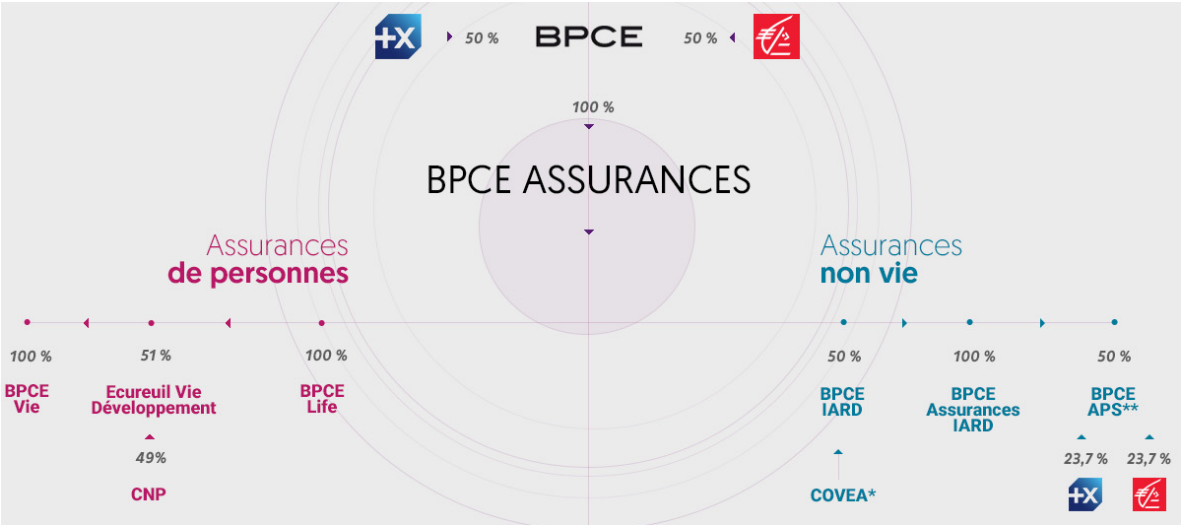


FIGURE 1.1 – Organigramme de BPCE Assurances

## 1.2 Les risques en Prévoyance Individuelle

Ce mémoire est mené au sein de la Direction Offre et Pilotage Commercial dont l'Actuariat Produits Prévoyance est un des départements. Il est important de définir les risques de l'assurance prévoyance. Les contrats de Prévoyance Individuelle s'inscrivent dans la Protection Sociale. Ils aident l'assuré à faire face aux conséquences financières de la survenance des risques suivants :

- Arrêt de travail (Incapacité Temporaire de Travail, Invalidité Permanente Partielle, Invalidité Permanente Totale)
- Perte Totale et Irréversible d'Autonomie (PTIA)
- Dépendance
- Décès à la suite d'une maladie ou d'un accident

La notion de maladie est définie comme tout problème de santé constaté par un médecin ou une autorité médicale. L'accident est défini quant à lui comme toute atteinte corporelle non provoquée par l'assuré mais provenant d'un facteur extérieur et donc étranger à la volonté de l'assuré.

**Arrêt de travail**

- Incapacité Temporaire de Travail (ITT) : l'assuré se retrouve temporairement dans l'impossibilité d'exercer son activité professionnelle. Au bout de trois ans passés dans cet état, l'assuré passe en Invalidité Permanente Totale (IPT).

- Invalidité Permanente Partielle (IPP) : l'assuré est en IPP lorsqu'il se trouve, à la suite d'une maladie ou d'un accident, dans un état physique ou mental le mettant dans l'impossibilité totale, permanente et présumée définitive d'exercer son activité professionnelle ou toute autre activité rémunératrice. Le taux d'incapacité, observé par un médecin, est compris entre 33 % et 66 %.
- Invalidité Permanente Totale (IPT) : l'assuré est en IPT s'il est dans l'impossibilité totale et définitive d'exercer une activité professionnelle avec un taux d'incapacité supérieur à 66 %. L'assuré entre également en IPT après trois ans consécutifs en ITT.

**Perte Totale et Irréversible d'Autonomie (PTIA)**

La Perte Totale et Irréversible d'Autonomie concerne toute invalidité, constatée par un médecin, physique ou mentale de l'assuré causant une incapacité totale et définitive d'exercer son activité professionnelle ou toute autre activité rémunératrice. L'assuré est dans l'obligation de recourir à l'assistance d'une tierce personne pour accomplir les actes de la vie quotidienne, et ce de façon constante pour le reste de sa vie.

**Décès**

Il existe deux garanties pour le risque de décès : « décès toutes causes » et « décès accidentel ». La garantie « décès toutes causes » entre en jeu pour n'importe quelle cause de décès : maladie, accident, mort naturelle, suicide. La garantie « décès accidentel » n'intervient qu'en cas de décès accidentel dont la définition varie selon les contrats d'assurance.

**Dépendance**

La dépendance partielle ou totale de l'assuré fait référence à l'état d'une personne qui ne peut plus agir en toute autonomie pour effectuer des gestes quotidiens.

## 1.3 L'offre Famille

Les produits de Prévoyance Individuelle sont rassemblés en familles de produits, selon leur nature, leurs conditions générales de vente et leur gestion. Dans ce mémoire, nous nous intéresserons aux familles de produits correspondant à des offres dites «Famille». La nouvelle offre Famille de BPCE Assurances est à destination des clients souhaitant se prémunir des conséquences financières d'une maladie redoutée ou d'une PTIA et protéger leurs proches en cas de décès. Elle comporte deux familles de produits : SECUR'Famille et Assurance Famille. La famille de produits SECUR'Famille est elle-même composée de deux produits appelés SECUR'Famille et SECUR'Famille 2. La famille de produits Assurance Famille n'est composée que du produit du même nom.

### 1.3.1 SECUR'Famille

SECUR'Famille était un produit distribué par les Caisses d'Epargne entre 2016 et 2020. Il s'agit d'un contrat d'assurance temporaire décès d'une durée d'un an, renouvelable par tacite reconduction. Il peut être souscrit par toute personne physique valide entre 18 et 64 ans. Ce contrat prévoit le versement d'un capital d'au moins 15 000€ en cas de décès ou de PTIA, avec possibilité d'indemnisation sous forme de rente éducation d'un minimum de 600€ par an pour le décès. Il est également possible qu'en cas de maladie redoutée [1], 20% du capital choisi précédemment soit versé à l'assuré, les 80% restants étant versés en cas de décès ou de PTIA. Une garantie assistance est aussi présente.

### 1.3.2 Assurance Famille et SECUR'Famille 2

Les produits Assurance Famille et SECUR'Famille 2 sont commercialisés depuis 2019 dans les Banques Populaires et depuis 2020 dans les Caisses d'Epargne respectivement. Ces deux produits sont identiques, seul leur réseau de distribution diffère. Comme SECUR'Famille, ce sont des contrats d'assurance temporaire décès d'un an renouvelables par tacite reconduction. Lors de l'adhésion, il est possible de choisir parmi trois formules :

- Formule « Essentiel » : accessible à toute personne physique valide entre 18 et 85 ans. Cette formule prévoit le versement d'un capital d'au moins 20 000€ en cas de décès ou de PTIA. En cas de décès, il est possible que le capital soit versé sous la forme d'une rente éducation, ou d'un panachage entre capital et rente éducation si cette dernière s'avère supérieure à 600€ par an. Une avance de fonds de 5 000€ peut également être versée sous 48 heures sous certaines conditions (elle est ensuite à déduire du montant du capital garanti).
- Formule « Confort » : accessible à toute personne physique valide entre 18 et 69 ans. Cette formule reprend les garanties de la formule précédente, avec en supplément la possibilité de verser 20% du capital garanti à l'assuré en cas de maladie redoutée. Les 80% restants seront versés en cas de décès ou de PTIA.
- Formule « Premium » : accessible à toute personne physique valide entre 18 et 69 ans. Cette formule reprend les garanties de la formule précédente avec en supplément le doublement du capital garanti et/ou de la rente éducation en cas de décès ou PTIA suite à un accident.

Une garantie assistance est présente peu importe la formule choisie par l'assuré. Une formule « Accident » existe également en cas d'impossibilité d'assurer le décès toutes causes.

1. Les maladies redoutées sont les suivantes : infarctus du myocarde, chirurgie des artères coronaires, accident vasculaire cérébral, cancer, hémiplégie, paraplégie, tétraplégie et brûlures graves.

# Chapitre 2

# Analyse exploratoire

## 2.1 Contexte

Un des leviers priorisés dans le plan stratégique de BPCE Vie est d'augmenter le taux d'équipement en prévoyance individuelle, notamment la nouvelle offre Famille, qui représentait 48 % du portefeuille de Prévoyance Individuelle fin 2022. Deux risques *business* sont alors à surveiller et prévenir : une production d'affaires nouvelles insuffisante et une résiliation trop élevée. L'objectif de ce chapitre est d'étudier la résiliation des nouveaux produits Famille, déterminer les facteurs pouvant expliquer cette résiliation afin d'améliorer le taux de rétention, et également nous assurer que la nouvelle offre Famille est correctement calibrée et cible la bonne population.

En effet, la Directive sur la Distribution d'Assurance (DDA) en vigueur depuis le 1er octobre 2018 exige que les produits d'assurance soient conçus de manière à répondre aux besoins et aux intérêts des cibles de marché que les concepteurs ont préalablement identifiées. Le marché cible est défini dans le règlement délégué POG (*Product Oversight and Governance*) comme «un groupe de clients partageant des caractéristiques communes à un niveau abstrait et généralisé, dans le but de permettre au concepteur d'adapter les particularités du produit aux besoins, caractéristiques et objectifs de ce groupe de clients»[Jou17].

Concernant les distributeurs d'assurance, la DDA indique également[LE 18] qu'ils doivent identifier les besoins et exigences des clients avant de leur proposer un produit d'assurance. Cela signifie que les offres doivent être proposées en fonction des informations spécifiques fournies par les clients concernant leur situation financière, leurs objectifs et leurs besoins.

La directive vise également à prévenir les conflits d'intérêt et à éviter que les distributeurs d'assurance ne poussent les clients à souscrire des produits qui ne leur conviennent pas. Ils doivent donc fournir des informations claires et objectives sur les produits d'assurance afin que les clients puissent prendre une décision en toute connaissance de cause.

Notre objectif est donc de mener une étude afin de découvrir des caractéristiques pouvant expliquer la résiliation des contrats de la nouvelle offre Famille. Cette étude est menée au sein de la Direction Offre et Pilotage Commercial, et plus précisément dans le Département Actuariat Produits Prévoyance, dont les missions sont entre autres la tarification, la rentabilité des différents produits de prévoyance, ainsi que la construction de lois d'expérience biométriques et comportementales. Parmi ces études comportementales, la résiliation est un sujet majeur.

Il est important de préciser que la résiliation évoquée dans ce mémoire renvoie à une résiliation suite à la demande du client ou d'une résiliation suite à une mise en demeure pour non-paiement de la prime d'assurance. Les contrats abandonnés ou sans suite, que ce soit du fait du client ou de l'assureur, ne sont pas compris dans cette définition car ils n'ont jamais donné lieu à un versement de prime. En cas de défaut de paiement de prime, BPCE Vie en informe l'adhérent par lettre recommandée de mise en demeure envoyée au plus tôt 10 jours après son échéance. L'adhérent dispose alors d'un délai de 40 jours à compter de l'envoi de cette lettre pour régulariser le paiement de ses primes. Le défaut de paiement à l'expiration de ce délai entraînera la résiliation de l'adhésion et la fin de couverture. Dans le cas d'une résiliation à la demande de l'adhérent, celui-ci doit adresser à BPCE Vie une lettre recommandée avec accusé de réception au plus tard un mois avant l'échéance périodique de cotisation. L'adhésion cesse à l'échéance de cotisation qui suit la demande de résiliation de l'adhésion et les garanties cessent à compter de cette même date.

Nous démarrons par une phase exploratoire des données. Cette étape très chronophage débouchera sur l'énoncé de la problématique détectée et une présentation des outils et méthodes retenus pour l'étudier.

## 2.2 Les données

### 2.2.1 Le Data Warehouse Prévoyance Individuelle (DWH PI)

Le Data Warehouse Prévoyance Individuelle (appelé DWH PI ou DWH par la suite) est la source de données que nous utiliserons tout au long de cette étude.

Un *data warehouse*, ou entrepôt de données, est[Ora23] une base de données relationnelle hébergée sur un serveur dans un Data Center ou dans le Cloud. Il recueille des données provenant de sources variées et hétérogènes dans le but principal de soutenir l'analyse et faciliter la prise de décision.

Les données des Banques Populaires et des Caisses d'Epargne sont stockées dans plusieurs infocentres, puis centralisées dans un *Operational Data Store* (ODS) Oracle. La rationalisation de ces infocentres concernant le domaine Prévoyance Individuelle a fait l'objet d'une étude de cadrage, suivie par le lancement de la mise en place du DWH PI en 2020. La création du DWH PI vise à optimiser l'utilisation du système d'information en créant des vues métier à partir des données provenant des différents infocentres.

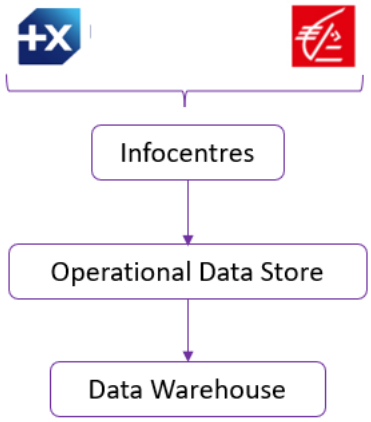


FIGURE 2.1 – Arborescence du DWH PI

L'IT s'occupe de développer et mettre à disposition les vues métiers, les diagrammes relationnels entre les différentes vues ainsi qu'un dictionnaire de données. Les diagrammes relationnels des vues du DWH PI utilisées durant ce mémoire sont disponibles en annexe.

Avoir un *data warehouse* à disposition nous permet d'avoir un plus grand contrôle et accès à nos données. Cela s'accompagne en revanche d'un travail plus important en amont de chaque étude. En effet, des jointures entre différentes vues du DWH PI sont très souvent nécessaires, ce qui n'est pas facilité par la volumétrie très importante des données. La préparation et le traitement des données que nous évoquerons ensuite ont donc été très chronophages.

### 2.2.2 Préparation des données

Pour cette étude, les données proviennent en majeure partie du DWH PI. Ce choix a été fait dans le but d'en tirer profit au maximum.

Nous avons à disposition les données relatives aux contrats de la nouvelle offre Famille, historisées chaque mois depuis avril 2019, dans la vue nommée Suivi Contrats. Afin de pouvoir suivre l'évolution des contrats depuis leur souscription, nous gardons uniquement les données des contrats souscrits à partir d'avril 2019. Ces données ont été extraites en juin 2023 et représentent plus de **37 millions de lignes** et près de 1,4 million de contrats. Ainsi pour chaque contrat et pour chaque mois, nous connaissons entre autres l'état du contrat (en cours, résilié, clôturé...) et le montant de l'échéance de prime.

Nous allons ajouter d'autres variables présentes dans différentes vues du DWH PI dont nous pensons qu'elles pourraient aider à expliquer la résiliation des contrats. Voici le schéma des jointures effectuées avec ces différentes vues ainsi que les variables récupérées dans chaque vue :

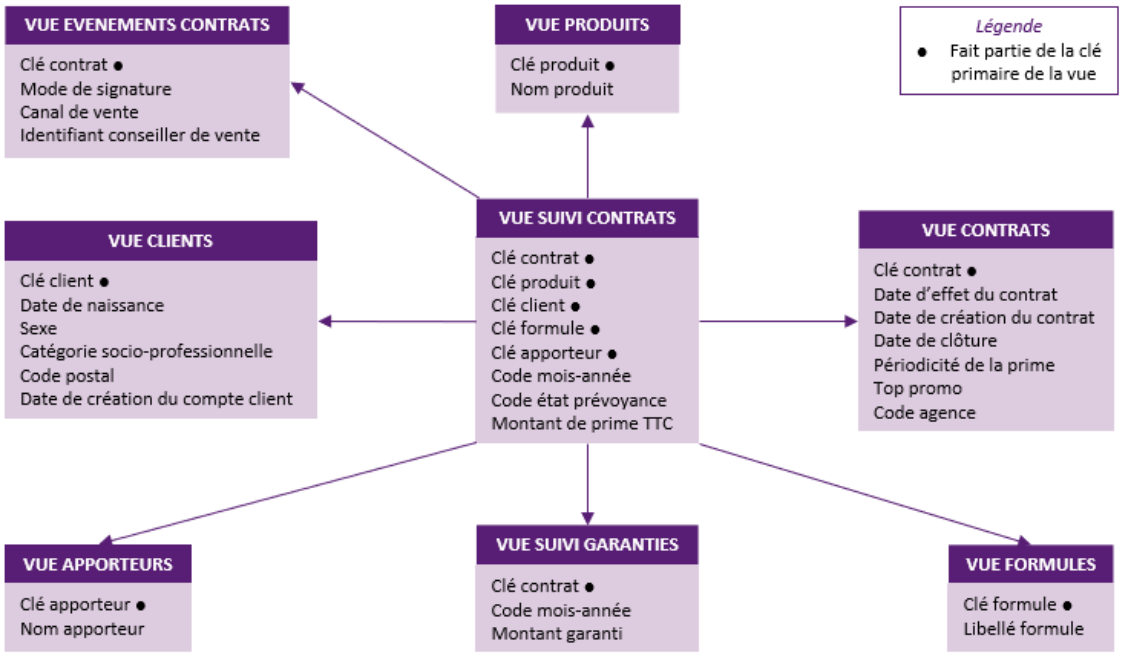


FIGURE 2.2 – Schéma des vues du DWH PI utilisées

La vue Évènements Contrats a nécessité une analyse plus poussée. Cette vue recueille les données relatives aux mouvements de chaque contrat (souscription, augmentation de tarif, clôture, ...). Nous devons nous assurer que tous les contrats se trouvant dans la vue Contrats se trouvent également dans la vue Évènements Contrats. Les statistiques suivantes portent sur tous les contrats de Prévoyance Individuelle.

| Année de création des contrats | Taux de correspondance des contrats |
|---|---|
| 2016 | 64.5% |
| 2017 | 62.3% |
| 2018 | 69.9% |
| 2019 | 100% |
| 2020 | 100% |
| 2021 | 100% |
| 2022 | 100% |
| 2023 | 100% |

TABLE 2.1 – Taux de correspondance entre les vues Évènements Contrats et Contrats

Nous observons que les données sont complètes pour les contrats souscrits en 2019 et après. Une étude sera menée sur l'incomplétude des contrats souscrits avant 2019. Ayant restreint notre périmètre sur les contrats Famille souscrits à partir d'avril 2019, nous pouvons en conclure que les données de la vue Évènements Contrats correspondantes sont bien présentes. Afin d'obtenir les données relatives à la souscription des contrats, il faut dédoublonner la vue Évènements Contrats. Le dédoublonnage se fait en prenant pour chaque contrat les informations correspondant à la version de contrat la plus ancienne.

Nous pouvons désormais effectuer les jointures entre toutes les vues. Chaque jointure a été vérifiée afin de s'assurer qu'aucun doublon n'a été introduit. Quelques retraitements de données s'imposent : les dates de création de contrat aberrantes sont supprimées et lorsque la date de création est supérieure à la date d'effet du contrat, cette dernière est considérée comme étant la vraie date de création du contrat. Il est également à noter que les dates exactes de résiliation des contrats sont inconnues, nous considérerons donc que la date de clôture d'un contrat résilié est sa date de résiliation.

Présentons rapidement les variables que nous avons rassemblées.

| **Variable** | **Détails** |
|---|---|
| Code mois-année | Mois et année d'observation du contrat |
| Code état prévoyance | État du contrat pour le mois d'observation. Les principaux sont « En cours », « En attente », « Sans effet », « Sans suite », « En contentieux », « Résilié » et « Décès » |
| Montant de prime | Prime annuelle TTC pour le mois d'observation |
| Montant garanti | Capital garanti du contrat pour le mois d'observation |
| Nom produit | « SECUR'Famille », « SECUR'Famille 2 » ou « Assurance Famille » |
| Date de création du contrat | |
| Date de clôture du contrat | Égale à la date de décès pour les contrats clos suite à un sinistre et à la date de résiliation pour les contrats résiliés |
| Périodicité de la prime | « A », « S », « T », « M » pour les primes annuelles, semestrielles, trimestrielles et mensuelles respectivement |
| Top promo | Vaut 1 si le contrat a été souscrit dans le cadre d'une promotion |
| Code agence | Code de l'agence dans laquelle le contrat a été souscrit |

| | |
|---|---|
| Formule | « Essentiel », « Confort » ou « Premium ». En cas d'impossibilité d'assurer (risque aggravé de santé), il existe également une formule « Accident » qui ne couvre que le risque Décès Accidentel |
| Apporteur | Nom de la Banque Populaire ou Caisse d'Epargne dont le contrat dépend |
| Canal de vente | Vaut « F » si la vente du contrat s'est effectuée en face à face (vente directe) et « D » si la vente s'est effectuée à distance |
| Mode de signature | Vaut « P » en cas de signature papier et « E » en cas de signature électronique |
| Identifiant conseiller de vente | |
| Date de naissance du client | |
| Sexe du client | « M » pour les hommes et « F » pour les femmes |
| CSP du client | |
| Code postal du client | |
| Date de création du compte client | |

TABLE 2.2 – Liste des variables extraites du DWH PI

Une fois nos données rassemblées, nous ajoutons les variables suivantes :

- Cohorte mois-année, obtenue avec le mois et l'année de souscription du contrat
- Ancienneté contrat, qui correspond au nombre de mois entre la cohorte mois-année du contrat et le code mois-année de la vue Suivi Contrats
- Âge du client à l'adhésion
- Ancienneté client, qui correspond au nombre de mois entre la date de création du compte client et la date de création du contrat
- Région, obtenue avec le code postal du client et correspond à la région où celui-ci habite
- Top prévoyance, qui vaut 1 lorsque le client possède plusieurs contrats de prévoyance individuelle et 0 sinon

Nous enlevons les lignes où l'âge du client à l'adhésion a une valeur aberrante (inférieur à 18 ans ou supérieur à 85 ans). Pour les produits Famille, il convient de préciser que le souscripteur/client et l'assuré sont la même personne.

## 2.3 Étude préliminaire

### 2.3.1 Statistiques descriptives

Maintenant que nos données ont été rassemblées, nous allons pouvoir présenter quelques statistiques descriptives. Regardons tout d'abord la répartition des contrats Famille par produit.

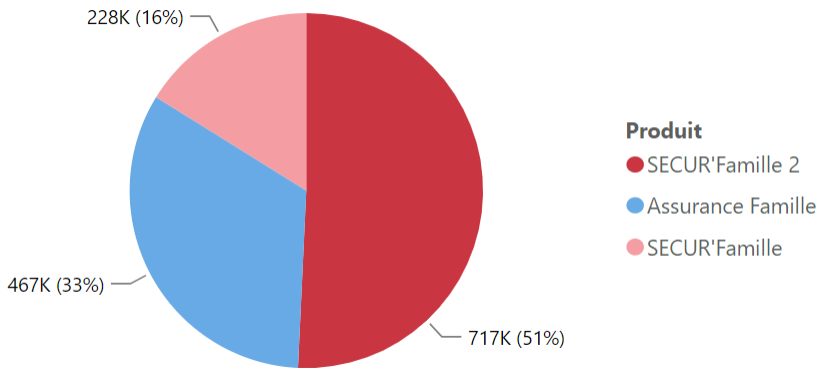


FIGURE 2.3 – Nombre de contrats par produit

La majorité des contrats Famille appartiennent au réseau Caisse d'Epargne. Le nombre de contrats SECUR'Famille n'est pas très élevé car ce produit a cessé d'être commercialisé en 2020 pour être remplacé par SECUR'Famille 2. Intéressons-nous maintenant à la proportion d'hommes et de femmes entre les deux réseaux.

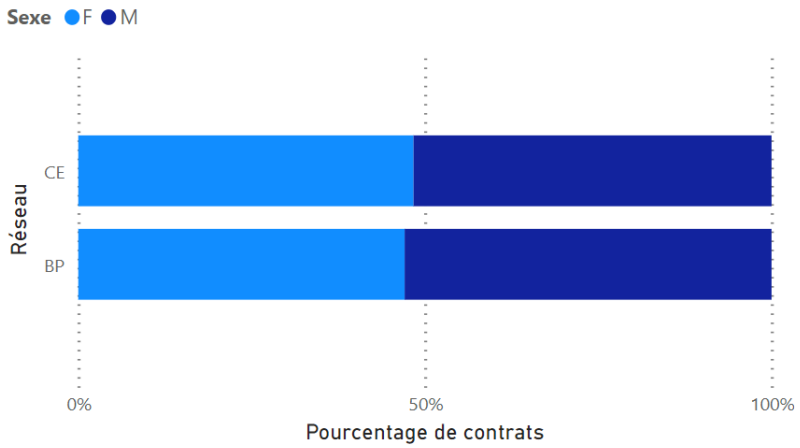


FIGURE 2.4 – Proportion de contrats par sexe et par réseau

Les proportions d'hommes et de femmes pour chaque réseau sont quasiment identiques. Regardons l'évolution du nombre de contrats en cours et résiliés, ainsi que l'évolution du taux de résiliation du portefeuille Famille au cours du temps.

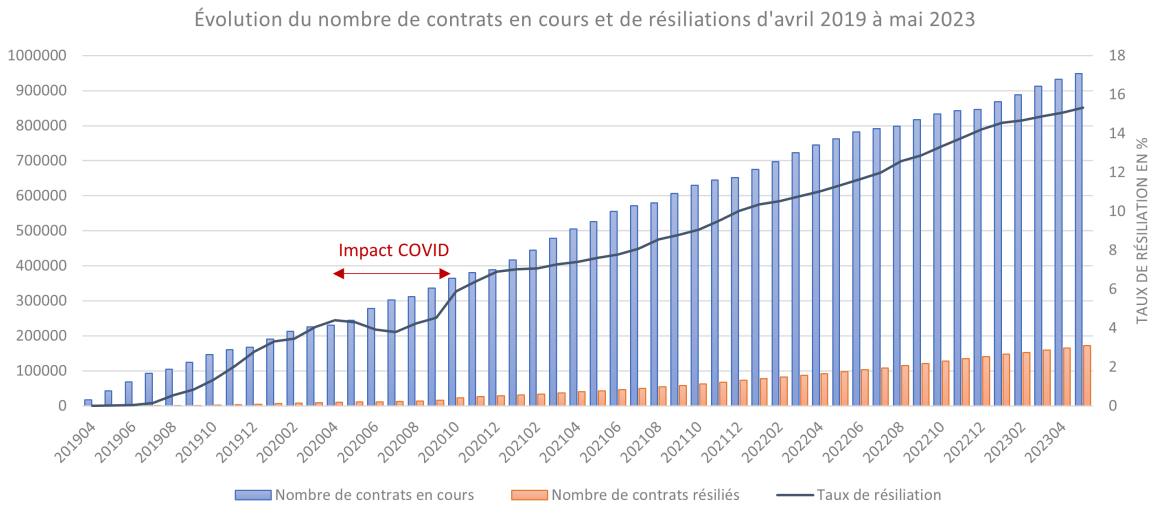


FIGURE 2.5 – Croissance du portefeuille Famille

Nous pouvons observer la croissance du portefeuille Famille depuis avril 2019 (pour des contrats souscrits à partir d'avril 2019). Le nombre de résiliations suit la même tendance que le nombre de contrats en cours, sauf vers mi-2020 où l'impact de la COVID-19 est clairement visible sur le taux de résiliation global.

Les statistiques suivantes sont obtenues avec les données des contrats Famille datant du 31 mai 2023. Regardons la répartition des contrats par état de contrat.

| État de contrat | Proportion de contrats |
|---|---|
| En cours | 71,9% |
| En attente | 0,1% |
| Sélection médicale | 0,5% |
| Sans effet | 8,8% |
| Sans suite | 5,1% |
| En contentieux | 0,6% |
| Résiliation MED | 9,5% |
| Résiliation client | 3,5% |
| Décès ou IAD | 0,1% |

TABLE 2.3 – Proportion de contrats par état de contrat au 31 mai 2023

Sans prendre en compte les contrats n'ayant jamais été actifs (sans effet et sans suite), le taux de résiliation global du portefeuille Famille est de **15%** , ce qui représente environ 172 000 contrats. 56% des contrats résiliés ont été souscrits par des hommes et près de la moitié des contrats résiliés au 31 mai 2023 sont des contrats appartenant à la formule « Essentiel ».

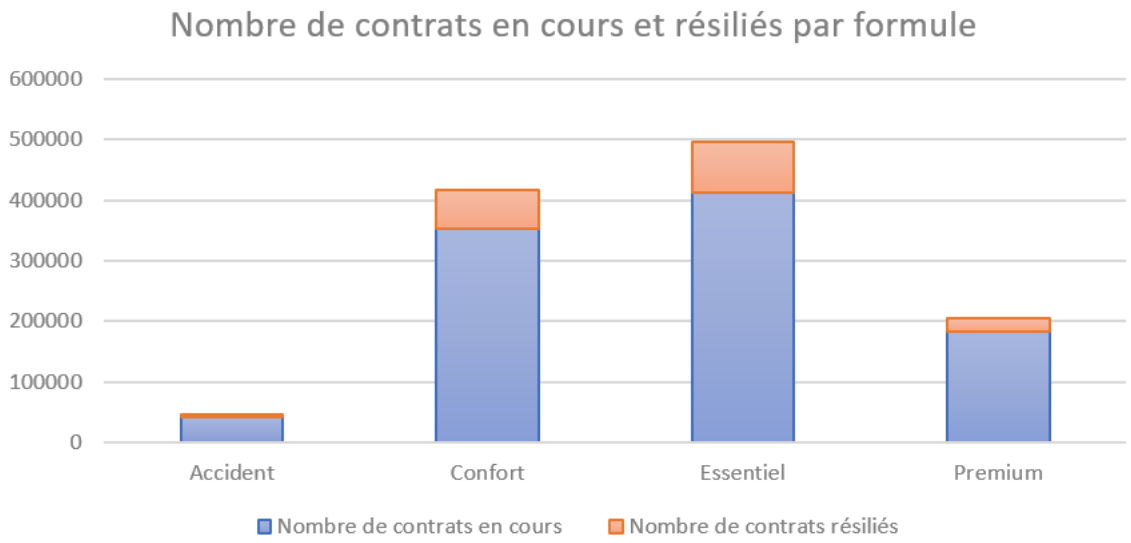


FIGURE 2.6 – Nombre de contrats en cours et résiliés par formule

Regardons le taux de résiliation par apporteur.

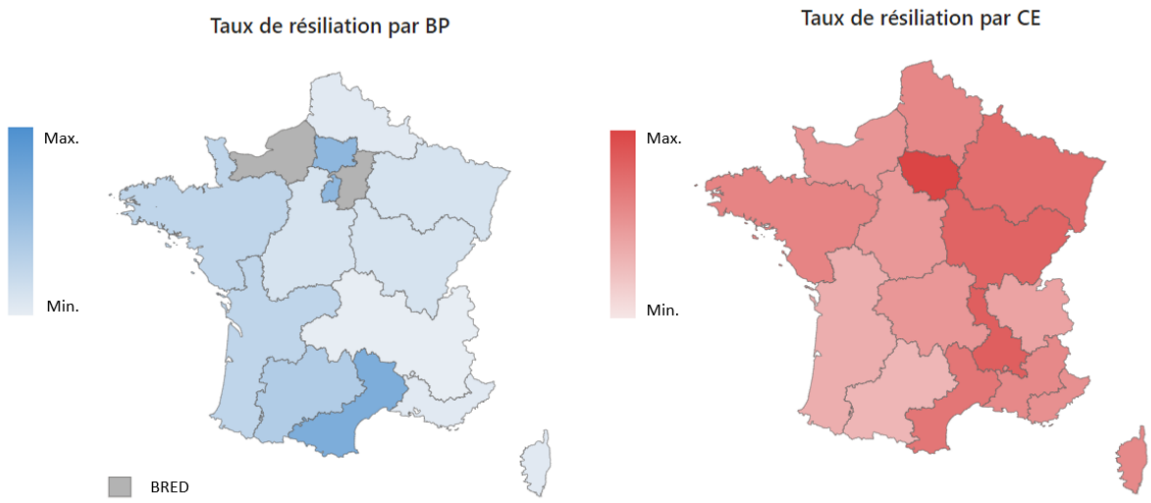


FIGURE 2.7 – Taux de résiliation par apporteur

Sur les deux cartes de la figure 2.7, les fonds de carte ont été créés pour coller au maillage de chaque réseau bancaire. Ainsi, chaque polygone correspond à une Banque Populaire (à gauche) ou une Caisse d'Epargne (à droite) régionale. Les valeurs des taux ne sont pas affichées pour cause de confidentialité. La zone grisée sur la carte des Banques Populaires correspond à la BRED, qui ne vend pas de contrats Famille. Nous pouvons voir que les taux de résiliation sont très différents pour les deux réseaux. Le réseau Caisse d'Epargne, plus étendu que le réseau Banque Populaire, semble également avoir des taux de résiliation plus élevés. Pour cette raison, les statistiques suivantes sont séparées entre réseau Banque Populaire et réseau Caisse d'Epargne.

| **Statistique** | **BP** | **CE** |
|---|---|---|
| Durée de détention moyenne des contrats | 15 mois | 13 mois |
| Durée de détention médiane des contrats | 12 mois | 11 mois |
| Âge moyen à la résiliation | 34 ans | 32 ans |
| Âge médian à la résiliation | 32 ans | 29 ans |
| Prime annuelle moyenne | 60€ | 54€ |
| Prime annuelle médiane | 33€ | 30€ |
| Capital moyen | 31 800€ | 30 500€ |
| Capital médian | 24 000€ | 25 000€ |

TABLE 2.4 – Statistiques par réseau sur les contrats Famille résiliés au 30 mai 2023

La moitié des contrats résiliés ont un an d'ancienneté ou moins et correspondent à des primes et capitaux garantis peu élevés. Les clients qui résilient sont jeunes (la moitié a moins de 32 ans). Ces statistiques sont en accord avec les taux de résiliation des contrats SECUR'Famille et SECUR'Famille 2 plus élevés que ceux des contrats Assurance Famille.

### 2.3.2 Définition de la problématique

D'après nos statistiques descriptives, la résiliation à un an des contrats Famille est loin d'être négligeable. En effet, la moitié des contrats Famille résiliés sont détenus pendant moins d'un an. Il peut être intéressant de regarder l'évolution moyenne de la proportion des contrats Famille par état de contrat et par mois d'ancienneté. Les statistiques suivantes ont été obtenues en calculant la moyenne des pourcentages sur toutes les cohortes de contrats.

| | Ancienneté du contrat (en mois) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **État du contrat** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 12 | 18 | 24 | 36 |
| En cours | 83,8 | 80,7 | 78,5 | 77,8 | 77,0 | 76,3 | 75,8 | 71,7 | 68,7 | 69,2 | 68,7 |
| En attente | 5,4 | 0,9 | 0,3 | 0,3 | 0,2 | 0,1 | 0,1 | 0,1 | 0,0 | 0,0 | 0,0 |
| Sélection médicale | 7,7 | 5,6 | 2,5 | 1,3 | 0,4 | 0,2 | 0,2 | 0,1 | 0,1 | 0,1 | 0,1 |
| Sans effet | 2,0 | 7,7 | 10,1 | 10,2 | 10,2 | 10,3 | 10,3 | 10,3 | 10,3 | 10,0 | 9,8 |
| Sans suite | 0,6 | 4,4 | 7,8 | 8,9 | 9,8 | 10,0 | 10,1 | 10,3 | 10,5 | 6,4 | 0,0 |
| En contentieux | 0,6 | 0,8 | 0,8 | 1,3 | 1,3 | 1,2 | 1,1 | 0,8 | 0,5 | 0,5 | 0,4 |
| Résiliation MED | 0,0 | 0,0 | 0,0 | 0,4 | 1,1 | 1,7 | 2,3 | 5,3 | 7,7 | 9,9 | 14,2 |
| Résiliation client | 0,0 | 0,0 | 0,1 | 0,2 | 0,2 | 0,3 | 0,4 | 1,5 | 2,2 | 3,7 | 6,6 |
| Décès ou IAD | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,1 | 0,1 | 0,1 | 0,2 |

FIGURE 2.8 – Évolution moyenne de la proportion de contrats (%) - Réseaux BP et CE

| | Ancienneté du contrat (en mois) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **État du contrat** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 12 | 18 | 24 | 36 |
| En cours | 85,9 | 91,7 | 95,5 | 95,6 | 96,0 | 95,6 | 94,9 | 90,1 | 86,6 | 82,9 | 76,1 |
| En attente | 5,5 | 1,0 | 0,4 | 0,4 | 0,2 | 0,1 | 0,1 | 0,1 | 0,0 | 0,0 | 0,0 |
| Sélection médicale | 7,9 | 6,4 | 3,0 | 1,6 | 0,5 | 0,3 | 0,3 | 0,1 | 0,1 | 0,2 | 0,1 |
| En contentieux | 0,6 | 0,9 | 0,9 | 1,6 | 1,6 | 1,5 | 1,4 | 1,1 | 0,7 | 0,6 | 0,5 |
| Résiliation MED | 0,0 | 0,0 | 0,0 | 0,5 | 1,4 | 2,1 | 2,9 | 6,7 | 9,7 | 11,9 | 15,7 |
| Résiliation client | 0,0 | 0,0 | 0,1 | 0,3 | 0,2 | 0,4 | 0,5 | 1,9 | 2,8 | 4,4 | 7,3 |
| Décès ou IAD | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,1 | 0,1 | 0,1 | 0,2 |

FIGURE 2.9 – Sans les contrats sans effet et sans suite - Réseaux BP et CE

Nous pouvons observer qu'en moyenne, une partie non négligeable des contrats est soit en attente, soit en processus de sélection médicale lors des premiers mois après la souscription. Évidemment, la résiliation (pour mise en demeure ou à la demande du client) est extrêmement faible sur cette période. Pour une ancienneté de contrat de 12 mois, les contrats résiliés représentent **8,6%** des contrats si on ne prend pas en compte les contrats sans suite et sans effet. Observons les taux de résiliation à ancienneté 12 mois par cohorte.

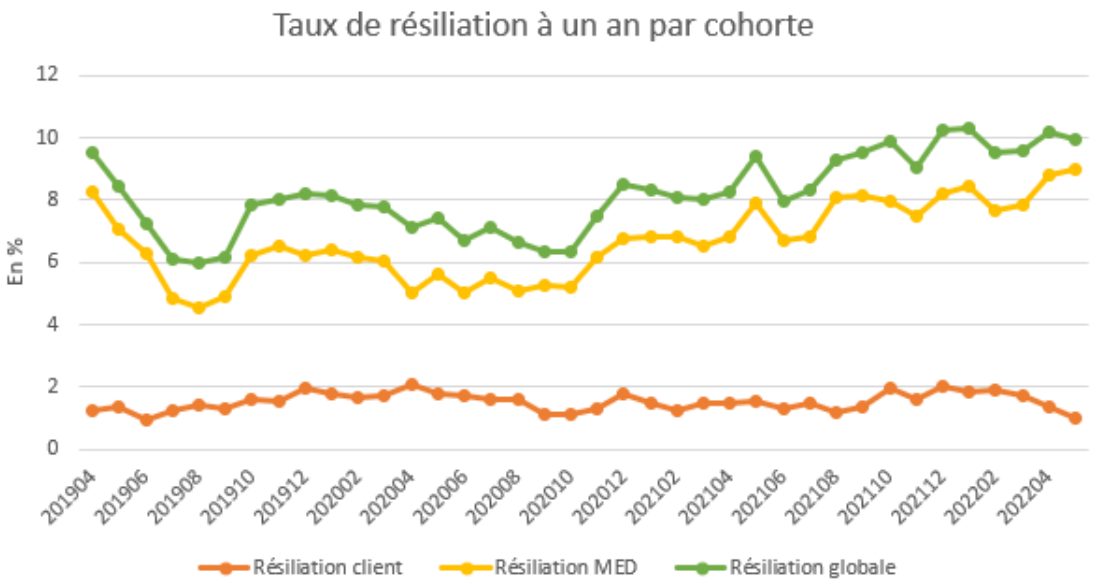


FIGURE 2.10 – Taux de résiliation à un an par cohorte

Nous constatons que les taux de résiliation à la demande du client sont relativement stables par cohorte. Les taux de résiliation suite à une mise en demeure pour non paiement des primes sont plus sensibles à la temporalité. En effet, les contrats souscrits entre avril et octobre 2020 ont des taux de résiliation à un an plus faibles, ce qui coïncide avec les premiers mois de la pandémie de COVID-19. De plus, les taux de résiliation à un an semblent augmenter depuis la cohorte de mai 2021. Cela peut être expliqué a priori par la hausse significative de l'inflation depuis le début de l'année 2022, mais aussi par d'autres facteurs que nous tenterons de découvrir au travers de notre modélisation.

**Nous décidons alors d'étudier plus en détail la résiliation des contrats Famille entre 2 et 12 mois après la souscription du contrat, ce que nous appellerons par la suite résiliation précoce**. Seuls les contrats en cours à 2 mois après la souscription seront pris en compte dans la suite. Comme la période d'observation de chaque contrat doit être de 12 mois, nous filtrons nos données sur les cohortes de contrats inférieures à mai 2022. En effet, les cohortes ultérieures sont censurées car il n'est pas possible de les observer pendant 12 mois.

Regardons maintenant le taux de résiliation précoce par apporteur. Pour des raisons de confidentialité, les apporteurs ne sont pas mentionnés.

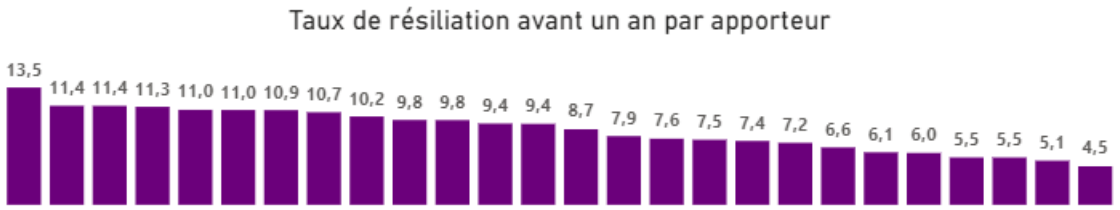


FIGURE 2.11 – Taux de résiliation précoce par apporteur

La grande variabilité des taux de résiliation avant un an entre les différents établissements nous conduit à nous interroger sur les causes des taux élevés.

Il est à noter que le développement commercial et le marketing des contrats SECUR'Famille et SECUR'Famille 2 distribués par les Caisses d'Epargne sont effectués par une autre entité, Ecureuil Vie Développement, détenue à 49% par CNP Assurances. BPCE Vie n'a des leviers d'action directs que pour les contrats Assurance Famille distribués par les Banques Populaires. Nous ne nous intéresserons donc qu'au réseau Banque Populaire à partir de maintenant. Ci-dessous l'évolution moyenne de la proportion des contrats Assurance Famille par état de contrat et par mois d'ancienneté.

| | Ancienneté du contrat (en mois) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **État du contrat** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 12 | 18 | 24 | 36 |
| En cours | 75,0 | 79,6 | 78,5 | 78,0 | 77,5 | 77,0 | 76,7 | 73,5 | 70,9 | 68,5 | 78,8 |
| En attente | 13,5 | 1,9 | 0,4 | 0,4 | 0,2 | 0,2 | 0,1 | 0,1 | 0,1 | 0,0 | 0,1 |
| Sélection médicale | 8,3 | 5,9 | 2,8 | 1,4 | 0,4 | 0,2 | 0,2 | 0,1 | 0,1 | 0,1 | 0,1 |
| Sans effet | 2,4 | 3,3 | 4,1 | 4,1 | 4,2 | 4,2 | 4,2 | 4,4 | 4,5 | 4,6 | 2,5 |
| Sans suite | 0,9 | 9,3 | 13,6 | 14,8 | 15,8 | 16,0 | 16,0 | 16,3 | 16,5 | 16,9 | 0,1 |
| En contentieux | 0,0 | 0,0 | 0,7 | 0,8 | 0,9 | 0,8 | 0,7 | 0,6 | 0,4 | 0,4 | 0,4 |
| Résiliation MED | 0,0 | 0,0 | 0,0 | 0,5 | 1,0 | 1,4 | 1,7 | 3,5 | 5,2 | 6,2 | 11,0 |
| Résiliation client | 0,0 | 0,0 | 0,1 | 0,2 | 0,2 | 0,3 | 0,4 | 1,5 | 2,3 | 3,2 | 7,0 |
| Décès ou IAD | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,1 | 0,1 | 0,2 |

FIGURE 2.12 – Évolution moyenne de la proportion de contrats (%) - Réseau BP

| | Ancienneté du contrat (en mois) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **État du contrat** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 12 | 18 | 24 | 36 |
| En cours | 77,5 | 91,1 | 95,2 | 96,1 | 96,7 | 96,5 | 96,1 | 92,5 | 89,7 | 87,3 | 80,8 |
| En attente | 13,9 | 2,2 | 0,5 | 0,4 | 0,3 | 0,2 | 0,2 | 0,1 | 0,1 | 0,1 | 0,1 |
| Sélection médicale | 8,6 | 6,7 | 3,3 | 1,7 | 0,5 | 0,3 | 0,2 | 0,1 | 0,1 | 0,1 | 0,1 |
| En contentieux | 0,0 | 0,0 | 0,8 | 1,0 | 1,1 | 0,9 | 0,9 | 0,8 | 0,5 | 0,4 | 0,4 |
| Résiliation MED | 0,0 | 0,0 | 0,0 | 0,6 | 1,2 | 1,7 | 2,1 | 4,6 | 6,6 | 7,9 | 11,3 |
| Résiliation client | 0,0 | 0,0 | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 | 2,0 | 2,9 | 4,1 | 7,2 |
| Décès ou IAD | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,1 | 0,1 | 0,2 |

FIGURE 2.13 – Sans les contrats sans effet et sans suite - Réseau BP

L'évolution de la proportion des contrats Assurance Famille est semblable à celle des contrats des deux réseaux. Les contrats résiliés représentent **6,6%** des contrats sans prendre

en compte les contrats sans suite et sans effet. Nous restreignons désormais nos données aux contrats Assurance Famille souscrits avant mai 2022 et en cours au deuxième mois post-souscription. ce qui représente près de 280 000 contrats. Nous enlevons les lignes correspondant aux contrats clôturés suite à un sinistre car nous les considérons censurés.

**Ajout de variables externes**

Il est a priori intéressant d'ajouter à nos données le nom de l'agence dans laquelle le contrat a été vendu ainsi que la typologie du conseiller qui s'est occupé de la vente. D'après une responsable régionale des ventes, les contrats de Prévoyance Individuelle ne sont pas vendus de la même manière selon les agences et les conseillers car les objectifs de vente et les stratégies marketing peuvent drastiquement changer. Il existe plusieurs types de conseillers de vente dans le réseau Banque Populaire :

- Conseillers Particuliers : ils gèrent des clients grand public et sont également les interlocuteurs des clients patrimoniaux au quotidien. Cependant, ils ne peuvent vendre de contrats directement aux clients patrimoniaux, ils doivent pour cela faire appel à un conseiller en gestion de patrimoine.
- Conseillers en Gestion de Patrimoine : ils s'occupent principalement de clients patrimoniaux.
- Conseillers Premium : ils gèrent des clients aisés et n'ont pas besoin de passer par l'intermédiaire d'un conseiller en gestion de patrimoine.
- Conseillers Privés : ils s'occupent de clients fortunés ou de chefs d'entreprise.
- Conseillers Pros : ils gèrent les clients dits Pros, c'est-à-dire des chefs d'entreprise, des micro-entrepreneurs ou des Travailleurs Non Salariés exerçant une profession libérale.
- Directeurs d'agence : ils managent les conseillers de leur agence mais ont également un portefeuille de clients Pros.
- Autres Métiers : dans cette catégorie sont regroupés tous les autres métiers non évoqués ci-dessus, ainsi que les stagiaires et les alternants.

Les noms des agences et la typologie des conseillers peuvent se trouver dans les rapports hebdomadaires de chaque Banque Populaire. Ces rapports étant publiés par les Banques Populaires et utilisés notamment par des responsables commerciaux et marketing, nous pouvons nous assurer qu'ils resteront disponibles dans le futur.

Les agences et les conseillers étant amenés à apparaître et disparaître au cours du temps, nous allons prendre les noms d'agence provenant du rapport hebdomadaire le plus récent. Cependant, pour ajouter la typologie des conseillers de vente, nous utilisons le rapport

hebdomadaire le plus ancien, datant de juin 2022. Pour rappel, les données que nous avons désormais sont celles des contrats souscrits avant mai 2022.

### 2.3.3 Conformité RGPD

Nous avons désormais rassemblé toutes les données qui nous semblent pertinentes pour notre étude et auxquelles nous avons facilement accès. Vérifions si cela respecte le RGPD.

Le RGPD[CNI18], ou Règlement Général sur la Protection des Données, est une réglementation de l'Union Européenne qui vise à renforcer et à unifier les règles pour la collecte, le traitement et la conservation des données personnelles des citoyens au sein de l'UE. Le RGPD s'applique à toute organisation publique ou privée qui traite des données personnelles pour son compte ou non, dès lors qu'elle est établie sur le territoire de l'UE ou que son activité cible directement des résidents européens. Le RGPD concerne aussi les sous-traitants qui traitent des données personnelles pour le compte d'autres organismes.

Les données personnelles se définissent par toute information se rapportant à une personne physique permettant directement ou non grâce à un ou plusieurs éléments de l'identifier. C'est le cas par exemple d'un nom, d'un prénom, d'un numéro de téléphone, d'une adresse électronique, d'un numéro de carte d'identité et/ou de sécurité sociale, d'une adresse IP ou d'une photo. Un traitement de données personnelles est quant à lui une opération ou ensemble d'opérations portant sur des données personnelles, quel que soit le procédé utilisé (collecte, conservation, modification, extraction, consultation, utilisation, diffusion, rapprochement).

Le RGPD[Ent22] établit plusieurs principes fondamentaux pour le traitement des données personnelles, notamment :

- Consentement : Les données personnelles ne peuvent être traitées qu'avec le consentement explicite et informé de la personne concernée, sauf dans certains cas spécifiques.
- Finalité : Les données ne peuvent être collectées que pour des finalités spécifiques et légitimes, et ne peuvent pas être traitées de manière incompatible avec ces finalités.
- Minimisation des données : Seules les données nécessaires pour atteindre la finalité du traitement doivent être collectées et utilisées, afin de minimiser la quantité de données personnelles traitées.
- Exactitude : Les données doivent être exactes et tenues à jour. Des mesures doivent être prises pour rectifier ou effacer les données inexactes.

- Limitation de la conservation : Les données personnelles ne doivent pas être conservées plus longtemps que nécessaire pour atteindre la finalité du traitement.
- Intégrité et confidentialité : Des mesures de sécurité appropriées doivent être mises en place pour protéger les données personnelles contre tout accès non autorisé, perte ou altération.
- Transparence : Les personnes concernées doivent être informées de manière claire et transparente sur la manière dont leurs données personnelles sont traitées.
- Droits des personnes concernées : Les individus ont le droit d'accéder à leurs données personnelles, de les rectifier, de les effacer et de s'opposer à leur traitement dans certaines circonstances.

Ces principes visent à garantir que les données personnelles sont traitées de manière éthique et à protéger les droits des individus quant à l'utilisation de leurs informations personnelles.

À notre niveau, nous pouvons vérifier nous-même si les principes de finalité, minimisation des données, exactitude et confidentialité sont bien respectés. Nous pouvons examiner brièvement les variables que nous avons :

| **Variable** |
|---|
| Région de résidence du client |
| Nom d'agence |
| Type de conseiller |
| Formule |
| Top promo |
| Top multi-équipé |
| Mode de signature |
| Canal de vente |
| Mois de création du contrat |
| Périodicité de paiement de la prime |
| Age du client à l'adhésion |
| Ancienneté du client à l'adhésion |
| Sexe du client |
| Catégorie socio-professionelle |
| Montant de prime |
| Montant du capital garanti |

TABLE 2.5 – Liste des variables

Une étude actuarielle est un motif légitime pour utiliser ces données. Nous n'avons recueilli que des données dont nous pensons qu'elles pourraient expliquer la résiliation des contrats et étant donné qu'elles proviennent du DWH PI, nous considérons que ces données sont fiables. Aucun nom, prénom ou adresse précise n'a été gardé et aucune donnée sensible n'est présente. Toutefois, une variable que nous devons regarder avec attention est la catégorie socioprofessionnelle.

En effet, le Code de la Défense encadre strictement les traitements de données personnelles de militaires. En application des dispositions de l'article L4123-9-1 dudit Code[Cod19], et en application du décret n°2018-932 du 29 octobre 2018 modifiant les dispositions du code de la défense relatives à la sécurité des traitements de données à caractère personnel comportant la mention de la qualité de militaire : « le responsable d'un traitement, automatisé ou non, ne peut traiter les données dans lesquelles figure la mention de la qualité de militaire des personnes concernées que si cette mention est strictement nécessaire à l'une des finalités du traitement. A l'exclusion des traitements mis en œuvre pour le compte de l'Etat, des collectivités territoriales et de leurs groupements ainsi que des associations à but non lucratif, les responsables des traitements informent le ministre compétent de la mise en œuvre de traitements comportant [...] la mention de la qualité de militaire ». En cas de non-respect, le risque s'élèverait à 300 000 euros d'amende par responsable de traitement concerné, hors impacts d'image et risque pénal pour l'entreprise.

La catégorie socioprofessionnelle que nous utilisons est celle de niveau 1 définie par l'INSEE [1], dans laquelle la qualité de militaire n'est pas mentionnée. Nos données sont donc bien conformes au RGPD.

### 2.3.4 Probabilités et *Odds ratios*

Il est intéressant de quantifier l'impact de nos variables et de leurs modalités sur la variable cible, à savoir la résiliation précoce. Pour cela, deux mesures sont souvent utilisées : les risques relatifs (*Relatives Risks*) et les rapports de cote (*Odds Ratios*)[Com19].

1. Institut National Statistique et des Études Economiques

Une probabilité $P$ correspond à la vraisemblance d'un évènement et varie de 0 (évènement impossible) à 1 (évènement certain). Une estimation de la probabilité d'un évènement est :

$$\widehat{P} = \frac{Nb\ occurrences}{Nb\ observations}$$

Le rapport de cote d'un évènement est le ratio entre la probabilité de l'évènement et son complémentaire.

$$Odds = \frac{Nb\ occurrences}{Nb\ observations - Nb\ occurrences} = \frac{\widehat{P}}{1 - \widehat{P}}$$

Les rapports de cote compris entre 0 et 1 correspondent à des évènements qui ont plus de chances de ne pas se produire que de se produire, et vice versa pour ceux supérieurs à 1 (ils ont plus de chance de se produire que de ne pas se produire). Nous utiliserons les *Odds ratios* pour estimer l'impact des modalités des différentes variables à notre disposition sur notre variable cible, et donc leur intérêt pour le modèle et leur influence sur la résiliation.

Les *odds ratios* suivants mettent en évidence l'impact de la catégorie socioprofessionnelle, du sexe et de l'âge du client sur la résiliation précoce des contrats :

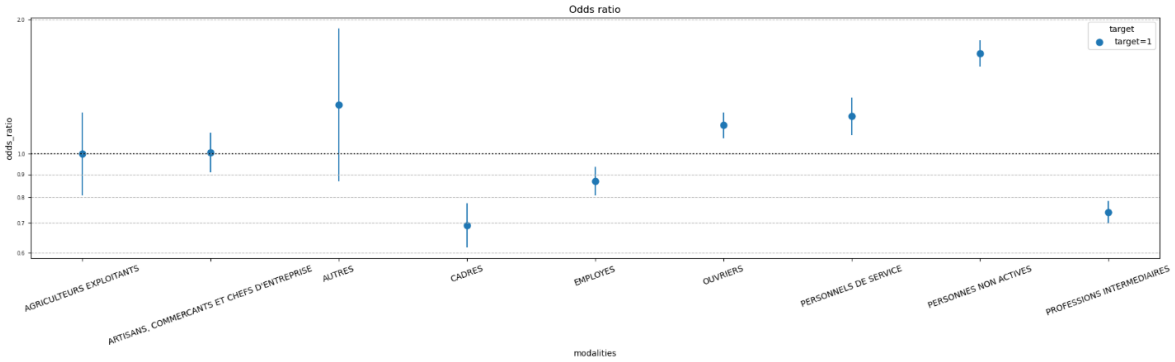


FIGURE 2.14 – *Odds ratios* - Catégorie socioprofessionnelle

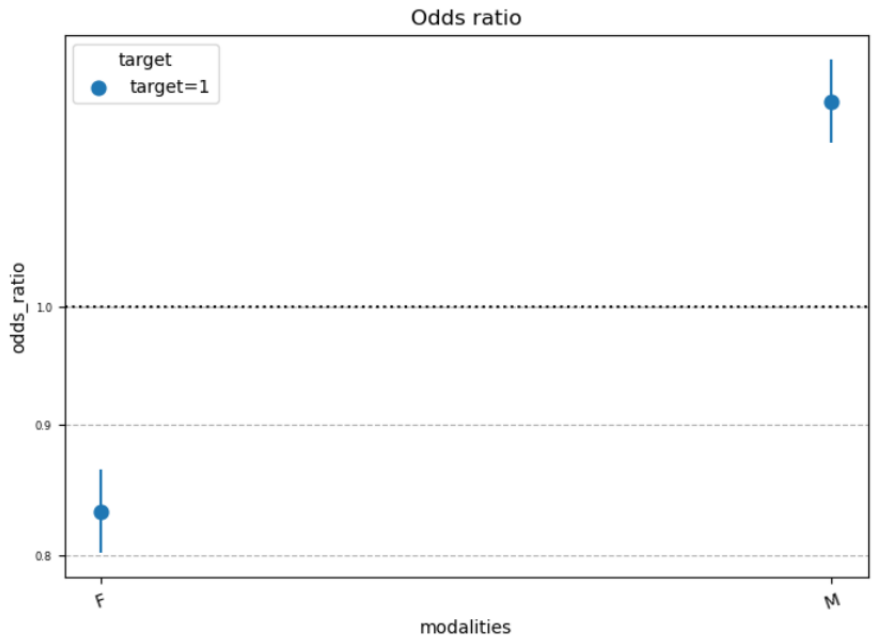


FIGURE 2.15 – *Odds ratios* - Sexe

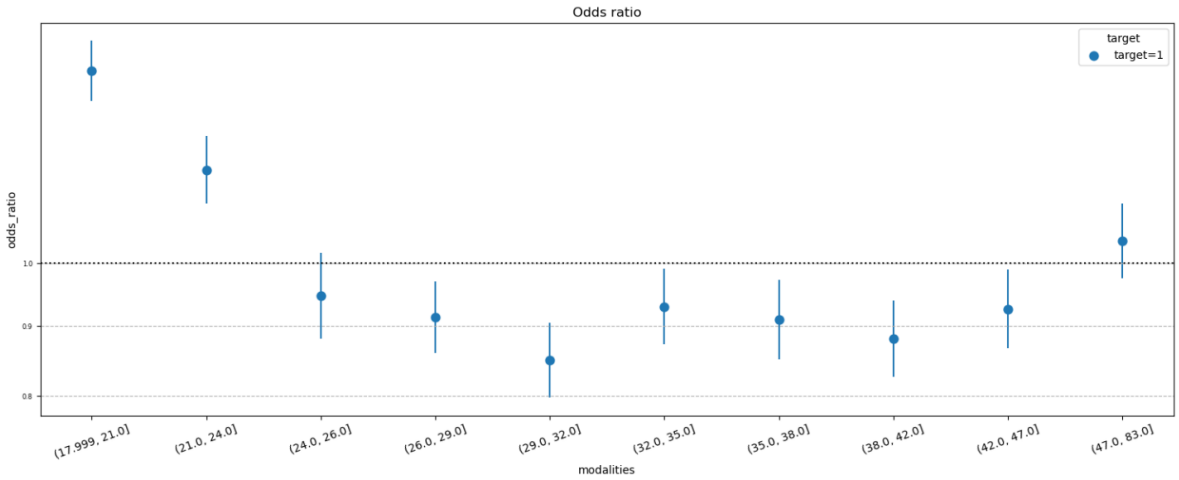


FIGURE 2.16 – *Odds ratios* - Âge

D'après les figures ci-dessus, les hommes, les jeunes de moins de 24 ans et les catégories socioprofessionnelles les moins aisées (ouvriers, personnels de service et personnes non actives) sont plus susceptibles de résilier leur contrat. Cela peut s'expliquer par leur situation économique plus précaire. Observons également l'impact des variables ayant trait aux conditions de souscription des contrats.

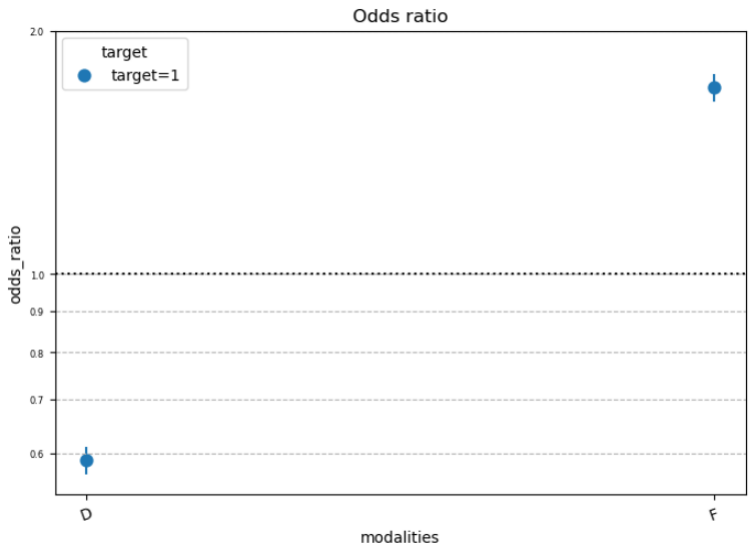


FIGURE 2.17 – *Odds ratios* - Canal de vente

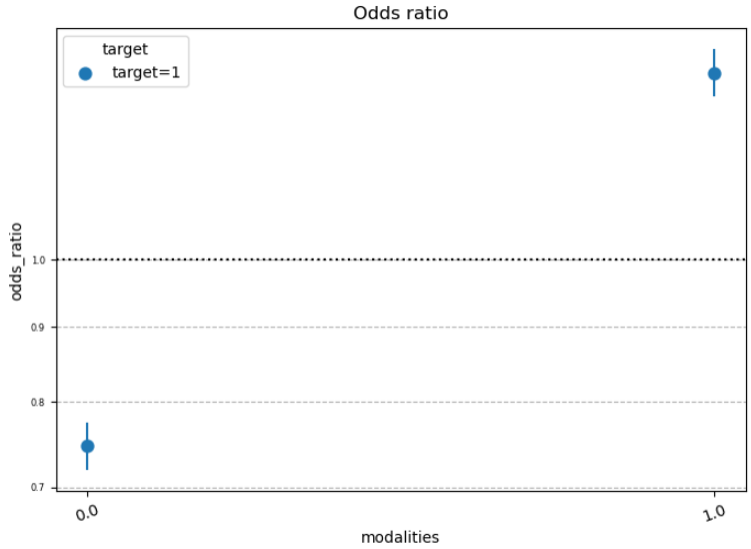


FIGURE 2.18 – *Odds ratios* - Top promo

Les contrats vendus en face à face et ceux vendus dans le cadre d'une offre promotionnelle sont plus susceptibles d'être résiliés. Cela paraît plausible, car une fois l'offre terminée, la prime augmente ce qui peut donc entraîner une volonté de résilier.

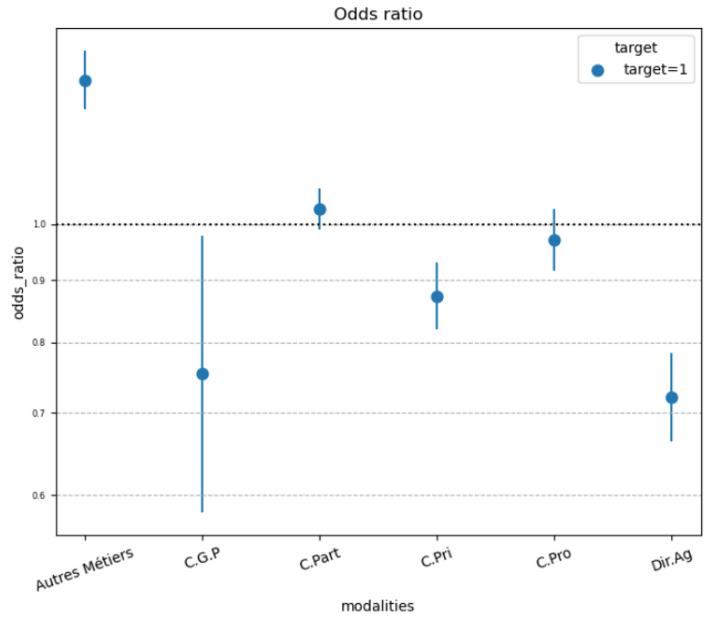


FIGURE 2.19 – *Odds ratios* - Type de conseiller

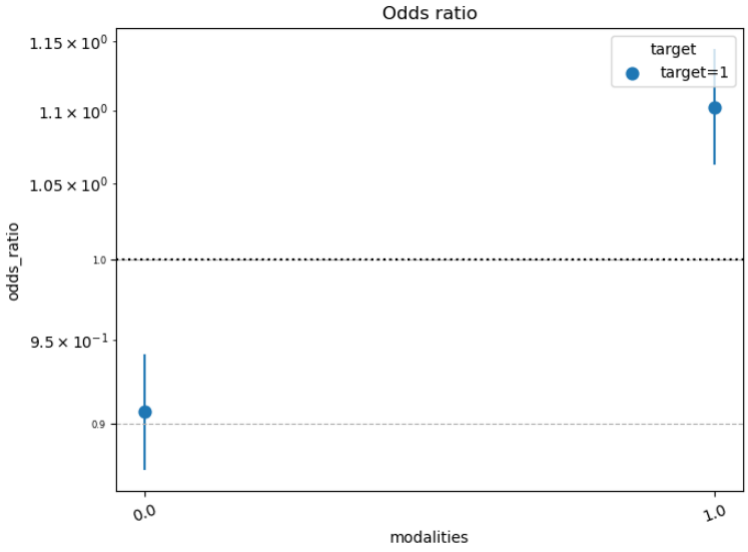


FIGURE 2.20 – *Odds ratios* - Top prévoyance

Les contrats vendus par des directeurs d'agence, des conseillers privés et des conseillers en gestion de patrimoine ont tendance à être moins résiliés. Une hypothèse est que ces conseillers s'adressent à un type bien précis de clientèle, qui résilie moins leurs contrats. De plus, les clients ayant déjà un ou plusieurs contrats de prévoyance individuelle résilient davantage leur contrat Assurance Famille. Ce n'est pas forcément étonnant car la présence d'un autre contrat de prévoyance peut indiquer une redondance dans les risques pour lesquels

le client est couvert. Il peut aussi résilier car il a l'impression qu'il est suffisamment couvert et qu'il n'a pas besoin d'un contrat supplémentaire, ou encore pour des raisons financières.

Nous avons désormais une bonne image de l'influence des modalités de nos variables sur la résiliation précoce. Cependant, nous ne savons pas quelles variables sont les plus déterminantes. Compte tenu du volume important de données auxquelles nous avons accès, il est intéressant d'utiliser des algorithmes d'apprentissage automatique (*machine learning*) pour continuer cette étude.

Afin de tenir compte de l'hétérogénéité de la résiliation précoce entre les différents apporteurs, la stratégie est la suivante : développer un modèle d'apprentissage automatique pour un établissement du réseau BPCE appelé «établissement pilote», puis généraliser ce modèle aux autres établissements. Nous allons également tester la mise en production du modèle, ce qui implique de n'utiliser que des données que nous savons être à notre disposition de façon pérenne, ce qui est notre cas. Dans le chapitre suivant, nous présentons les différents modèles d'apprentissage automatique, les métriques auxquels nous nous intéresserons ainsi que la mise en place des modèles.

# Chapitre 3

# Détection de l'attrition précoce

Nous souhaitons mettre en place différents modèles d'apprentissage automatique afin d'étudier la résiliation précoce des contrats Assurance Famille.

## 3.1 Présentation des modèles et des métriques

Il existe trois principaux types d'apprentissage :

- l'apprentissage supervisé qui implique l'utilisation de données étiquetées pour entraîner un modèle à faire une prédiction. Les algorithmes de classification et de régression sont des exemples d'apprentissage supervisé. Un modèle de classification prédit l'appartenance d'une observation à une classe tandis qu'un modèle de régression prédit une valeur numérique ou continue.
- l'apprentissage non supervisé qui regroupe les techniques cherchant à identifier des structures et des relations dans des données non étiquetées (clustering).
- l'apprentissage par renforcement qui consiste à apprendre quelles décisions choisir dans un environnement pour maximiser l'obtention d'une récompense.

La problématique de notre étude est l'explication et accessoirement la prédiction de la résiliation précoce des contrats Assurance Famille. La variable cible vaut 1 (classe positive) si le contrat a été résilié entre deux et douze mois après sa souscription et vaut 0 (classe négative) sinon. Il s'agit d'un cas d'apprentissage supervisé classique, à savoir la classification binaire. Nous allons désormais présenter les algorithmes de classification que nous mettrons en place par la suite, ainsi que les métriques utilisées pour évaluer leurs performances.

### 3.1.1 Arbres de classification

#### Introduction aux arbres de décision

Les arbres de décision[CHE23] sont un algorithme d'apprentissage supervisé utilisé pour traiter aussi bien les problèmes de régression que de classification. Leur popularité est due à leur simplicité et leur facilité d'interprétation. Ils sont couramment utilisés dans des domaines d'exploration de données ou d'aide à la décision. Leur objectif est de constituer des groupes d'individus les plus homogènes possibles selon un ensemble de variables prédictives par rapport à la variable cible. Ils se développent à partir du sommet de l'arbre, appelé nœud racine, en utilisant des règles de décision pour diviser de manière binaire les données en sous-ensembles.

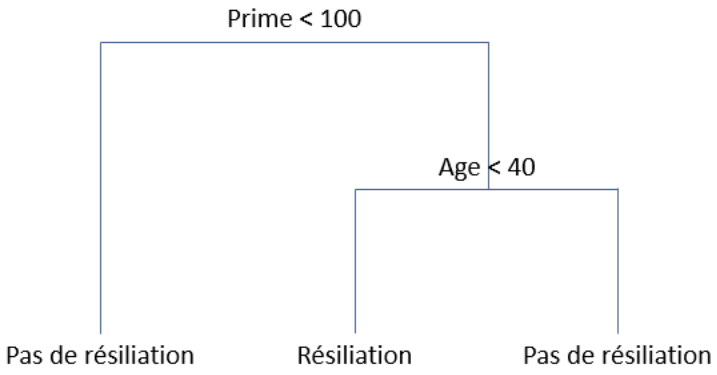


FIGURE 3.1 – Exemple d'arbre de décision

La figure ci-dessus présente un arbre de classification binaire contenant deux nœuds internes et trois nœuds terminaux. Les nœuds terminaux, également appelés feuilles, représentent la classe prédite par l'arbre. Dans cet exemple, le nœud racine (nœud parent) répartit les observations selon la variable Prime à un seuil de 100. Une observation dont la valeur de la variable Prime est inférieure à ce seuil donné se retrouve dans la feuille gauche. Si, au contraire, une observation a une valeur de Prime supérieure à 100, elle est assignée à l'autre nœud interne (nœud enfant) qui répartit les observations par la variable Age à un seuil de 40. Ce processus de division se poursuit pour toutes les instances jusqu'à ce qu'elles se retrouvent dans l'une des feuilles.

#### Terminologie

- Nœud (ou racine) : représente l'échantillon entier ou le sous-échantillon qui sera ensuite divisé en deux ensembles homogènes ou plus.

- Nœud de décision (nœud interne) : lorsqu'un sous-nœud se divise en plusieurs sous-nœuds, il est appelé nœud de décision.
- Partitionnement : processus de division d'un nœud en deux ou plusieurs sous-nœuds.
- Nœud parent et enfant : un nœud qui est divisé en sous-nœuds est appelé nœud parent. Ses descendants sont appelés nœuds enfants.
- Feuille (nœud terminal) : les nœuds qui n'ont pas de descendant.
- Branche (sous-arbre) : correspond à une sous-section de l'arbre entier.

**Critères de segmentation**

Le partitionnement des observations dans un nœud de décision se fait grâce au calcul d'un indicateur de qualité[ROK05]. Plus cet indicateur est élevé et plus l'homogénéité des données est élevée. La variable retenue pour le partitionnement sera celle qui optimise cet indicateur.

L'indicateur de qualité le plus couramment utilisé est le coefficient de Gini. Il mesure la dispersion d'une distribution dans une population donnée.

$$Gini_t = \sum_{i \neq j} P(i|t)P(j|t)$$

où $P(i|t)$ désigne la proportion d'éléments de la classe $i$ affectée au nœud $t$. Cet indicateur prend des valeurs comprises entre 0 et 1. Plus la valeur du coefficient de Gini est proche de 1 et plus la population est homogène. Pour partitionner les observations, l'arbre de classification va retenir une variable dont le coefficient de Gini est maximal.

L'entropie est une autre mesure de l'homogénéité d'une population. Selon la théorie de l'information, une classe pure nécessite peu d'informations pour être décrite et inversement, une classe impure nécessite plus d'informations. Pour définir le niveau d'«impureté», la mesure d'entropie est :

$$Entropie = -p\log_2(p) - q\log_2(q)$$

où $p$ et $q$ sont les probabilités de succès et d'échec respectivement dans le nœud de division. Si la classe est complètement homogène, alors l'entropie est nulle et si la classe est divisée de manière proportionnelle (par exemple 50% de résiliations et 50% de non-résiliations), l'entropie est égale à 1. La partition qui a l'entropie la plus faible par rapport au nœud parent et aux autres divisions sera choisie.

**Echantillonnage et dilemme biais-variance**

Avant de mettre en place un modèle, il est important de diviser nos données en trois échantillons : un jeu de données d'entraînement (*train dataset*), un jeu de données de test (*test dataset*) et un jeu de données de validation (*validation dataset*). Les données d'entraînement sont utilisées par les modèles pour apprendre à classer les observations, les données de test pour optimiser leurs paramètres et les données de validation pour calculer les taux d'erreur des classifieurs optimisés et ainsi vérifier les capacités de généralisation du modèle. Pour prédire les performances d'un modèle, nous devons utiliser des données qui n'ont pas contribué à son apprentissage. Les données d'entraînement et de validation doivent donc être indépendantes.

Il existe plusieurs techniques d'échantillonnage. La méthode la plus simple consiste à tirer aléatoirement sans remise un certain nombre d'observations qui formeront l'échantillon d'apprentissage. Les observations restantes constitueront l'échantillon de test. Pour obtenir un bon modèle prédictif, il est nécessaire de disposer de beaucoup de données d'entraînement, mais pour avoir une bonne estimation de l'erreur et pour paramétrer le modèle, il faut de nombreuses données de test.

Si la taille de la base d'apprentissage est petite, le modèle aura tendance à sur-apprendre les données. Toutefois, si la taille de la base d'apprentissage est trop grande, un phénomène de sous-apprentissage apparaît. La caractéristique notable de la courbe d'apprentissage est la convergence vers un certain score tant que la taille d'échantillon d'apprentissage augmente. Lorsque l'échantillon d'apprentissage contient suffisamment d'observations pour qu'un modèle converge, l'ajout de nouvelles données d'entraînement ne pourra pas augmenter la performance du modèle. La seule façon d'améliorer les résultats sera d'utiliser un autre modèle, souvent plus complexe.

En apprentissage statistique, le dilemme biais-variance repose sur la minimisation simultanée de deux sources d'erreurs : le biais et la variance. L'erreur de biais est considérée comme la différence entre la prédiction donnée par le modèle et la valeur réelle de l'observation. La source de cette erreur provient généralement d'hypothèses erronées dans le modèle. Si le biais est très élevé, cela signifie que l'algorithme manque de relations pertinentes dans les données d'apprentissage (sous-apprentissage). Quant à l'erreur due à la variance, elle est considérée comme la variabilité d'une prédiction du modèle pour une observation donnée. La variance provient de la sensibilité du modèle à de petites fluctuations des données d'apprentissage. Une variance élevée indique un sur-apprentissage du modèle.


FIGURE 3.2 – Exemple de courbe d'apprentissage (source : [KRY18])

Soit Y la variable cible que nous cherchons à prédire à partir de variables explicatives $X_j$. Pour une observation $i$, nous supposons qu'il existe une fonction $f$ telle que $y_i = f(x_i) + \epsilon_i$, où $\epsilon$ est un terme de bruit normalement distribué suivant une loi $\mathcal{N}(0, \sigma^2)$. Il est possible d'estimer une fonction $\hat{f}$ de $f$ à partir d'un modèle d'apprentissage supervisé. L'erreur de prédiction attendue de l'échantillon test $x$ se décompose ainsi :

$$\begin{aligned} Erreur(x) &= E[(Y - \hat{f}(x))^2] \\ &= E[\hat{f}(x) - f(x)]^2 + E[(\hat{f}(x) - E[\hat{f}(x)])^2] + \sigma^2 \\ &= Biais[\hat{f}(x)]^2 + Var[\hat{f}(x)] + Erreur\ Irréductible \end{aligned}$$

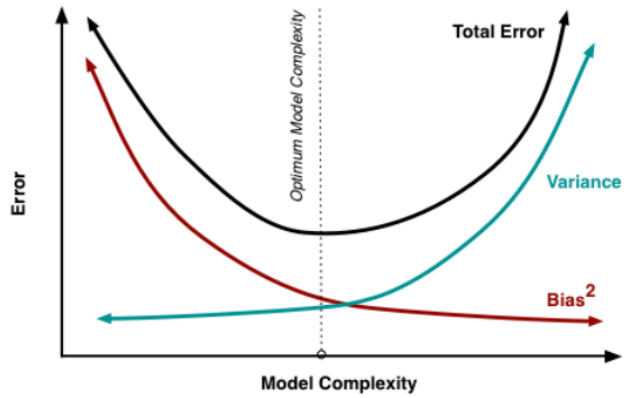


FIGURE 3.3 – Illustration du dilemme biais-variance (source : [KRY18])

L'espérance de l'erreur de prédiction peut se décomposer comme une somme de trois termes : le biais, la variance et le bruit, appelé erreur irréductible. Cette décomposition a

été initialement formulée pour un problème de régression des moindres carrés mais il est possible de trouver une décomposition similaire en classification. En effet, en reformulant le problème de classification comme classification probabiliste, l'erreur quadratique attendue des probabilités prédites par rapport aux véritables probabilités peut être décomposée comme précédemment.

Le grand enjeu de tout modèle est donc de minimiser simultanément le biais et la variance du modèle. Plusieurs méthodes sont possibles : changement de la taille de l'échantillon d'apprentissage, élagage pour les arbres de décision, hyperparamétrisation du modèle ou utilisation de méthodes ensemblistes.

**Paramétrage d'un arbre**

Un des principaux défis rencontrés lors de la modélisation basée sur les arbres de décision est le sur-apprentissage. En effet, si un arbre de décision n'a aucune contrainte sur son développement, celui-ci donnera une précision de 100% sur l'ensemble d'information, avec une feuille pour chaque observation. Empêcher le sur-apprentissage est donc essentiel afin de minimiser le biais et surtout la variance du modèle. Cela est possible en paramétrant l'arbre correctement.

Les paramètres principaux de la taille de l'arbre peuvent être définis manuellement :

- Nombre minimal d'observations dans un nœud de division : Nombre minimal d'observations requis dans un nœud pour qu'il puisse être considéré comme un nœud de décision. Des valeurs élevées du paramètre empêchent au modèle d'apprendre des relations qui peuvent être très spécifiques à l'échantillon. Cependant, des valeurs trop élevées peuvent entraîner un sous-apprentissage du modèle.
- Nombre minimal d'observations dans une feuille : Nombre d'observations dans le nœud terminal. Si la variable cible contient des classes déséquilibrées, il est préférable de choisir des petites valeurs pour ce paramètre car la classe minoritaire est plus rare.
- Taille de l'arbre : Profondeur maximale de l'arbre de décision. Un arbre avec une grande profondeur peut contenir des relations trop spécifiques et donc entraîner un sur-apprentissage du modèle.
- Nombre de feuilles maximal : Peut être défini à la place de la taille de l'arbre.

Il est important de noter que le paramétrage initial du modèle est rarement celui qui fournit les meilleurs résultats. L'ajustement des paramètres doit être fait par validation croisée.

Il est à noter que si les arbres de décision sont simples à mettre en place et très faciles d'interprétation, ils sont également instables et sensibles, que ce soit aux données d'apprentissage ou aux paramètres du modèle comme nous l'avons vu précédemment. Il existe plusieurs méthodes qui permettent de réduire ces problèmes d'instabilité. Parmi ces techniques, nous distinguons l'agrégation des modèles (ou prédicteurs) issus d'une même famille. Le but de l'agrégation est de faire converger les résultats des prédicteurs issus du même modèle vers une solution commune. Les méthodes d'agrégation les plus populaires sont le *boosting* et le *bagging*.

### 3.1.2 Forêts aléatoires

Le principe du *bagging* consiste à tirer un grand nombre d'échantillons de façon indépendante (bootstrap) et d'appliquer à chacun d'eux un modèle de prédiction (appelé «règle de base»). Les résultats des prédictions de chaque échantillon sont ensuite agrégés par une moyenne ou par un vote majoritaire. Dans le cas d'un problème de classification, il s'agit d'agrégation des probabilités de réponse de chaque modalité. Le rééchantillonnage peut être effectué par tirage aléatoire avec ou sans remise. L'un des principaux avantages de l'agrégation par *bagging* est la réduction de la variance du modèle de base, et donc l'atténuation de sur-apprentissage. En effet, si chaque échantillon a la même variance $\sigma^2$, alors la variance agrégée du *bagging* est de $\frac{\sigma^2}{n}$.

L'algorithme des forêts aléatoires[BRE01] (ou *Random Forest*) fait partie de la famille des méthodes de *bagging*. La principale différence réside dans le fait que la règle de base est appliquée non pas aux échantillons bootstrapés, mais aux arbres qui sont créés via le tirage aléatoire des observations et des $k$ variables prédictives. L'algorithme *Random Forest* peut être décrit ainsi :

- Tirer aléatoirement avec remise $k$ variables prédictives parmi les $K$ variables totales de la base d'apprentissage et de $n$ observations parmi les $N$ de la base d'apprentissage. Le nombre de variables tirées à chaque itération ainsi que le nombre de tirages peuvent être spécifiés manuellement.
- Appliquer la règle de base (arbre de classification) aux échantillons créés.
- Agréger les résultats de chaque arbre construit via la moyenne des probabilités de réponse de chaque modalité, (pour un problème de classification).

Les forêts aléatoires sont construites à partir de faibles prédicteurs qui vont former un prédicteur agrégé plus puissant en termes de stabilité et de capacité de prédiction. Cependant, elles sont souvent assimilées à une boîte noire : il n'est pas possible de savoir tout ce qui

se passe à l'intérieur de l'algorithme et les méthodes d'interprétabilité de ce modèle sont limitées (valeurs de Shapley, LIME...).

### 3.1.3 CatBoost

Le *boosting* désigne quant à lui la transformation des faibles prédicteurs en forts prédicteurs. Les prédicteurs appartenant à ce type de famille sont formés de façon séquentielle et non parallèle comme pour le *bagging*. La technique du *boosting* se base sur les erreurs des prédicteurs précédents. De plus, des observations ont des probabilités différentes de contribuer à la construction du prédicteur. En effet, les observations classées initialement de façon incorrecte ont plus de chance d'apparaître dans les prédicteurs suivants. Le choix des observations d'apprentissage est donc basé sur les erreurs du prédicteur et non sur le processus de ré-échantillonnage. Cependant, le nombre d'itérations de l'algorithme de *boosting*, qui est le critère d'arrêt, doit être bien choisi pour ne pas faire sur-apprendre le modèle.

Le modèle CatBoost[MAC22] (*Categorical Boosting*) est une implémentation de l'algorithme de *gradient boosting*. C'est une technique d'apprentissage supervisé qui, dans le cas d'une classification, vise à minimiser le coût des erreurs de classement (fonction de perte). Ce problème de minimisation de coût revient à minimiser l'erreur résiduelle du prédicteur précédent en implémentant deux techniques simultanément : le *boosting* et la descente de gradient.

La descente de gradient est une méthode itérative utilisée pour trouver la solution d'un problème d'optimisation. Dans le cas du *gradient boosting*, le problème d'optimisation revient à minimiser la fonction de perte $L(y, f)$ au regard des paramètres du modèle $f$, $y$ étant la variable cible. Cette fonction de perte peut être présentée comme :

$$\sum_{i=1}^{N} l(y_i, f(x_i))$$

où $l$ est une fonction de coût entre la valeur réelle de la variable cible et la prédiction du modèle $f$ pour une observation $i$ donnée. La formule récursive de descente du gradient est alors :

$$f_{k+1}(x_i) = f_k(x_i) - \eta \frac{l(y_i, f(x_i))}{f(x_i)}$$

où $k < K$, $K$ étant le nombre d'itérations total. Les paramètres $K$ et $\eta$ (constante d'apprentissage) sont choisis manuellement.

L'implémentation la plus connue du *gradient boosting* est le XGBoost. CatBoost diffère de ce dernier notamment par sa capacité à traiter nativement les variables catégorielles. En effet, la plupart des algorithmes d'apprentissage automatique nécessitent que les données d'apprentissage ne contiennent que des variables numériques : les variables catégorielles doivent être encodées. CatBoost permet donc de diminuer le pré-traitement manuel des données.

Une autre différence entre CatBoost et XGBoost est que CatBoost utilise des arbres de décision inconscients (*oblivious*). Les arbres de décision inconscients sont des arbres de décision pour lesquels tous les nœuds au même niveau sont partitionnés en fonction de la même variable. CatBoost est aussi connu pour avoir une vitesse d'apprentissage plus rapide et moins sur-apprendre que la plupart des autres modèles de *boosting*. Toutefois, un modèle CatBoost est plus chronophage qu'un algorithme plus simple tel que les arbres de décision. C'est un paramètre important à prendre en compte, notamment lors de l'optimisation des hyperparamètres. Ce modèle est également, comme les forêts aléatoires, de type boîte noire.

### 3.1.4 Régression logistique

La régression logistique[MAC22] est une méthode de classification couramment utilisée pour prédire une variable réponse binaire $Y \in \{0, 1\}$. Soit $X$ un ensemble $m$-dimensionnel de variables où $x_i$ est le vecteur de variables appartenant à la $i$-ème observation. La régression logistique modélise la probabilité conditionnelle de l'appartenance de l'observation à une classe spécifique (0 ou 1). Cette probabilité est modélisée à l'aide de la fonction logistique :

$$P(Y_i = 1|X_i = x_i) = \frac{e^{\beta_0+\beta^T x_i}}{1 + e^{\beta_0+\beta^T x_i}}$$

où $\beta_0$ est l'ordonnée à l'origine et $\beta = [\beta_1, \beta_2, ..., \beta_m]^T$ est le vecteur des coefficients associés aux variables prédictives. Cette équation peut également s'écrire sous la forme :

$$\ln\frac{P(Y_i = 1|X_i = x_i)}{1 - P(Y_i = 1|X_i = x_i)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} = \beta_0 + \beta^T x_i$$

Le terme de gauche est appelé le rapport de cote ou *log odds ratio* (les *odds ratios* ayant été présentés précédemment). Cette quantité permet de mesurer la relation entre la variable explicative $X_i$ et la réponse $Y$. Nous pouvons encore écrire :

$$\text{logit}(p_i) = \beta_0 + \beta^T x_i$$

où $p_i$ est la probabilité que la $i$-ème observation appartienne à la classe 1.

Puisque le modèle logistique renvoie une valeur entre 0 et 1, une valeur seuil est choisie afin de prédire la classe à laquelle une observation appartient. Ce seuil est généralement fixé à 0.5. Ainsi, si la probabilité $p_i$ est supérieure à 0.5, le modèle prédira que l'observation $i$ appartient à la classe 1 et inversement pour la classe 0.

Afin d'ajuster le modèle, nous devons maximiser la fonction de vraisemblance. En posant $\theta = (\beta_0, \beta)$ et $p(x_i, \theta) = P(Y_i = 1 | X_i = x_i, \theta)$, nous avons :

$$p(x_i, \theta) = \frac{1}{1 + e^{-\theta^T x_i}}$$

Dans le cas d'une classification binaire, la distribution de Bernoulli est souvent utilisée comme fonction de vraisemblance car $Y_i | X_i = x_i$ suit une loi de Bernoulli de paramètre $p(x_i, \theta)$. L'expression de la fonction de vraisemblance est donc :

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i, \theta)^{y_i} (1 - p(x_i, \theta))^{1 - y_i}$$

et celle de la log-vraisemblance est :

$$l(\theta) = \sum_{i=1}^{n} y_i \ln p(x_i, \theta) + (1 - y_i) \ln (1 - p(x_i, \theta))$$

Nous en déduisons l'entropie croisée :

$$J(\theta) = -\frac{1}{n} \sum_{i=1}^{n} y_i \ln p(1 | x_i, \theta) + (1 - y_i) \ln (1 - p(0 | x_i, \theta))$$

ce qui donne sous forme matricielle :

$$J(\theta) = -\frac{1}{n} \sum_{i=1}^{n} y^T \ln p(1 | x, \theta) + (1 - y)^T \ln (1 - p(0 | x, \theta))$$

C'est la fonction que nous voulons minimiser en trouvant les coefficients optimaux $\theta$. Pour cela, plusieurs méthodes sont possibles. La méthode itérative de Newton-Raphson est communément utilisée : en pratique, les coefficients sont mis à jour à chaque itération par la formule de récurrence suivante :

$$\theta_{t+1} = \theta_t - H^{-1}(\theta_t) \nabla_\theta J(\theta_t)$$

Ici, $\nabla_\theta J(\theta_t)$ est un vecteur de dérivées partielles de la fonction d'entropie croisée vectorisée, par rapport aux composantes de $\theta$, à la $t$-ème itération. $H^{-1}$ est une matrice hessienne qui contient les dérivées secondes de la fonction d'entropie croisée binaire vectorisée, également

dérivée par rapport aux composantes de $\theta$. Le processus se répète au fur et à mesure que $\hat{\theta}$ converge vers ses vraies valeurs, jusqu'à ce qu'un critère d'arrêt soit respecté.

Les modèles qui seront mis en place sont maintenant présentés. Nous allons désormais nous intéresser à leur interprétabilité.

### 3.1.5 Interprétabilité des modèles

En général, l'interprétabilité d'un algorithme[JOH19] de *machine learning* est essentielle pour de nombreuses raisons :

- Avoir une plus grande confiance dans le modèle en tant qu'utilisateur lorsque ce modèle est utilisé comme outil d'aide à la décision
- Générer une analyse à forte valeur ajoutée pour le décisionnaire
- Respecter le nouveau cadre légal de la Commission Européenne qui introduit un droit d'explication des algorithmes
- Prendre en compte d'autres critères de validation des modèles. En effet, des contraintes d'ordre éthique, juridique ou opérationnel peuvent jouer un rôle dans la validation finale du modèle, par exemple vérifier l'absence de « biais encapsulé » au sein de l'algorithme d'apprentissage afin d'assurer son équité.

Les modèles de régression linéaire et d'arbre de décision sont considérés comme interprétables étant donné leur complexité faible, la théorie mathématique disponible et la manipulabilité possible de leurs structures et résultats. Cependant, les algorithmes d'apprentissage automatique de type « boîte noire » comme les forêts aléatoires et les algorithmes de *boosting* posent quelques contraintes d'interprétabilité de par leur nature.

Deux méthodes communes pour interpréter ce type d'algorithmes sont les méthodes LIME et Shapley. L'algorithme LIME (*Local Interpretable Model-agnostic Explanations*) crée un modèle autour d'une prédiction donnée afin de l'approximer localement. Plus précisément, LIME génère de nouvelles données proches de la prédiction à expliquer, puis les apprend à l'aide d'un modèle interprétable (régression linéaire ou arbre) et de la classification faite par un modèle « boîte noire » quelconque (méthode agnostique). L'inconvénient de la méthode LIME est que celle-ci ne permet pas de généraliser l'interprétabilité issue du modèle local à un niveau plus global.

La méthode Shapley permet quant à elle d'expliquer une décision locale tout en proposant, contrairement à LIME, une théorie axiomatique pour généraliser l'interprétabilité. La

méthode Shapley renvoie un classement des contributions des variables explicatives selon des principes issus de la théorie des jeux. La méthode étant très coûteuse en calcul, une variante SHAP (*SHapley Additive exPlanations*) a été proposée sur les mêmes bases. SHAP permet de traduire la prédiction d'un individu en explication sous forme de somme de contributions de chacune des variables.

C'est la méthode SHAP[PEL21] que nous utiliserons pour interpréter les algorithmes *Random Forest* et CatBoost. En théorie des jeux coopératifs, le cadre est le suivant : $n$ joueurs collaborent ensemble et obtiennent un gain $G$. On cherche à répartir équitablement le gain entre les $n$ joueurs, en prenant en compte la contribution individuelle des joueurs dans l'obtention du gain, mais également leur contribution au groupe, lorsqu'ils interagissent avec les autres joueurs.

Soit une fonction $c : \mathbb{P}(N) \mapsto \mathbb{R}$ (où $N = \{1, .., n\}$) caractéristique du jeu, indiquant pour chaque sous-ensemble de joueurs le gain maximal qu'ils peuvent obtenir en collaborant. Le gain total $G$ est le gain $c(N)$ qu'obtiennent tous les joueurs lorsqu'ils collaborent ensemble. D'après le théorème de Shapley, il existe une unique répartition satisfaisant quatre propriétés, assurant que la répartition du gain entre les joueurs est équitable. Cela signifie qu'il existe une unique fonction $\phi : \left| \begin{array}{ccc} N & \longrightarrow & \mathbb{R} \\ i & \longmapsto & \phi_i(c) \end{array} \right.$ vérifiant les contraintes suivantes :

| Propriété | Théorie | Interprétation | Exemple |
|---|---|---|---|
| Efficience | $\sum_{i \in N} \phi_i(c) = c(N) - c(\emptyset)$ | La **somme** des parts de chaque joueur doit **être égale au gain total.** En général c(∅) = 0, mais ce n'est pas toujours le cas, en particulier en intelligibilité ! | Le trésor (gain) est partagé en intégralité entre les n pirates (joueurs) |
| Symétrie | $\forall Z\ c(Z \cup \{i\}) = c(Z \cup \{j\}) \implies \phi_i(c) = \phi_j(c)$ | Si **deux joueurs contribuent de la même façon** dans toutes les coalitions dans lesquelles ils apparaissent, **leurs parts doivent être égales.** | Deux pirates qui contribuent de la même façon dans l'obtention du trésor, obtiennent la même part du butin. |
| Joueur nul | $\forall Z\ c(Z \cup \{i\}) = c(Z) \implies \phi_i = 0$ | Si toutes les coalitions dans lesquelles un **joueur est présent ont le même gain avec et sans lui, alors la part de ce joueur est nulle** | Un pirate qui n'apporte rien, quel que soit le groupe de pirates avec lequel il collabore, n'obtient rien. |
| Additivité | $\phi_i(c_1 + c_2) = \phi_i(c_1) + \phi_i(c_2)$ $\phi_i(ac_1) = a\phi_i(c_1)$ | **Additivité** des indices de Shapley par rapport à la **fonction caractéristique du jeu** | Si le groupe de n pirates obtient deux trésors, l'un après l'autre, les partager successivement, revient au même que de distribuer simultanément les deux trésors aux n pirates |

FIGURE 3.4 – Propriétés vérifiées par les valeurs de Shapley (Source : Quantmetry)

Cette fonction est définie par :

$$\phi_i(c) = \sum_{Z \subseteq N \setminus \{i\}} \frac{|Z|!(n - |Z| - 1)!}{n!} \times [c(Z \cup \{i\}) - c(Z)]$$

Pour obtenir $\phi_i$ il faut donc calculer pour chaque coalition $Z$ dans laquelle le joueur $i$ n'apparaît pas la différence de gain $c(Z \cup \{i\}) - c(Z)$. Cela permet de comparer le gain obtenu de la coalition avec et sans ce joueur, afin de mesurer son impact lorsqu'il collabore avec cet ensemble de joueurs. Une différence positive signifie que le joueur $i$ contribue positivement à cette coalition. A l'inverse, une différence négative signifie que le joueur $i$ pénalise le groupe. Enfin, une différence nulle indique que le joueur $i$ n'apporte rien à ce groupe. On calcule ensuite la moyenne de ces écarts sur toutes les coalitions dans lesquelles le joueur $i$ apparaît.

On peut réécrire l'expression de $\phi_i$ de la manière suivante, en regroupant les coalitions par tailles $k$ :

$$\phi_i = \frac{1}{n} \sum_{k=0}^{n-1} \binom{n-1}{k}^{-1} \sum_{\substack{Z \subset N \setminus \{i\} \\ |Z|=k}} [c(Z \cup \{i\}) - c(Z)]$$

Le dénombrement est le suivant : on classe les coalitions $Z$ par cardinal, puis on fait la moyenne des écarts pour toutes les coalitions $Z$ de mêmes cardinaux. Pour une taille de coalition $k$ il y a $\binom{n-1}{k}$ coalitions possibles. On moyenne ensuite ces résultats intermédiaires.

En *machine learning*, on souhaite expliquer la prédiction $f(x)$ associée à une observation $x$. À chaque couple $(x, f(x))$, est associé le jeu suivant :

- Les joueurs sont les valeurs $x_i$ prises par $x$ sur chaque variable prédictive $i$.
- Le gain à répartir équitablement entre tous ces joueurs est la différence entre la prédiction $f(x)$ et la moyenne des prédictions $\mathbb{E}[f(X)]$.
- La fonction caractéristique du jeu est $c(u) = \mathbb{E}[f(X)|X_u = x_u]$, définie pour toute coalition $u$ de valeurs $x_u$ de $x$. Si $u$ est la coalition $\{1, 2, 4\}$ (première, deuxième et quatrième variables prédictives), la notation $X_u = x_u$ correspond à l'évènement $X_1 = x_1, X_2 = x_2, X_4 = x_4$.

La valeur de Shapley associée à une modalité $x_i$ est donc définie par :

$$\phi_i(f, x) = \sum_{u \subseteq \{1,\ldots,n\} \setminus i} \frac{(n - |u| - 1)!|u|!}{n!} \left[\mathbb{E}[f(X)|X_{u \cup \{i\}} = x_{u \cup \{i\}}] - \mathbb{E}[f(X)|X_u = x_u]\right]$$

La somme des valeurs de Shapley d'une observation $x$ est égale à l'écart entre la prévision $f(x)$ et la moyenne des prévisions $\mathbb{E}[f(X)]$ :

$$f(x) - \mathbb{E}[f(X)] = \sum_{i=1}^{p} \phi_i$$

Il faut comprendre que ce n'est pas la prévision $f(x)$ elle-même que l'on répartit entre les joueurs (valeurs) $x_i$, mais la quantité $f(x) - \mathbb{E}[f(X)]$. Autrement dit, le coefficient $\phi_i$ explique comment les valeurs $x_i$ contribuent à décaler la prévision $f(x)$ de la moyenne $\mathbb{E}[f(X)]$ des prévisions.

L'intérêt de l'approche basée sur le théorème de Shapley est de tenir compte des effets d'interactions des valeurs $x_i$. Cependant, la méthode SHAP a également des limites. Tout d'abord, le calcul des valeurs de Shapley peut être coûteux lorsque le nombre de variables prédictives ou d'observations est grand. De plus, les valeurs de Shapley peuvent être sensibles aux données d'entraînement et au modèle d'apprentissage utilisé. Enfin, il n'existe pas de mesure pour évaluer les résultats d'un algorithme d'interprétabilité. Si un expert métier peut valider ou invalider certaines prédictions fournies par l'algorithme de *machine learning*, il demeure impossible de quantifier la précision et la justesse de l'algorithme d'intelligibilité a posteriori. Nous échangerons donc avec une responsable régionale des ventes afin d'obtenir son avis sur l'interprétabilité de nos modèles.

Nous avons présenté les différents modèles que nous allons mettre en place, ainsi qu'une méthode d'interprétabilité des modèles de type « boîte noire ». Afin de comparer les performances des modèles, plusieurs métriques seront introduites dans la section suivante.

### 3.1.6 Indicateurs de performance des modèles

Plusieurs métriques sont couramment utilisées en *machine learning* pour évaluer les performances des modèles de classification[JEN13]. Une des métriques les plus utilisées est la matrice de confusion. Dans cette matrice se trouvent le nombre de données dans l'échantillon de validation du modèle qui ont été correctement prédites ou non. Les lignes de la matrice correspondent aux données réellement observées et les colonnes aux données prédites par le modèle.

| | Prédictions | |
|---|---|---|
| Observations | Négatif | Positif |
| Négatif | VN | FP |
| Positif | FN | VP |

FIGURE 3.5 – Exemple de matrice de confusion

Dans cet exemple, VN signifie Vrais Négatifs, FP Faux Positifs, FN Faux Négatifs et VP Vrais Positifs. De cette matrice de confusion découlent d'autres métriques :

### Précision totale (Accuracy)

La précision totale (ou *accuracy* en anglais) est le rapport entre le nombre de prédictions correctes et le nombre total d'échantillons.

$$Accuracy = \frac{VN + VP}{VN + VP + FN + FP}$$

### Précision

La précision est le rapport entre le nombre de vrais positifs et le nombre total d'échantillons prédits comme positifs.

$$Précision = \frac{VP}{VP + FP}$$

### Rappel

Le rappel (ou *recall* en anglais) est le rapport entre le nombre de vrais positifs et le nombre total d'échantillons réellement positifs.

$$Rappel = \frac{VP}{VP + FN}$$

### F1-score

Le F1-score est la moyenne harmonique entre la précision et le rappel.

$$F1\text{-}score = 2 \times \frac{Précision \times Rappel}{Précision + Rappel}$$

### AUC-ROC et AUC-PR

La courbe ROC[FAW04] (*Receiving Operating Characteristic*) représente le taux de vrais positifs (sensibilité) en fonction du taux de faux positifs (spécificité) pour différents seuils de classification. Le taux de vrais positifs est le rappel, tandis que le taux de faux positifs est défini par $TFP = \frac{FP}{FP+VN}$. La courbe ROC est construite de la manière suivante : pour chaque observation, le modèle de classification produit une probabilité que celle-ci appartienne à la classe positive. Les observations sont ensuite ordonnées par probabilité croissante. Le taux de vrais positifs et le taux de vrais négatifs sont calculés pour différents seuils. Un seuil de classification agit comme un seuil de confiance au delà duquel une observation est considérée appartenir à la classe positive. Si la probabilité donnée par le modèle pour une observation est supérieure au seuil, cette observation sera classée positive, sinon elle sera classée négative. La plupart du temps, le seuil par défaut est égal à 0,5.

Chaque point sur la courbe ROC représente le compromis entre la capacité à détecter les vrais positifs et la capacité à éviter les faux positifs . Une courbe ROC idéale serait celle qui se rapproche du coin supérieur gauche du graphique, indiquant à la fois un taux de vrais positifs élevé et un taux de faux positifs faible.

Une aire sous la courbe ROC (AUC-ROC) proche de 1 est donc un indicateur de bonne performance, tandis qu'une valeur de l'AUC-ROC proche de 0.5 indique que le modèle n'est pas plus performant que le hasard. Dans ce cas, la courbe ROC se rapproche d'une droite d'équation $y = x$. La figure suivante présente un exemple de courbe ROC.

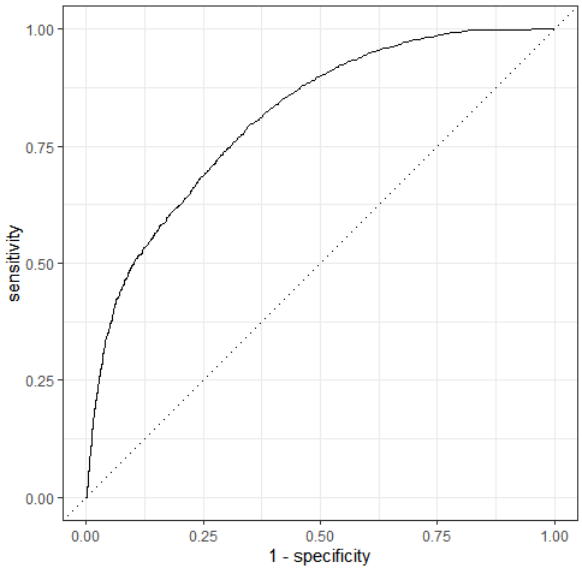


FIGURE 3.6 – Exemple de courbe ROC

La courbe Précision-Rappel[DAV04] est un autre outil graphique utilisé pour évaluer la performance d'un modèle de classification. Elle représente la précision en fonction du rappel pour différents seuils de classification. Une courbe Précision-Rappel idéale a une aire sous la courbe (AUC-PR) proche de 1.

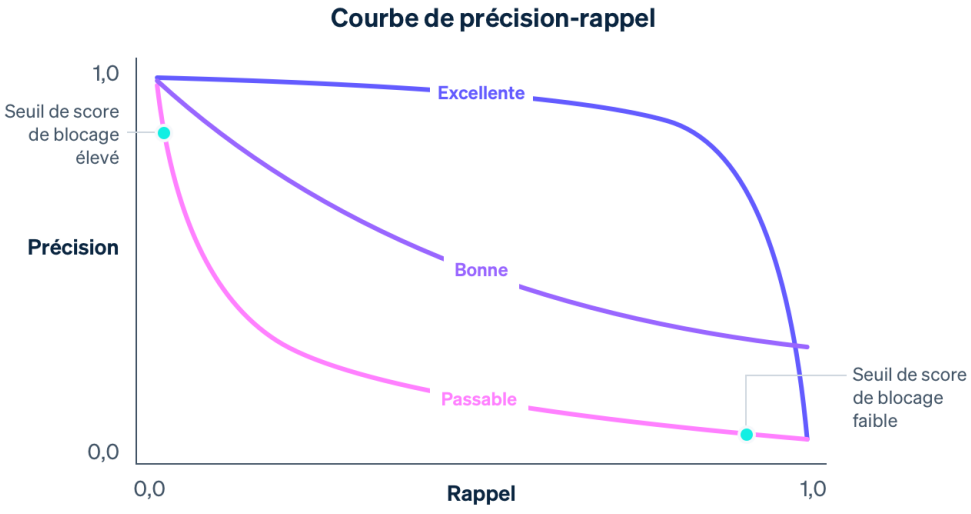


FIGURE 3.7 – Exemple de courbe Précision-Rappel

Il est davantage conseillé d'utiliser l'AUC-PR que l'AUC-ROC[SAI15] lorsque l'on est en présence d'un jeu de données déséquilibré (c'est-à-dire que la variable cible a une distribution d'observations non proportionnelle). En effet, la courbe ROC a tendance à dépeindre une performance du modèle trop optimiste, notamment à cause de la présence du nombre de vrais négatifs dans la construction de la courbe. Le modèle va facilement prédire la classe majoritaire et cela peut faire croire que le modèle est très performant, alors qu'il faut également que le modèle puisse bien prédire la classe minoritaire. La courbe Précision-Rappel se focalise donc davantage sur la prédiction de la classe minoritaire. Par le même raisonnement, nous pouvons déjà conclure que la précision totale (*accuracy*) n'est pas non plus un bon indicateur de prédiction de la classe positive. Nous nous intéresserons surtout au F1-score et à l'aire sous la courbe Précision-Rappel (AUC-PR).

### Courbe de gain

C'est un outil d'évaluation couramment utilisé en apprentissage automatique[SER20] qui permet de comparer la capacité prédictive d'un modèle par rapport à une prédiction aléatoire. Elle est construite à partir du taux de vrais positifs (précision) en fonction du taux de positifs prédits par le modèle. Elle est généralement utilisée dans le contexte de problèmes de classification où le jeu de données est déséquilibré (la classe positive est relativement plus rare que la classe négative). La courbe de gain montre comment les performances du modèles s'améliorent par rapport à une prédiction aléatoire (représentée par une droite d'équation $y = 1$). Un modèle avec une courbe de gain supérieure à cette droite est considéré comme ayant une meilleure performance. La figure ci-dessous présente un exemple de courbe de gain.

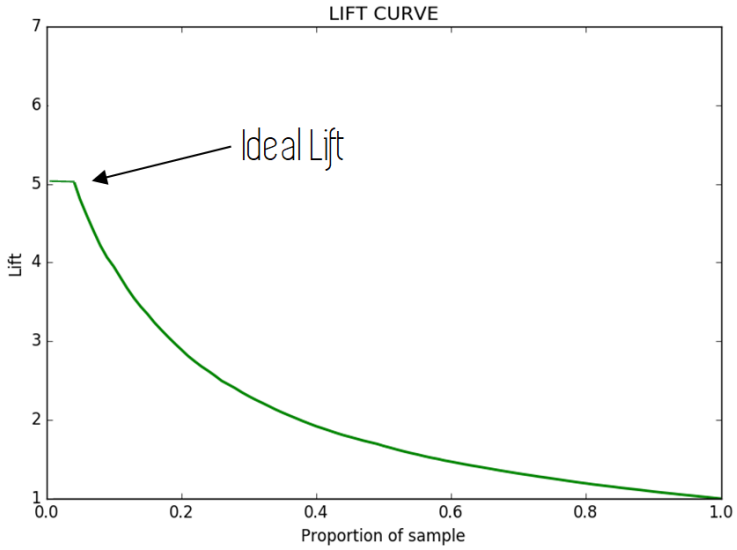


FIGURE 3.8 – Exemple de courbe de gain

La courbe de gain est également un bon outil pour décider du seuil de probabilité de classification qui correspond à des objectifs commerciaux spécifiques. Elle peut aider à prendre des décisions pour allouer les ressources de manière plus efficace en les concentrant sur les exemples les plus susceptibles d'appartenir à la classe positive, en trouvant un compromis entre la détection de vrais positifs et la réduction des faux positifs. Il est donc intéressant que le maximum de la courbe de gain soit atteint pour des proportions faibles de l'échantillon lorsque les campagnes marketing ont des budgets et des ressources très limités et que le nombre de clients est élevé. Ainsi, les ressources peuvent être allouées aux quelques clients les plus « appétents » à la résiliation tout en gardant une certaine confiance dans le pouvoir prédictif du modèle.

Maintenant que nous avons présenté les différents modèles auxquels nous allons nous intéresser et les métriques avec lesquelles nous allons les comparer, nous pouvons désormais les appliquer.

## 3.2 Application des modèles

Nous allons d'abord entraîner nos quatre modèles (CART, Random Forest, régression logistique et CatBoost) sur les données d'une seule Banque Populaire qui servira de banque pilote. Nous voulons choisir une banque ayant un profil moyen par rapport aux autres, afin de ne pas entraîner notre modèle sur une banque avec un comportement trop spécifique.

Regardons le nombre de contrats et le taux de résiliation précoce par Banque Populaire. Les noms des banques ont été anonymisées par souci de confidentialité.

| Établissement | Nombre de contrats | Taux de résiliation précoce |
|---|---|---|
| BP A | 31 262 | 5,5% |
| BP B | 31 240 | 6,6% |
| BP C | 30 836 | 6,1% |
| BP D | 30 037 | 5,5% |
| BP E | 28 930 | 7,6% |
| BP F | 27 607 | 7,3% |
| BP G | 25 172 | 10,9% |
| BP H | 24 243 | 11,1% |
| BP I | 18 586 | 5,4% |
| BP J | 18 087 | 4,5% |
| BP K | 15 875 | 6,1% |

TABLE 3.1 – Comparaison des Banques Populaires (nombre de contrats et taux de résiliation)

La Banque Populaire F est celle qui présente le profil le plus représentatif parmi tous les établissements du réseau. C'est donc sur les données de cette banque que nous allons entraîner nos modèles. Pour rappel, nous disposons des données des contrats Assurance Famille souscrits avant mai 2022, en cours à 2 mois d'ancienneté, ce qui représente environ 280 000 contrats.

Nous isolons les contrats des cohortes de juin 2021 à mai 2022 afin qu'elles servent de données de validation. Il est important de prendre les données d'une année entière afin de capter la saisonnalité des résiliations. Les données de validation représente près de 30% des données totales, soit 85 000 contrats. Les données des contrats souscrits entre avril 2019 et mai 2021 ont été séparées en échantillon d'apprentissage (*train dataset*) et en échantillon de test (*test dataset*) dans les proportions de 80% et 20% respectivement, en s'assurant que le taux de résiliation précoce soit le même dans les deux échantillons. Nous utilisons notamment les libraires Python *sklearn* et *catboost*.

Dans un premier temps, nous regardons les corrélations entre les différentes variables. C'est une étape préalable importante car cela permet d'éviter la multicollinéarité : lorsque deux variables sont fortement corrélées, elles peuvent apporter des informations redondantes

aux modèles d'apprentissage automatique. Cela peut rendre le modèle plus instable et compliquer l'interprétation des résultats. Nous utilisons la fonction *ProfileReport* de la librairie *pandas* pour générer une analyse complète de nos données d'entraînement et de test. Un extrait du rapport généré est disponible en annexe.

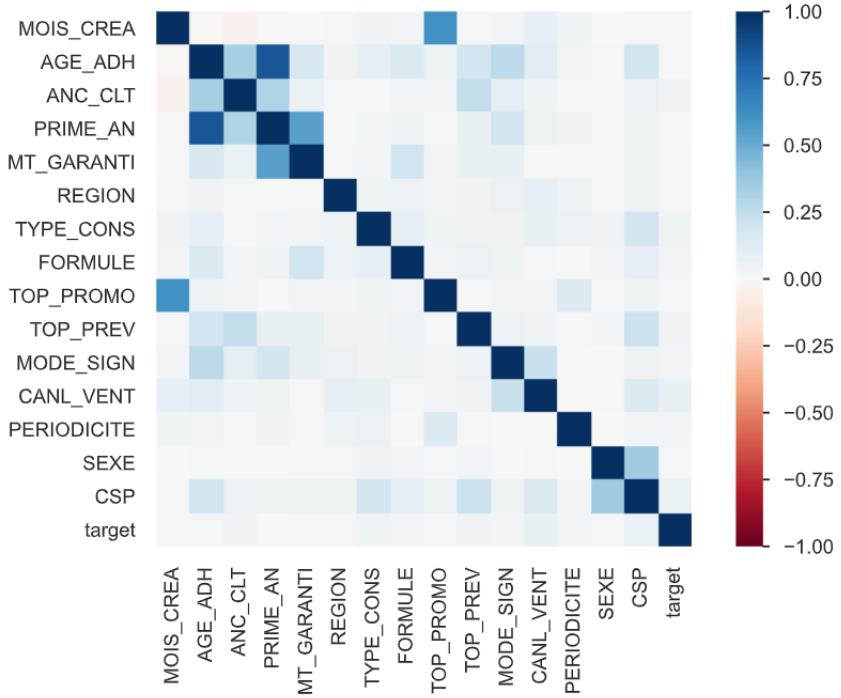


FIGURE 3.9 – Corrélations - BP F

Nous pouvons voir que très peu de variables sont corrélées à la variable cible (*target*). Cela indique que les liens entre les variables prédictives et la résiliation précoce ne sont pas évidents et sont probablement complexes. Le montant de prime annuelle et l'âge du client à l'adhésion sont fortement corrélés positivement, ce qui est normal. Le montant garanti est également corrélé positivement à la prime. Un point intéressant est que le top promo (i.e. si le contrat a été souscrit dans le cadre d'une promotion) est corrélé avec le mois de création du contrat, montrant donc l'impact des offres promotionnelles dans la vente des contrats.

### 3.2.1 Pré-traitement des données

Parmi les modèles que nous avons choisi de mettre en place, certains nécessitent des retraitements de données particuliers. En effet, les arbres de classification (*DecisionTreeClassifier*) ne prennent pas en charge les données manquantes, ni les variables catégorielles, du moins pas nativement. En revanche la mise à l'échelle (*scaling*) des variables n'est pas utile. Ce n'est pas le cas de la régression logistique, qui nécessite tous les retraitements de données

mentionnés. L'algorithme CatBoost, quant à lui, ne requiert aucun retraitement spécifique à part le traitement des données manquantes. Voyons comment nous allons pré-traiter nos données.

### Traitement des données manquantes

Pour retraiter convenablement les valeurs manquantes, il faut bien regarder quelles variables sont concernées et comprendre la raison de leur présence. Il existe de nombreuses méthodes pour traiter les valeurs manquantes dans un ensemble de données[KUM20] :

- Suppression des lignes ou colonnes

  Cette méthode est très simple à mettre en place mais entraîne forcément une perte d'information qui peut s'avérer importante.

- Imputation moyenne/médiane

  Cela permet de conserver la taille de l'échantillon mais peut introduire un biais si les valeurs manquantes sont significatives.

- Imputation par l'exemple le plus proche

  Plus difficile à mettre en place, cette méthode utilise des observations similaires pour estimer les valeurs manquantes, ce qui peut être plus précis mais également sensible à la distribution des données.

- Création d'une catégorie « valeur manquante »

  Cette technique s'avère utile lorsque l'on est en présence d'un grand nombre de données manquantes. Cependant elle peut aussi introduire du bruit, ce qui n'est pas désirable.

Regardons le taux de complétude des différentes variables présentes dans nos données (pour toutes les Banques Populaires).

| Variable | Taux de complétude |
|---|---|
| Région de résidence du client | 100% |
| Nom d'agence | 99,4% |
| Type de conseiller | 97,5% |
| Formule | 100% |
| Top promo | 100% |
| Top prévoyance | 100% |
| Mode de signature | 93,8% |
| Canal de vente | 99,7% |
| Mois de création du contrat | 100% |
| Périodicité de paiement de la prime | 100% |

| | |
|---|---|
| Age du client à l'adhésion | 100% |
| Ancienneté du client à l'adhésion | 100% |
| Sexe du client | 100% |
| Catégorie socio-professionelle | 74,7% |
| Montant de prime | 100% |
| Montant du capital garanti | 100% |

TABLE 3.2 – Taux de complétude des données

Concernant le nom de l'agence et le type du conseiller, les valeurs manquantes peuvent donner une information sur la variable cible. En effet ces valeurs manquantes sont présentes car elles correspondent à des contrats dont les agences et les conseillers étaient présents à leur souscription mais qui ne le sont plus maintenant. Nous choisissons donc de créer une modalité « Non renseigné » lorsque ces variables sont manquantes. Le taux relativement élevé de valeurs manquantes de la catégorie socioprofessionnelle nous pousse à adopter la même stratégie, toutefois pour des raisons différentes : le taux semble trop élevé pour remplacer les données manquantes sans introduire de bruit.

Pour toutes les autres variables, les quelques valeurs manquantes seront remplacées par la modalité la plus fréquente. Cela se justifie par leur taux faible de valeurs manquantes.

**Encodage des variables catégorielles**

De nombreuses méthodes d'encodage différentes sont utilisées pour transformer des données catégorielles en format numérique compréhensible par les algorithmes de *machine learning*. Voici les méthodes d'encodage les plus courantes[GAR22] :

- Encodage par étiquette (*Label Encoding*)
  Chaque modalité est assignée à une valeur numérique unique. C'est une méthode simple à mettre en place et qui fonctionne bien pour les variables ordinales où il existe un ordre naturel entre les modalités. Cependant, cela peut créer une relation d'ordre artificielle dans les variables nominatives, ce qui peut ensuite biaiser les modèles.
- Encodage à chaud (*One-Hot Encoding*)
  Chaque catégorie est transformée en un vecteur binaire. Cela évite la création d'une relation d'ordre artificielle, mais entraîne une augmentation du nombre de dimensions, notamment pour les variables catégorielles avec un grand nombre de modalités.
- Encodage par fréquence (*Frequency Encoding*)

Les catégories sont remplacées par la fréquence de leur modalité dans les données. Cette méthode peut aider à capter des informations sur la fréquence des modalités mais elle peut également entraîner une perte d'information si beaucoup de modalités ont la même fréquence.

- Encodage par cible (*Target Encoding*)
  Pour chaque catégorie, la valeur moyenne de la variable cible est calculée. C'est une méthode très pratique pour les variables catégorielles présentant de nombreuses modalités. Cela peut toutefois introduire du bruit ou conduire à un sur-apprentissage.

Regardons les variables catégorielles que nous avons ainsi que le nombre de modalités de chaque variable catégorielle après imputation des données manquantes. Selon les variables, la méthode d'encodage ne sera pas forcément la même.

| **Variable** | **Variable catégorielle** | **Nombre de modalités** |
|---|---|---|
| Région de résidence du client | Oui | > 10 |
| Nom d'agence | Oui | > 10 |
| Type de conseiller | Oui | 7 |
| Formule | Oui | 4 |
| Top promo | Non | |
| Top prévoyance | Non | |
| Mode de signature | Oui | 2 |
| Canal de vente | Oui | 2 |
| Mois de création du contrat | Non | |
| Périodicité de paiement de la prime | Oui | 4 |
| Age du client à l'adhésion | Non | |
| Ancienneté du client à l'adhésion | Non | |
| Sexe du client | Oui | 2 |
| Catégorie socio-professionelle | Oui | > 10 |
| Montant de prime | Non | |
| Montant du capital garanti | Non | |

TABLE 3.3 – Liste des variables catégorielles

Nous choisissons le *Label Encoding* pour les variables « Formule », « Périodicité de paiement de la prime » et « Sexe du client ». Les variables « Mode de signature », « Canal de vente » sont transformées via *One-Hot Encoding*. La variable « Type de conseiller » est

encodée par *Frequency Encoding*. Les autres variables catégorielles ayant un grand nombre de modalités, la technique d'encodage choisie est le *Weight of Evidence Encoding*.

Cette méthode est très proche du *Target Encoding*. En classification binaire, la variable cible vaut 1 lorsqu'un certain évènement se produit (ici la résiliation précoce) et 0 si l'évènement ne se produit pas. Pour chaque modalité $m$ d'une variable catégorielle $X$, le *Weight of Evidence Encoding* va attribuer son *log odds ratio* :

$$\ln \frac{P(Y = 1|X = m)}{1 - P(Y = 1|X = m)}$$

avec $Y$ la variable cible. Cette méthode d'encodage permet d'exploiter la relation entre la variable cible et la distribution de chaque modalité. Comparativement au *Target Encoding*, elle réduit le risque de sur-apprentissage en réduisant le bruit des observations, ce qui la rend plus robuste, bien que l'interprétation des valeurs obtenues soient moins intuitives à interpréter.

### Ré-échantillonnage

En présence d'un jeu de données déséquilibré, les modèles peuvent avoir du mal à prédire correctement la classe minoritaire, dû à leur faible présence dans l'échantillon. Cela conduit à des prédictions biaisées et affecte donc les performances des modèles. C'est pourquoi le ré-échantillonnage est nécessaire dans ce contexte. Les méthodes de ré-échantillonnage les plus courantes sont[ALE17] :

- Sous-échantillonnage (*Undersampling*)

  Le nombre d'observations de la classe majoritaire est réduit, aidant les modèles à éviter de surestimer cette classe. Cela peut néanmoins entraîner une perte d'information importante à cause de la suppression des données.

- Sur-échantillonnage (*Oversampling*)

  Le nombre d'observations de la classe minoritaire est augmenté, permettant aux modèles d'apprentissage de mieux capter les informations de cette classe. Cependant, en créant des duplicatas de données existantes, cela augmente le risque de sur-apprentissage (*overfitting*) et introduit du bruit dans les données d'entraînement.

- SMOTE (*Synthetic Minority Oversampling Technique*)

  Cette technique génère des observations synthétiques pour la classe minoritaire en se basant sur des observations similaires. Cela limite l'*overfitting* mais peut introduire beaucoup de bruit.

- Utilisation de poids de classes (*class weight*)

Chaque classe se voit affecter un poids différent afin de tenir compte du déséquilibre. Toutefois cette méthode peut ne pas être suffisante si la séparation des classes est très difficile.

Regardons comment ces différentes méthodes de ré-échantillonnage affectent un modèle. Le modèle choisi ici est un arbre de décision avec un paramétrage par défaut. Les données d'entraînement sont celles de la Banque Populaire F et les indicateurs de performance ci-dessous ont été calculés sur la base de test.

| **Méthode** | **Accuracy** | **Précision** | **Rappel** | **F1-score** | **AUC-ROC** | **AUC-PR** |
|---|---|---|---|---|---|---|
| Aucune | 82,6% | 12,2% | 23,9% | 16,1% | 55,5% | 8,2% |
| *Undersampling* | 46,3% | 8,9% | 72,2% | 15,9% | 58,3% | 8,4% |
| SMOTE | 27,6% | 7,8% | 82,6% | 14,3% | 52,9% | 7,7% |
| *Oversampling* | 86,2% | 9,3% | 11,1% | 10,5% | 51,5% | 7,3% |
| Poids | 85,8% | 9,2% | 11,5% | 10,2% | 51,5% | 7,3% |

TABLE 3.4 – Impact du ré-échantillonnage sur les performances d'un arbre de décision

La métrique qui nous intéresse le plus est le F1-score. La méthode de ré-échantillonnage obtenant le F1-score le plus élevé est celle sans ré-échantillonnage particulier. Dans la suite, nous testerons différents poids de classes afin de trouver les modèles optimaux.

### 3.2.2 Application à une Banque Populaire

Nous pouvons désormais entraîner nos modèles. Le nombre de variables n'étant pas très important, il n'est pas nécessaire de faire une étape préalable de sélection de variables. Nous souhaitons, pour chaque modèle, trouver les hyperparamètres qui permettent de maximiser le F1-score. L'optimisation des hyperparamètres se fait par un principe de validation croisée grâce à la fonction *GridSearchCV*.

La validation croisée consiste à diviser une base de données en $k$ parties : $k-1$ serviront pour apprendre les données et la dernière partie sera utilisée lors de la phase de test. Le découpage est répété aléatoirement $k$ fois. Les métriques calculées via cette opération sont obtenues par la moyenne des métriques sur chacun des découpages.

### Arbre de classification

Le premier modèle que nous cherchons à optimiser est un arbre de classification (*DecisionTreeClassifier*). Le pré-traitement des données adéquat a été effectué sur les données d'entraînement et les données de test séparément, à savoir traitement des valeurs manquantes et encodage des variables catégorielles. Les hyperparamètres que nous voulons optimiser pour ce modèle sont :

- *criterion* : le critère de partitionnement des noeuds. Le critère par défaut est le coefficient de Gini.
- *max_depth* : la profondeur maximale de l'arbre. Pas de valeur par défaut.
- *min_samples_split* : le nombre d'observations minimal requis pour un partitionnement. La valeur par défaut est 2.
- *min_samples_leaf* : le nombre d'observation minimal requis dans une feuille. La valeur par défaut est 1.
- *class_weight* : poids associés à chaque classe. Pas de valeur par défaut.

Suite aux résultats de l'hyperparamétrisation, le modèle avec les hyperparamètres suivants a été retenu :

- *criterion* = 'entropy'
- *max_depth* = 5
- *min_samples_split* = 2
- *min_samples_leaf* = 50
- *class_weight* = {0 :1, 1 :8}

Les indicateurs de performance de ce modèle, calculés sur la base de test, sont présentés ci-dessous.

| | **Prédictions** | | |
|---|---|---|---|
| **Observations** | Non résilié | Résilié | Total |
| Non résilié | 2 302 | 800 | 3 102 |
| Résilié | 82 | 152 | 234 |
| Total | 2 384 | 952 | 3 336 |

TABLE 3.5 – Matrice de confusion - Arbre de classification

| Accuracy | Précision | Rappel | F1-score | AUC-ROC | AUC-PR |
|---|---|---|---|---|---|
| 73,6% | 15,9% | 64,9% | 25,6% | 74,3% | 17,2% |

TABLE 3.6 – Indicateurs de performance - Arbre de classification

Nous pouvons déjà remarquer que ce modèle prédit un grand nombre de faux positifs. Les métriques obtenues ne sont pas très élevées, notamment le F1-score, cependant cela ne veut pas dire que les prédictions faites par le modèle ne sont pas exploitables. Nous allons regarder la courbe de gain du modèle (*lift curve*).

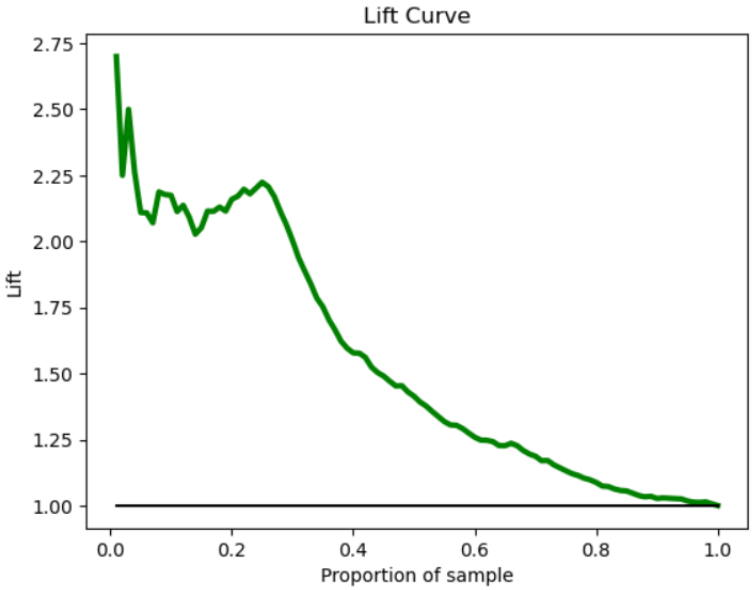


FIGURE 3.10 – Courbe de gain - Arbre de classification

D'après la figure ci-dessus, l'arbre de classification optimisé peut prédire la résiliation précoce jusqu'à 2,7 fois mieux que le hasard. La courbe de gain atteint son maximum si on prend un très faible pourcentage des observations de l'échantillon de test ayant la plus forte probabilité de résiliation. Cela signifie que si nous voulons lancer une campagne marketing avec pour but d'empêcher les nouveaux clients de résilier, le mieux serait de cibler un faible pourcentage de clients que le modèle prédit comme les plus « appétents » à la résiliation précoce.

Maintenant que nous avons pu juger de la qualité prédictive de l'arbre de décision, concentrons-nous sur son interprétation. L'arbre optimisé est représenté ci-dessous, bien que ce soit une version légèrement simplifiée afin de le rendre plus lisible.

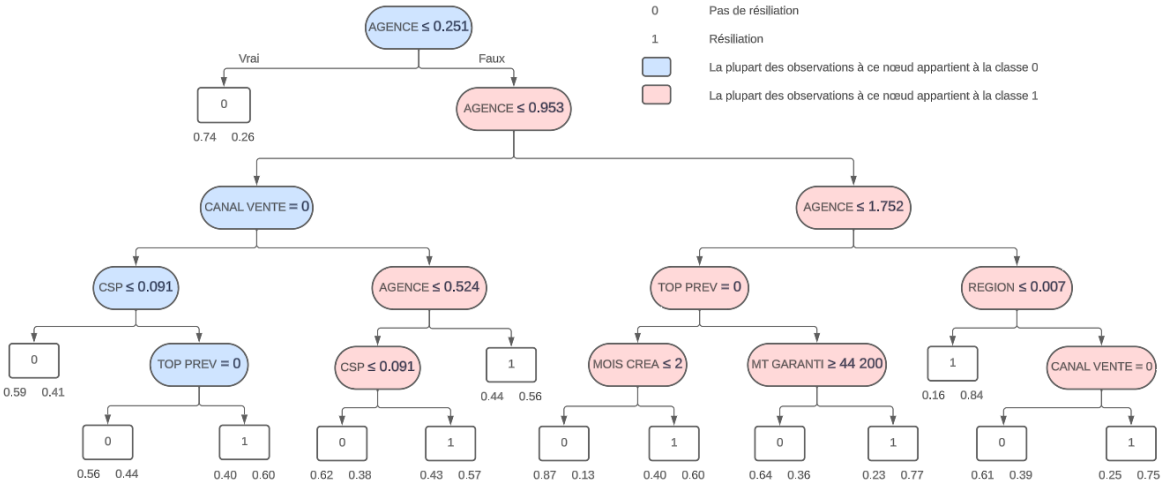


FIGURE 3.11 – Arbre de décision simplifié - Banque Populaire F

Nous pouvons voir que l'arbre sépare les agences en plusieurs catégories selon la valeur de leur *log odds ratio*. Parmi les agences dont le *log odds ratio* est inférieur à 0.251 (ce que l'on pourrait qualifier de « bonnes » agences), 74% des contrats souscrits dans ces agences n'ont pas été résiliés au cours de leur première année d'effet. Parmi les contrats souscrits dans des agences « moyennes » (*log odds ratio* inférieur à 0.953), 59% ne sont pas résiliés si la vente du contrat s'est effectuée à distance et si la catégorie socioprofessionnelle correspond à « Cadres », « Employés » ou « Professions Intermédiaires ». Pour les agences et les régions de résidence des clients dont les contrats sont les plus susceptibles de résilier de manière précoce, 75% sont résiliés si la vente s'est effectuée en face à face.

Regardons maintenant les variables les plus importantes de l'arbre de décision. L'importance d'une variable est définie par le nombre de fois où cette variable sert de critère de partitionnement pour tous les noeuds de l'arbre de décision.

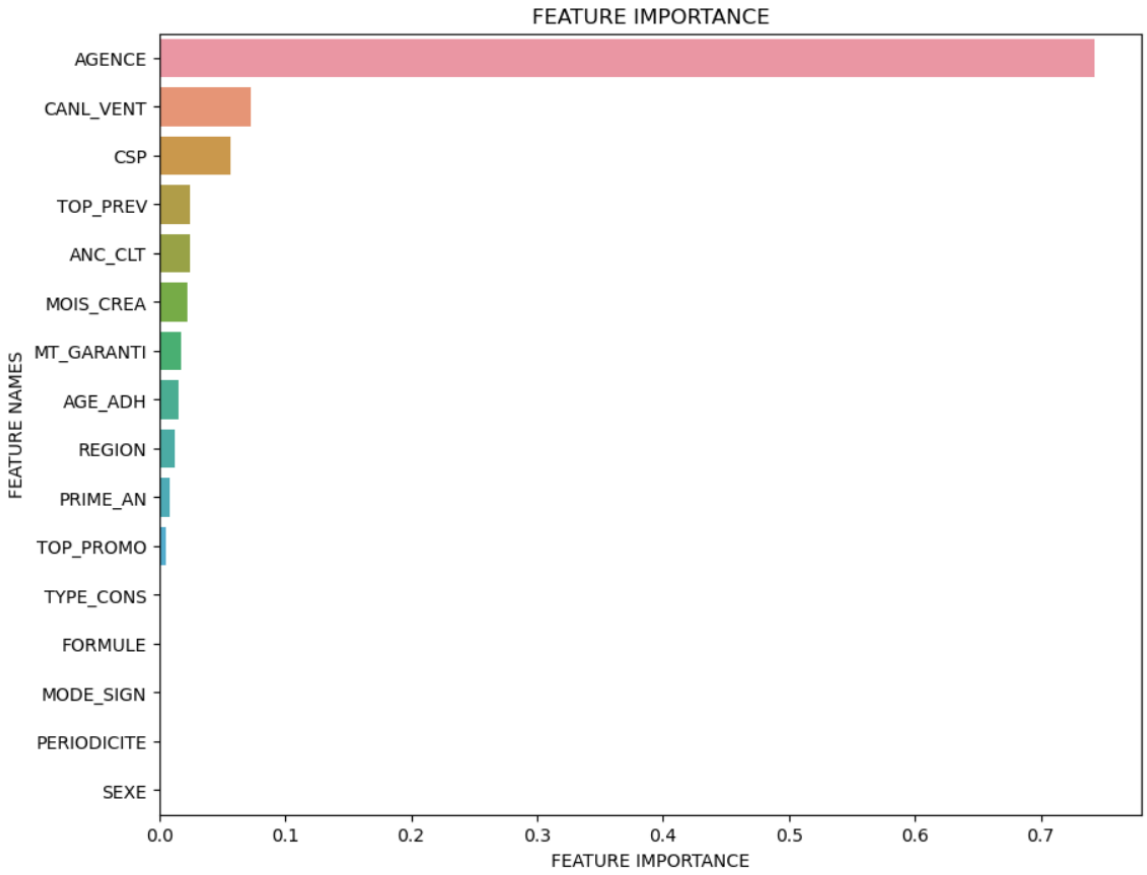


FIGURE 3.12 – Importance des variables - Arbre de décision

Nous observons que les variables les plus importantes pour le modèle sont l'agence, le canal de vente, la catégorie socioprofessionnelle. Nous allons pouvoir comparer ces résultats avec les autres modèles.

### Random Forest

Les hyperparamètres que nous avons cherché à optimiser pour ce modèle sont :

- *n_estimators* : le nombre d'arbres à entraîner. La valeur par défaut est 100.
- *criterion* : le critère de partitionnement des noeuds. Le critère par défaut est le coefficient de Gini.
- *max_depth* : la profondeur maximale de l'arbre. Pas de valeur par défaut.
- *min_samples_split* : le nombre d'observations minimal requis pour un partitionnement. La valeur par défaut est 2.
- *min_samples_leaf* : le nombre d'observation minimal requis dans une feuille. La valeur par défaut est 1.

- *class_weight* : poids associés à chaque classe. Pas de valeur par défaut.

Suite aux résultats de l'hyperparamétrisation, le modèle avec les hyperparamètres suivants a été retenu :

- *criterion* = 'gini'
- *max_depth* = 20
- *min_samples_split* = 2
- *min_samples_leaf* = 20
- *class_weight* = {0 :1, 1 :9}

Les indicateurs de performance de ce modèle, calculés sur la base de test, sont présentés ci-dessous.

| | **Prédictions** | | |
|---|---|---|---|
| **Observations** | Non résilié | Résilié | Total |
| Non résilié | 2 653 | 449 | 3 102 |
| Résilié | 127 | 107 | 234 |
| Total | 2 780 | 556 | 3 336 |

Table 3.7 – Matrice de confusion - Random Forest

| **Accuracy** | **Précision** | **Rappel** | **F1-score** | **AUC-ROC** | **AUC-PR** |
|---|---|---|---|---|---|
| 82,7% | 19,2% | 45,7% | 27,1% | 78,7% | 22,3% |

Table 3.8 – Indicateurs de performance - Random Forest

Toutes les métriques sont plus élevées pour le Random Forest que pour l'arbre de classification. Cela n'est pas étonnant compte tenu de la nature de ces modèles. Nous allons également regarder la courbe de gain.

Le Random Forest optimisé peut prédire jusqu'à 3 fois mieux la résiliation précoce que le hasard, ce qui est légèrement mieux que l'arbre de classification. Ce maximum est atteint pour environ 5% des observations de l'échantillon de test ayant la plus forte probabilité de résiliation.

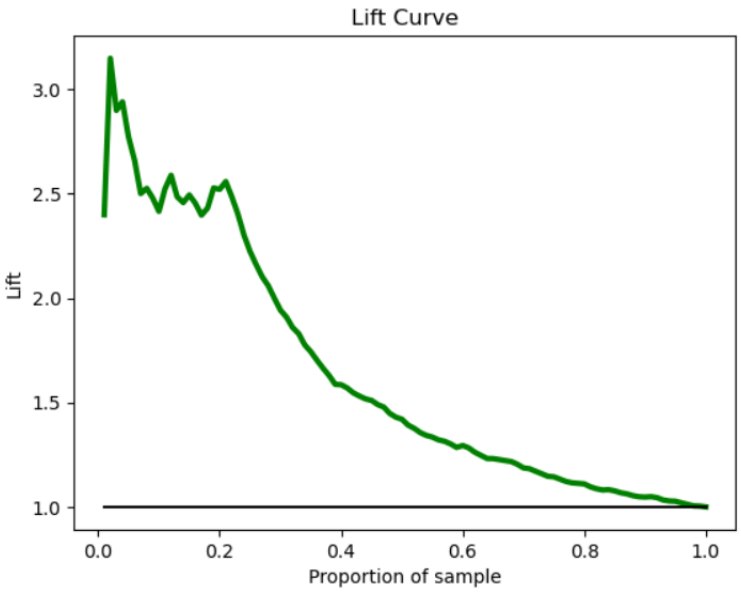


FIGURE 3.13 – Courbe lift - Random Forest

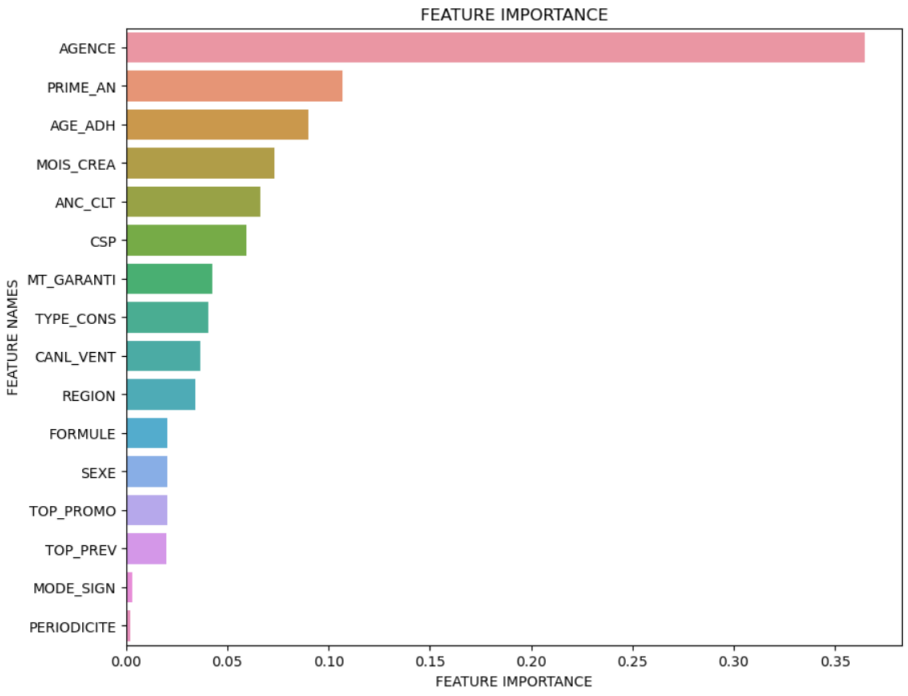


FIGURE 3.14 – Importance des variables - Random Forest

Encore une fois, la variable la plus importante pour ce modèle est l'agence de distribution du contrat. Elle est suivie du montant de prime, de l'âge du client à l'adhésion. Ces variables sont différentes de celles dont la *feature importance* était grande pour l'arbre de classification.

Nous allons continuer à regarder les importances des variables des autres modèles.

Intéressons-nous maintenant à l'interprétabilité du modèle. Les valeurs de Shapley ont été calculées grâce à la librairie *SHAP* de Python.

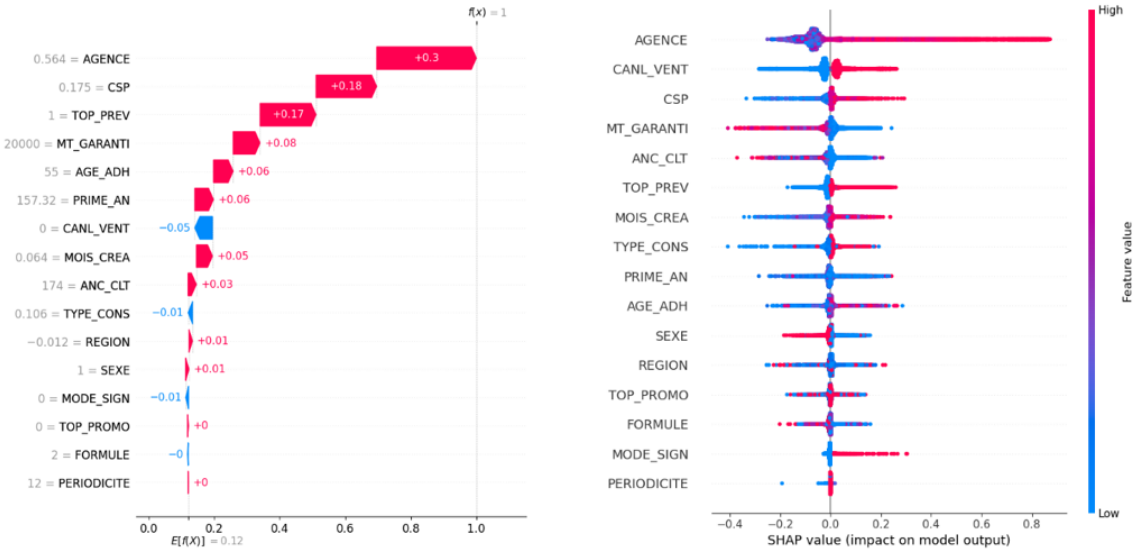


FIGURE 3.15 – Valeurs de Shapley - Random Forest

À gauche de la figure ci-dessus sont représentées les valeurs de Shapley pour un contrat résilié. Les variables qui ont le plus contribué à éloigner la prévision de ce contrat de l'espérance des prédictions sont l'agence, la catégorie socioprofessionnelle et le top prévoyance. Ces variables figuraient déjà dans les *features* les plus importantes de l'arbre de classification.

Le côté droit de la figure montre quant à lui l'impact des modalités des différentes variables prédictives sur les valeurs de Shapley. Les variables sont rangées par ordre croissant d'importance. Ainsi, les agences avec de grandes valeurs (*log odds ratios* élevés) ont beaucoup d'impact. Nous pouvons remarquer que ces impacts coïncident avec les statistiques descriptives et de l'étude des *odds ratios* effectuées précédemment. En effet, il en ressort que le canal de vente en face à face, les montants garantis faibles, les petites anciennetés client et les âges à l'adhésion élevés sont des caractéristiques conduisant à des valeurs de Shapley positives, indiquant qu'elles sont des facteurs de résiliation précoce pour ce modèle.

### CatBoost

Les hyperparamètres que nous avons cherché à optimiser pour ce modèle sont :

- *iterations* : le nombre d'itérations. La valeur par défaut est 1 000.

- *learning_rate* : la fonction qui mesure la qualité d'un partitionnement. La valeur par défaut est 0.03.
- *depth* : la profondeur maximale de l'arbre. La valeur par défaut est 6.
- *class_weights* : poids associés à chaque classe. Pas de valeur par défaut.

Il existe de très nombreux autres paramètres pour ce type de modèle. Cependant, par souci de temps d'exécution, nous allons nous limiter dans la recherche des hyperparamètres optimaux. Par exemple, le nombre d'itérations maximal que nous avons pris en compte est 500. Suite aux résultats de l'hyperparamétrisation, le modèle avec les hyperparamètres suivants a été retenu :

- *iterations* = 100
- *learning_rate* = 0.2
- *depth* = 5
- *class_weights* = {0 :1, 1 :10}

Les indicateurs de performance de ce modèle, calculés sur la base de test, sont présentés ci-dessous.

| | **Prédictions** | | |
|---|---|---|---|
| **Observations** | Non résilié | Résilié | Total |
| Non résilié | 2 697 | 405 | 3 102 |
| Résilié | 177 | 57 | 234 |
| Total | 2 874 | 462 | 3 336 |

Table 3.9 – Matrice de confusion - CatBoost

| **Accuracy** | **Précision** | **Rappel** | **F1-score** | **AUC-ROC** | **AUC-PR** |
|---|---|---|---|---|---|
| 82,6% | 12,3% | 24,4% | 16,4% | 64,2% | 17,1% |

Table 3.10 – Indicateurs de performance - CatBoost

Les métriques sont moins bonnes que pour l'arbre de classification et la forêt aléatoire. Cela peut être dû au fait que nous nous sommes limités dans le choix des valeurs des différents hyperparamètres, notamment le nombre d'itérations du modèle. Cela peut aussi montrer à quel point l'encodage des variables catégorielles, notamment le *Weight of Evidence encoding* peut impacter les capacités prédictives du modèle lorsque la frontière entre les classes positive et négative n'est pas bien délimitée.

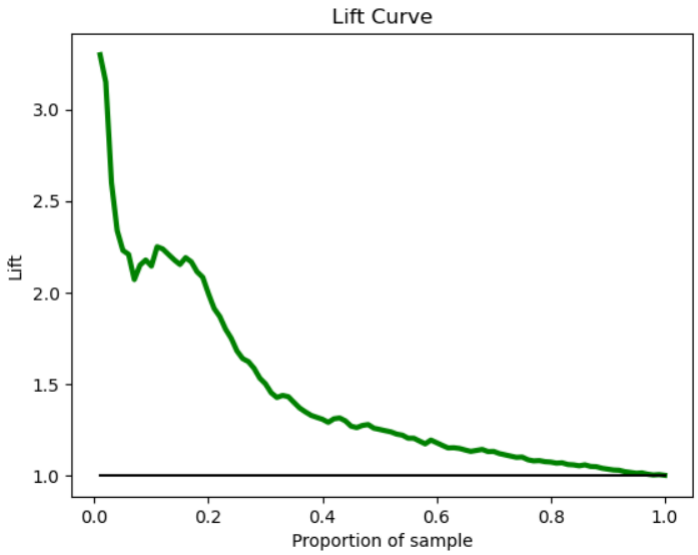


FIGURE 3.16 – Courbe lift - CatBoost

La courbe de gain atteint pour ce modèle un maximum de 3,2 fois mieux que le hasard pour de faibles pourcentages des observations de l'échantillon de test ayant la plus forte probabilité de résiliation (moins de 5%). C'est relativement mieux que pour les deux autres modèles que nous avons implémenté. Regardons désormais les variables les plus importantes de l'algorithme CatBoost.

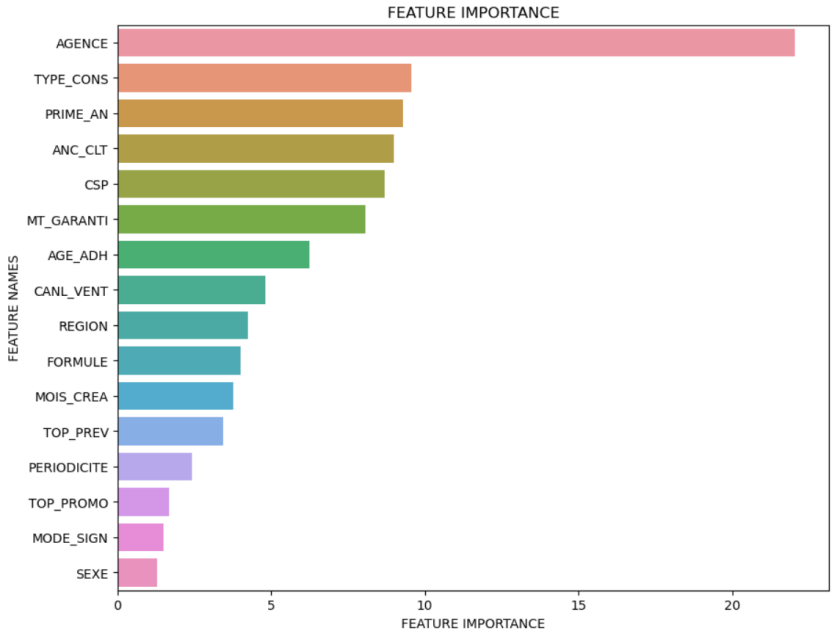


FIGURE 3.17 – Importance des variables - CatBoost

Les variables les plus importantes pour le modèle sont l'agence, le type de conseiller et le montant de prime annuelle. Ces variables figuraient déjà en partie parmi les plus importantes du *Random Forest*. Toutefois, pour interpréter ce modèle, analysons les valeurs de Shapley. L'observation présentée est la même que pour celle du *Random Forest*.

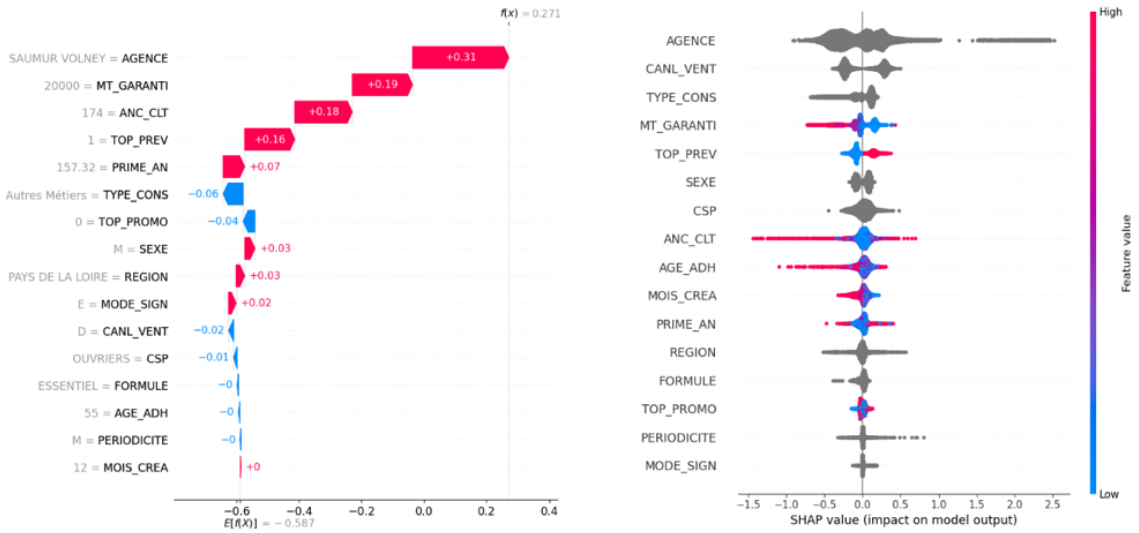


FIGURE 3.18 – Valeurs de Shapley - CatBoost

Pour cette observation, l'agence reste la variable qui influence le plus la différence entre la prédiction du CatBoost et l'ensemble des prédictions. Les autres variables importantes sont le montant garanti, l'ancienneté client et le top prévoyance. Toutefois, au global, les variables influençant le plus les valeurs de Shapley sont l'agence, le canal de vente et le type de conseiller.

### Régression logistique

Les hyperparamètres que nous avons cherché à optimiser pour ce modèle sont :

- *tol* : le critère d'arrêt de l'algorithme d'optimisation. La valeur par défaut est 1e-4.
- *C* : coefficient de régularisation (plus la valeur est petite et plus la régularisation est forte). La valeur par défaut est 1.
- *solver* : algorithme utilisé pour résoudre le problème d'optimisation. L'algorithme utilisé par défaut est Limited-memory BFGS.
- *max_iter* : nombre maximum d'itérations pour que le solveur converge. La valeur par défaut est 100.
- *class_weight* : poids associés à chaque classe. Pas de valeur par défaut

Suite aux résultats de l'hyperparamétrisation, le modèle avec les hyperparamètres suivants a été retenu

- *tol* = 1e-6
- *C* = 0.1
- *solver* = 'lbfgs'
- *max_iter* = 100
- *class_weight* = {0 :1, 1 :7}

Les indicateurs de performance de ce modèle, calculés sur la base de test, sont présentés ci-dessous.

| | **Prédictions** | | |
|---|---|---|---|
| **Observations** | Non résilié | Résilié | Total |
| Non résilié | 2 676 | 426 | 3 102 |
| Résilié | 116 | 118 | 234 |
| Total | 2 792 | 544 | 3 336 |

TABLE 3.11 – Matrice de confusion - Régression logistique

| **Accuracy** | **Précision** | **Rappel** | **F1-score** | **AUC-ROC** | **AUC-PR** |
|---|---|---|---|---|---|
| 83,8% | 21,7% | 50,4% | 30,3% | 80% | 25,6% |

TABLE 3.12 – Indicateurs de performance - Régression logistique

La régression logistique optimisée donne le F1-score et l'AUC-PR les plus élevés. C'est le meilleur modèle parmi les quatre que nous avons mis en place.

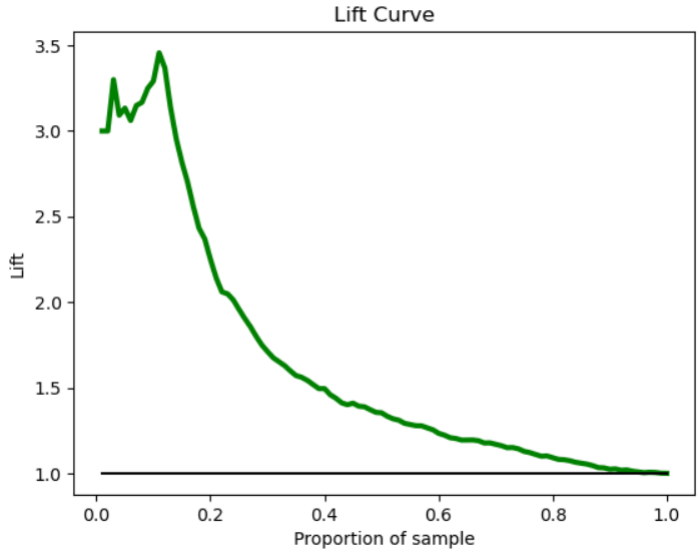


FIGURE 3.19 – Courbe lift - Régression logistique

Ce modèle prédit la résiliation précoce jusqu'à 3,5 fois mieux que le hasard, ce qui est une amélioration en comparaison des trois autres modèles. La courbe de gain atteint son maximum si on prend environ 15% des observations de l'échantillon de test ayant la plus forte probabilité de résiliation. Regardons maintenant la valeur des coefficients que la régression logistique a attribué à chaque variable.

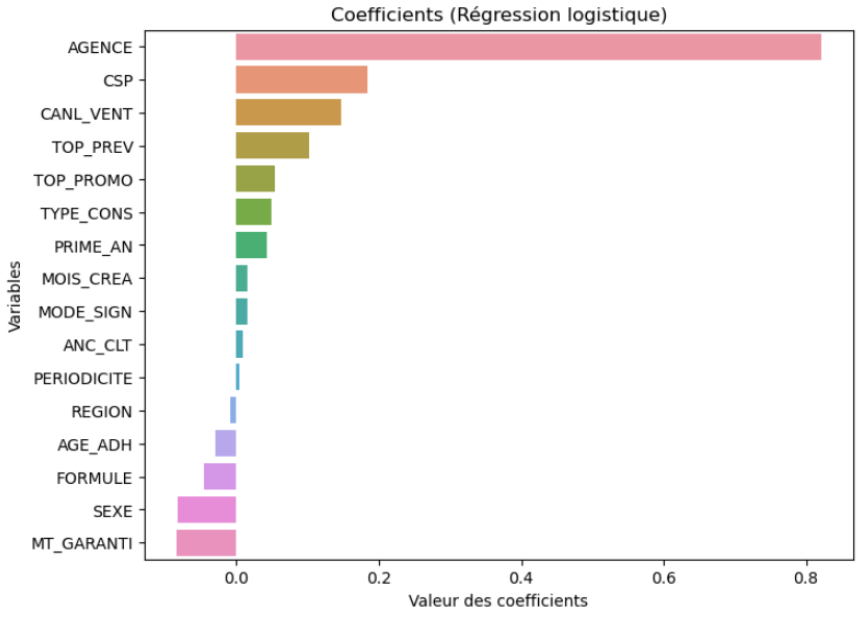


FIGURE 3.20 – Coefficients - Régression logistique

La variable avec le coefficient le plus élevé (en valeur absolue) est l'agence. Les autres variables dont le coefficient est élevé sont la catégorie socioprofessionnelle, le canal de vente et le top prévoyance individuelle.

Les coefficients de la régression logistique, et plus particulièrement leur signe positif ou négatif, nous donne une interprétation directe de l'impact des différentes variables sur la résiliation précoce. La vente en face à face, l'adhésion à de multiples contrats de prévoyance individuelle et les offres promotionnelles contribuent à une probabilité de résiliation précoce (donnée par le modèle) plus élevée. D'après ce modèle, un montant garanti élevé (et donc une formule avec davantage d'options) contribue à une probabilité de résiliation plus faible tandis qu'une prime plus élevée augmente cette probabilité. Ainsi, si la prime est trop élevée par rapport au capital garanti, la probabilité de résiliation augmente. Ces résultats sont cohérents avec les informations que nous avons pu trouver précédemment.

**Frugalité**

À mesure que le *machine learning* gagne en popularité, son impact environnemental est de plus en plus préoccupant. En effet, les modèles d'apprentissage automatique, et notamment de *deep learning*, exigent d'énormes quantités de calculs et entraînent une consommation d'énergie et des émissions de gaz à effet de serre considérables. Pour atténuer cet impact, il est important d'adopter des pratiques plus frugales comme l'utilisation de modèles moins complexes, la gestion de l'efficacité énergétique des infrastructures et l'utilisation d'outils comme la mesure de l'empreinte carbone. La libraire Python *codecarbon*[Mil20] permet de quantifier la consommation énergétique associées à l'exécution d'un code. Nous avons récupéré les consommations énergétiques de nos quatre modèles lors de leur hyperparamétrisation. En effet, un modèle en production sera amené à être entraîné et optimisé régulièrement. Les résultats ont été rassemblés dans le tableau ci-dessous :

| | **Arbre** | **Random Forest** | **CatBoost** | **Régression logistique** |
|---|---|---|---|---|
| Nombre de fits | 10 000 | 10 000 | 720 | 3 600 |
| Consommation | 0,02 kWh | 0,17 kWh | 0,32 kWh | 0.03 kWh |

Table 3.13 – Consommation énergétique de l'hyperparamétrisation des modèles

À titre de comparaison, 1 kWh équivaut à travailler une journée et demie avec un ordinateur portable. Sans surprise, les modèles avec les consommations énergétiques les plus faibles sont l'arbre de classification et la régression logistique, ces modèles étant beaucoup

moins complexes que le *Random Forest* et le CatBoost. Dans notre cas, le modèle le plus performant est donc aussi un modèle avec un faible impact environnemental.

### Calibration de la régression logistique

La régression logistique fournit des scores de probabilités de classification, mais celles-ci ne représentent pas toujours les probabilités réelles qu'une observation appartienne à une classe spécifique. La calibration[sci23] permet d'ajuster les probabilités prédites par le modèle afin qu'elles soient plus fiables. Cela permet d'augmenter la certitude dans nos prédictions.

La calibration par régression isotonique fait partie des techniques utilisées pour la calibration des modèles, notamment les régressions logistiques. Elle produit une série de prédictions $\hat{f}_i$ qui minimisent $\sum_{i=1}^{n}(y_i - \hat{f}_i)^2$ sous la contrainte $\hat{f}_i \geq \hat{f}_j$ lorsque $f_i \geq f_j$, $y_i$ étant la classe réelle de l'observation $i$ et $f_i$ la prédiction donnée par le modèle non calibré. Ces prédictions sont interpolées pour prédire les données non vues. Les prédictions de la régression isotonique forment donc une fonction linéaire par morceaux.

La calibration par régression sigmoïde est quant à elle basée sur le modèle logistique de Platt :

$$p(y_i = 1|f_i) = \frac{1}{1 + exp(Af_i + B)}$$

avec $y_i$ la classe réelle de l'observation $i$. $A$ et $B$ sont des nombres réels déterminés par maximum de vraisemblance. En général, cette méthode est plus efficace pour les échantillons de petite taille ou lorsque le modèle non calibré n'est pas suffisamment fiable et présente des erreurs de calibrage similaires pour les sorties élevées et faibles. Cette méthode de calibration est moins utilisée que la régression isotonique car celle-ci est plus générale du fait que la seule restriction est que la fonction de sortie soit monotone et croissante. Elle est donc plus puissante car elle peut corriger toute distorsion monotone du modèle non calibré. Cependant, la régression isotonique est plus encline à un sur-ajustement, en particulier sur les petits ensembles de données.

Observons ce que donnent les deux méthodes de calibration.

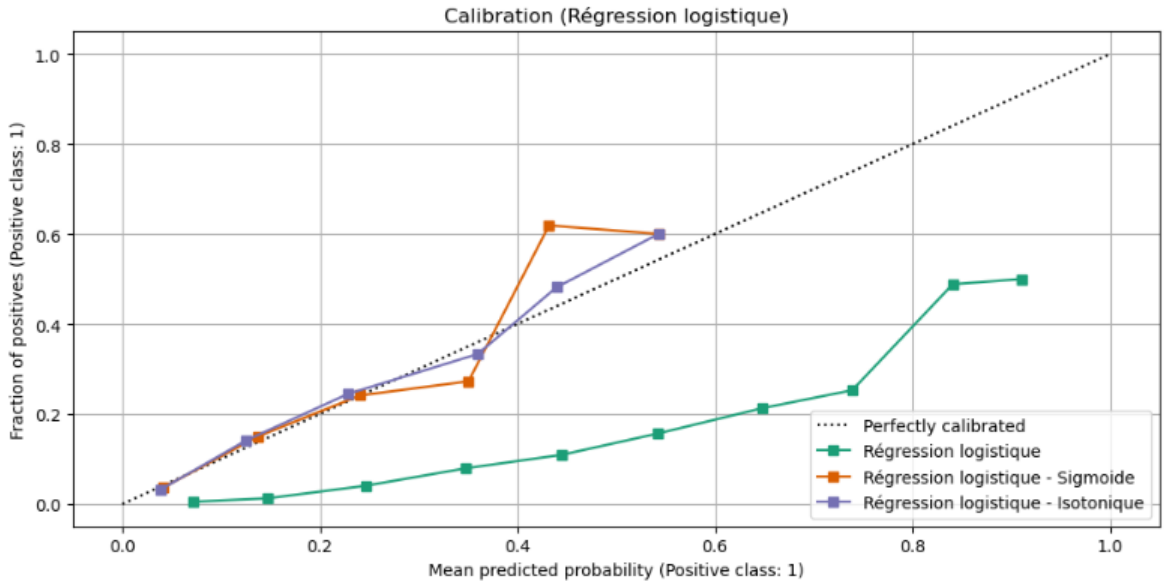


FIGURE 3.21 – Courbes de calibration - Régression logistique

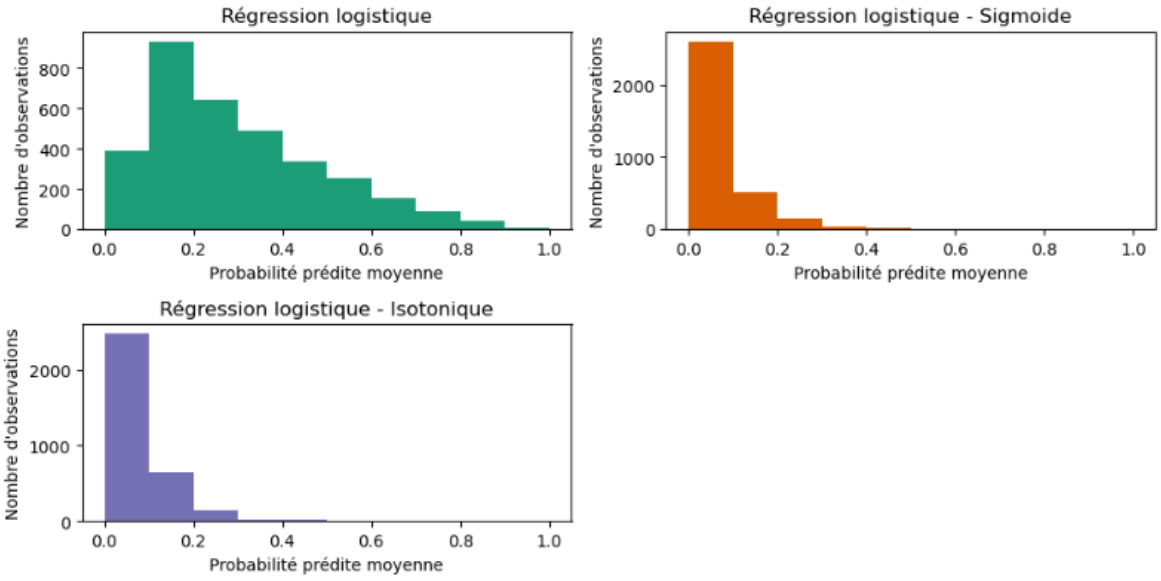


FIGURE 3.22 – Calibration - Régression logistique

Les histogrammes ci-dessus montrent la distribution des probabilités données par la régression logistique et les régressions calibrées. Nous remarquons que la régression logistique non calibrée ne fournit pas de probabilités de résiliation précoce proches de 1.

Nous comparons les deux méthodes de calibration grâce au score de Brier. Dans le cas d'une classification binaire, le score de Brier équivaut à l'erreur quadratique moyenne entre les probabilités du modèle non calibré et celles du modèle calibré.

| Régression Logistique | RL + Sigmoïde | RL + Isotonique |
|---|---|---|
| 0,123 | 0,059 | 0,058 |

Table 3.14 – Scores de Brier

Les deux méthodes de calibration sont très proches l'une de l'autre, toutefois la régression isotonique est légèrement meilleure.

### Conclusion

Pour la Banque Populaire F, le modèle le plus performant d'après nos critères est la régression logistique. En effet, c'est le modèle avec le F1-score et l'AUC-PR les plus élevés. C'est également le modèle avec le *lift* le plus élevé. C'est relativement étonnant car les modèles plus complexes comme le *Random Forest* ou le CatBoost sont connus pour avoir de meilleures performances prédictives. Toutefois la régression logistique donne l'avantage d'être un modèle frugal et facilement explicable. Les résultats pour chaque modèle sont récapitulés ci-dessous :

| Métrique | Arbre | Random Forest | CatBoost | Régression logistique |
|---|---|---|---|---|
| VN | 2 302 | 2 653 | 2 697 | 2 676 |
| VP | 152 | 107 | 57 | 118 |
| FN | 82 | 127 | 177 | 116 |
| FP | 800 | 449 | 405 | 426 |
| Accuracy | 73,5% | 82,7% | 82,6% | 83,8% |
| Précision | 15,9% | 19,2% | 12,3% | 21,7% |
| Rappel | 64,9% | 45,7% | 24,4% | 50,4% |
| F1-score | 25,6% | 27,1% | 16,4% | 30,3% |
| AUC-ROC | 74,3% | 78,7% | 64,2% | 80% |
| AUC-PR | 17,2% | 22,3% | 17,1% | 25,6% |

Table 3.15 – Indicateurs de performance - Tous les modèles

D'après le modèle de régression logistique, la variable « Agence » est la variable la plus déterminante dans le phénomène de résiliation précoce, suivie par la catégorie socioprofessionnelle, le canal de vente, le top prévoyance et le montant garanti. La variable « Agence » est désignée comme la plus importante par tous les modèles. Le canal de vente, la catégorie socioprofessionnelle et le top prévoyance font également partie des variables les plus importantes des quatre modèles.

Le fait que l'agence ressorte particulièrement pour tous les modèles est probablement dû en partie à la façon dont la variable a été encodée. En effet, le *Weight of Evidence encoding* lie directement les différentes modalités à la variable cible, facilitant ainsi sa prise en compte dans les modèles. Afin de nous assurer que l'importance de cette variable n'a pas biaisé les importances des autres variables, nous faisons tourner les mêmes modèles en enlevant l'agence des données d'apprentissage. Finalement, cela ne change pas le rang des importances des autres variables, confirmant ainsi nos résultats.

Ces résultats, ainsi que l'arbre de classification simplifié, ont été présentés au métier afin de savoir s'ils sont cohérents avec la réalité du terrain. Il en retourne que dans les faits, certaines modalités de variables ne sont pas toujours indépendantes. Par exemple, la variable « Canal de vente » a deux modalités : « F » pour la vente en face à face (en agence) et « D » pour la vente à distance. En pratique, il est possible qu'un contrat soit indiqué comme ayant été souscrit à distance bien que le client était en face d'un conseiller. Dans ce cas, la vente du contrat s'est en réalité faite en face à face mais le contrat apparaît comme ayant été souscrit à distance.

La même erreur est possible pour le type de conseiller. En effet, il est fréquent que des conseillers particuliers enregistrent la souscription d'un contrat à leur nom alors que le conseiller en gestion de patrimoine s'est en réalité occupé de la vente de ce contrat. Cela est dû au fait que ces deux types de conseillers ont des portefeuilles « miroir », c'est à dire qu'ils sont amenés à s'occuper des mêmes clients. Les résultats de notre étude peuvent donc être légèrement biaisés. Déterminons maintenant les capacités prédictives de nos modèles en les appliquant aux données de validation.

### Validation

L'étape de validation permet de vérifier les pouvoirs prédictifs des modèles sur des données indépendantes de la base d'apprentissage. Les données d'entraînements sont constituées des contrats Assurance Famille souscrits entre juin 2021 et mai 2022 à la Banque Populaire F.

Cela représente environ 15 000 contrats.

| Métrique | Arbre | Random Forest | CatBoost | Régression logistique |
|---|---|---|---|---|
| VN | 7 641 | 8 208 | 7 706 | 7 743 |
| VP | 283 | 197 | 210 | 287 |
| FN | 557 | 651 | 630 | 553 |
| FP | 1 218 | 643 | 1 153 | 1 116 |
| Accuracy | 81,7% | 86,7% | 81,6% | 82,8% |
| Précision | 18,9% | 23,2% | 15,4% | 20,5% |
| Rappel | 33,7% | 23,5% | 25% | 34,2% |
| F1-score | 24,2% | 23,3% | 19,1% | 25,6% |
| AUC-ROC | 69,6% | 71,4% | 61% | 72,3% |
| AUC-PR | 16,5% | 18,3% | 12,8% | 19,9% |

TABLE 3.16 – Validation - Tous les modèles

Nous constatons que les modèles performent moins bien que précédemment sur les données de test mais les métriques restent dans le même ordre de grandeur. Le modèle de régression logistique est celui qui prédit le mieux la résiliation précoce des contrats Assurance Famille sur la période juin 2021-mai 2022. Cependant, aucun modèle n'a une performance prédictive très élevée. Nous en déduisons que les modèles n'ont pas réellement réussi à identifier de fortes liaisons entre les variables explicatives et la résiliation des contrats. La faible prédiction des résiliations pourrait aussi s'expliquer par le fait que le comportement des clients a changé.

Mais ces modèles d'apprentissage procèdent d'une démarche itérative. Le premier apport a été de permettre, notamment à travers les arbres de classification, d'échanger avec le métier et d'entrer dans les subtilités du processus de vente des produits Famille et des produits de prévoyance en général. À ce stade, il s'agit d'une mise en place d'un modèle d'apprentissage le plus frugal possible, l'enrichissement avec d'autres variables sera envisagé par la suite.

Nous allons maintenant appliquer le modèle de régression logistique aux données des autres Banques Populaires.

### 3.2.3 Généralisation au réseau Banque Populaire

Nos modèles ont été entraînés à partir des données de la Banque Populaire F. Nous allons maintenant déterminer si le modèle optimal que nous avons trouvé précédemment est généralisable aux autres Banques Populaires en appliquant ce modèle à leurs données de validation (contrats souscrits entre juin 2021 et mai 2022). Cela nous permettra d'établir des profils de banques et de savoir pour quelles banques il serait possible de mettre le modèle en production.

Afin de nous assurer qu'il n'y ait pas de dérive de données (*data drift*), nous utilisons la librairie *evidently*. La dérive de données survient lorsque les nouvelles observations sur lesquelles prédit le modèle diffèrent de façon trop importante des données d'entraînement. Ce problème doit être détecté et anticipé car il dégrade les performances de prédiction au fur et à mesure que le temps passe. Le *data_drift_report* de la librairie *evidently* mesure la dérive des données variable par variable. Si la dérive des données est détectée sur moins de 50% des variables, les nouvelles données sont considérées non dérivées. Un exemple de rapport généré par *evidently* est disponible en annexe.

Aucune dérive de données n'est signalée d'après les différents rapports. Nous pouvons maintenant implémenter le modèle de régression logistique optimisé précédemment aux autres Banques Populaires. Les résultats ont été rassemblés dans le tableau ci-dessous.

| Banque | Accuracy | Précision | Rappel | F1-score | AUC-ROC | AUC-PR |
|---|---|---|---|---|---|---|
| BP A | 82% | 11,3% | 26,8% | 15,9% | 61,4% | 11,2% |
| BP B | 83,3% | 12,6% | 23,6% | 16,4% | 59,9% | 11,2% |
| BP C | 82,1% | 10,5% | 22,5% | 14,3% | 58,2% | 9% |
| BP D | 83,7% | 16,7% | 37,1% | 23% | 67,1% | 24,4% |
| BP E | 80,7% | 14,4% | 26% | 18,5% | 63% | 15% |
| BP F | 80,8% | 13,7% | 23% | 17,2% | 59,6% | 11,9% |
| BP G | 83,3% | 29,1% | 29% | 29% | 65% | 30,4% |
| BP H | 79,6% | 17,9% | 20,6% | 19,2% | 62,4% | 18,1% |
| BP I | 82,4% | 8,7% | 20,2% | 12,2% | 56,1% | 8,8% |
| BP J | 83,3% | 9,8% | 31,2% | 14,9% | 62,6% | 14% |
| BP K | 81,3% | 10,9% | 22,7% | 14,7% | 60,9% | 11,6% |

Table 3.17 – Indicateurs de performance - Toutes BP

D'après les différents indicateurs de performance, nous remarquons que les Banques Populaires G et D sont celles dont le modèle arrive le mieux à prédire la résiliation. La Banque Populaire G est également celle avec le taux de résiliation précoce le plus élevé. Ces résultats montrent bien l'hétérogénéité au sein du réseau Banque Populaire : afin de mieux prédire les résiliations précoces, il faudrait quasiment produire un modèle par banque.

## 3.3 Apports et limites de l'étude

Notre objectif était d'étudier et mieux comprendre la résiliation des contrats Famille, et notamment la résiliation précoce des contrats Assurance Famille. Pour cela, nous avons appliqué plusieurs modèles de *machine learning*. Le modèle optimal qui prédit le mieux la résiliation des contrats est la régression logistique. Les variables qui contribuent le plus à ce modèle sont l'agence, la catégorie socioprofessionnelle, le canal de vente et le top prévoyance. Ces variables sont pour la plupart liées aux conditions de souscription des contrats. La régression logistique, tout comme les arbres de classification, est très facile à interpréter, ce qui est un réel avantage lors d'échanges avec le métier (bien que certaines techniques d'encodage des variables catégorielles compromettent légèrement l'interprétabilité). Ces échanges ont permis de souligner de possibles inexactitudes dans les données (notamment pour le canal de vente et le type de conseiller).

Même si les métriques des différents modèles peuvent paraître faibles, celles relatives au gain s'avèrent intéressantes d'un point de vue marketing opérationnel. Nous pouvons toutefois avancer quelques limites. Tout d'abord, les données d'apprentissage correspondent aux contrats souscrits entre avril 2019 et mai 2021. Une grande partie de ces contrats ont donc été souscrits durant la pandémie de COVID-19, ce qui a pu conduire à un comportement de résiliation que nous ne retrouvons pas dans le futur. Ensuite, les données d'apprentissage ne sont pas très nombreuses par rapport à la totalité des données que nous avions à disposition, car l'offre Famille étant récente, le nombre de contrats Famille souscrits augmente beaucoup au fil du temps. Cela a pu conduire à un sur-apprentissage du modèle. De plus, le modèle de régression logistique renvoie des probabilités conditionnelles et classe les observations selon un seuil de 0,5. Ce seuil pourrait être modifié en fonction des besoins de ciblage.

Afin d'améliorer la performance du modèle, il existe plusieurs possibilités. Premièrement, nous pouvons continuer l'apprentissage du modèle, chose réalisable du fait d'une forte autonomie dans l'accès aux données. Ensuite, nous pourrions ajouter d'autres variables, comme par exemple l'appétence d'une Banque Populaire à la prévoyance individuelle (i.e. si la Banque Populaire a l'habitude de vendre ce type de produits) ou si le client a un ou

plusieurs contrats d'assurance emprunteur (ou d'autres contrats d'assurance BPCE). Cela demanderait un enrichissement avec des données externes au DWH PI.

Dans une optique de mise en production du modèle, nous pouvons choisir une autre banque pilote, par exemple la Banque Populaire G. Si des actions commerciales sont organisées afin de cibler les clients que le modèle aura prédit comme étant les plus susceptibles de résilier à court terme, nous pourrons observer si ces actions auront eu un effet positif ou non, et donc si le client a fini par résilier son contrat. Ces retours permettraient d'améliorer encore le modèle.

Cette étude aura permis de mieux comprendre la résiliation des contrats Famille. Elle aura également permis d'exploiter les données du DWH PI et d'introduire des analyses de résiliation par cohorte année-mois, ce qui n'avait pas été fait jusqu'à présent. Nous avons également pu dégager des facteurs d'explication de la résiliation précoce des contrats Assurance Famille. Nous allons désormais étudier la résiliation des contrats Famille à horizon durée de vie du contrat. Pour cela, nous utilisons une approche basée sur les modèles de survie.

# Chapitre 4

# Modèles de survie

## 4.1 Intérêt

Après avoir étudié la résiliation précoce, nous allons maintenant modéliser la résiliation des contrats Famille sur toute la durée de vie du contrat. Pour cela, nous établissons des lois d'expérience par âge, puis par ancienneté.

Les tables d'expérience sont des tables construites sur une population spécifique. Les compagnies d'assurance ont la possibilité de construire leurs propres tables, adaptées à leur portefeuille d'assurés. Pour cela, il est nécessaire de disposer d'un volume suffisant de données. Utiliser une table d'expérience permet de refléter le risque intrinsèque au portefeuille et ainsi de mieux calculer les provisions *Best Estimate* dans le cadre des référentiels prudentiels Solvabilité II et IFRS17. L'estimation des taux de résiliation des contrats Famille par âge et par ancienneté nous permettra également de raffiner les modèles actuels d'estimation de provisions et de projection de rentabilité.

Nous voulons construire deux lois de résiliation d'expérience : une loi par âge, ce que le modèle de projection des provisions *Best Estimate* utilise actuellement, et une loi par ancienneté du contrat. Ce choix est motivé par l'étude commencée dans le chapitre précédent. Nous décidons de construire la loi par ancienneté avec un pas mensuel afin de gagner en précision.

Les taux de résiliation sont estimés via une approche non-paramétrique. Dans ce chapitre, nous introduirons les données que nous utiliserons pour cette étude, puis les approches classiques d'estimation des taux de résiliation (méthodes d'estimation de Kaplan-Meier et de Hoem). Ensuite, une partie concernant le lissage de ces taux sera présentée, dans laquelle nous testerons la méthode de Whittaker-Henderson, ainsi que la méthode par noyaux discrets. Pour finir, nous allons valider les taux estimés sur un nouvel échantillon qui n'a pas participé à la construction des lois.

Tous les calculs dans ce chapitre ont été effectués avec le langage $R$, notamment la librairie *survival*.

## 4.2 Les données

La base de données utilisée est une base historisée des contrats de prévoyance individuelle extraite le 2 juin 2023. Les données sont à la maille client-contrat-garantie et représentent près de 7 millions de lignes. Nous avons filtré cette base sur les contrats Famille et effectué quelques retraitements, notamment le reformatage des données et la suppression des valeurs aberrantes.

La base historisée a été dédoublonnée afin de ne garder qu'une seule ligne par client et par contrat. La nouvelle base contient environ 1,5 million de lignes. Comme dans le chapitre précédent, nous considérons que la date de résiliation est la date de clôture du contrat.

Nous devons nous assurer qu'il n'y a pas de problème d'archivage des données. Il se pourrait en effet qu'une partie des données relatives à des contrats plus anciens soit purgée de la base historisée. Afin de vérifier cela, nous comparons notre base historisée dédoublonnée avec la même base historisée mais extraite en juin 2021.

| Année de clôture | Nombre de contrats 2021 | Nombre de contrats 2023 |
|---|---|---|
| 2016 | 1 877 | 1 877 |
| 2017 | 16 949 | 16 949 |
| 2018 | 36 000 | 35 999 |
| 2019 | 51 585 | 51 588 |
| 2020 | 51 824 | 51 838 |

Table 4.1 – Nombre de résiliations par année de clôture des contrats Famille

Nous remarquons qu'aucune purge n'a eu lieu et que toutes les résiliations sont bien présentes dans notre base (pas de troncature). Nous pouvons nous servir de cette base pour calculer nos lois de résiliation.

Les notions de base des modèles de survie sont présentées dans la section suivante.

## 4.3 Introduction à l'analyse de survie

L'analyse de survie, également appelée analyse de durée de vie, est une méthode statistique utilisée dans de nombreux domaines pour étudier le temps qu'il faut à un évènement particulier pour se produire. L'analyse de survie permet de comprendre les facteurs qui influencent le moment où cet évènement se produit, ainsi que de prédire les durées de vie des éléments étudiés. Dans ce mémoire, l'évènement d'intérêt est la résiliation d'un contrat et par la suite, la variable $X$ représente la durée de détention d'un contrat, à savoir la durée entre la date de création du contrat et sa date de résiliation.

L'analyse de survie est particulièrement utile lorsque l'on travaille avec des données incomplètes dues à des censures et/ou des troncatures[HUB94]. Elles peuvent être causées par de nombreux facteurs, indépendants ou non de l'évènement étudié. Ces deux phénomènes introduisent des biais dans les données observées, c'est pourquoi il est impératif d'en prendre compte.

Dans cette section, nous allons présenter les notions de censure et de troncature, ainsi que quelques méthodes d'estimation des taux bruts de résiliation.

### 4.3.1 Censures et troncatures

La durée de détention d'un contrat peut être modélisée par une variable aléatoire $X$. Cette variable est dite censurée par une variable aléatoire de censure $C$ lorsqu'il arrive d'observer $C$ au lieu de $X$. Il existe deux types de censure :

- Censure à gauche si $X < C$
- Censure à droite si $X > C$

La censure à gauche se produit notamment lorsque le contrat est résilié avant la date de début d'étude. Nous savons que la résiliation a eu lieu, sans précision supplémentaire. Ainsi, $X$ est inconnue et nous savons seulement que $X < C$ avec $C$ la durée de détention du contrat à la date de début de l'étude.

La censure à droite est quant à elle le cas où le contrat n'est pas résilié pendant la période d'observation, soit parce que le contrat sort du portefeuille avant la fin de cette période pour une autre raison, soit parce que le contrat est toujours en cours à la fin. La durée exacte de détention du contrat n'est donc pas disponible, nous savons seulement que $X > C$, avec $C$ la durée de détention du contrat au moment où celui-ci n'est plus observable. Une majorité de données en assurance vie sont censurées à droite, car la plupart des contrats restent en cours après la fin de l'étude. Ne pas considérer ce phénomène introduit un biais car utiliser la date de censure $C$ à la place de la date de résiliation reviendrait à sous-estimer $X$.

La troncature se produit si l'observation de la variable d'intérêt $X$ n'a lieu que conditionnellement à un événement $B$. On observe donc $X$ sachant $B$ au lieu de $X$. Lorsqu'il y a troncature, les valeurs de $X$ qui sont observées appartiennent donc à un sous-ensemble des valeurs réellement prises par $X$.

La troncature à gauche est le cas où $X$ n'est observée que si elle est supérieure à un seuil $s$. Par exemple, dans le cas d'un produit d'assurance avec une franchise en jours, les sinistres déclarés seront d'une durée supérieure à cette franchise. La troncature à droite se produit quant à elle dans le cas où $X$ n'est observable que si elle est inférieure à un seuil $s$. Par conséquent, il y a troncature à droite quand les seuls individus observables sont ceux ayant expérimenté l'événement d'intérêt avant une date $s$.

### 4.3.2 Estimateur binomial et estimateur de Hoem

Soient les notations suivantes :

- $x$ l'âge de l'assuré
- $n_x$ le nombre de contrats actifs avec des assurés de l'âge $x$ à l'âge $x+1$
- $D_x$ la variable aléatoire représentant le nombre de résiliations observées à l'âge $x$
- $d_x$ la réalisation de $D_x$
- $q_x$ : la probabilité de résilier dans l'année pour un assuré d'âge $x$

Nous faisons l'hypothèse que chaque résiliation est indépendante des autres et que $D_x$ suit une loi binomiale $\mathcal{B}(n_x, q_x)$. Ainsi, nous avons

$$P(D_x = d_x) = \binom{n_x}{d_x} \times q_x^{d_x} \times (1-q_x)^{n_x - d_x}$$

L'estimateur binomial de $q_x$ noté $\hat{q}_x^{Bin}$ est déterminé par la méthode du maximum de vraisemblance.

$$\mathcal{L}(q_x) = K \times q_x^{d_x} \times (1-q_x)^{n_x - d_x}$$

avec $K$ une constante indépendante de $q_x$. Nous en déduisons la log-vraisemblance :

$$l(q_x) = \ln(K) + d_x \times \ln(q_x) + (n_x - d_x) \times \ln(1 - q_x)$$

d'où :

$$\frac{\partial l}{\partial q_x}(\hat{q}_x^{Bin}) = \frac{d_x - \hat{q}_x^{Bin} \times n_x}{\hat{q}_x^{Bin} \times (1 - \hat{q}_x^{Bin})} = 0$$

L'estimateur binomial est donc :

$$\hat{q}_x^{Bin} = \frac{d_x}{n_x}$$

Notons que cet estimateur est également celui obtenu par la méthode des moments. En effet, comme le nombre de résiliations suit une loi $\mathcal{B}(n_x, q_x)$, alors $\mathbb{E}[D_x] = n_x \times q_x$ , d'où $q_x = \frac{\mathbb{E}[D_x]}{n_x} \approx \frac{d_x}{n_x}$.

L'estimateur binomial a l'avantage d'être facile à calculer. Toutefois, il n'est utilisable que lorsque tous les contrats sont totalement observables, c'est-à-dire que la résiliation doit être l'unique cause de sortie possible pour tout assuré d'âge $x$[PLA11]. Ce modèle est donc difficilement utilisable en pratique.

L'estimateur de Hoem[HOE76] généralise l'estimateur binomial en prenant en compte les censures et troncatures. En effet, il repose sur la notion d'exposition au risque : un contrat $i$ dont l'assuré est d'âge $x$ est exposé au risque de résiliation entre les dates $\alpha_i$ et $\beta_i$ avec $[\alpha_i, \beta_i] \subset [x, x+1]$. L'exposition au risque de ce contrat est la fraction d'années pendant lequel son assuré a l'âge $x$. Notons :

- $n_x$ le nombre de contrats actifs dont les assurés sont d'âge $x$
- $[\alpha_i, \beta_i]$ l'intervalle inclus dans $[x, x+1]$ dans lequel l'assuré $i$ est exposé au risque de résiliation
- $X_1, ..., X_{n_x}$ des variables de Bernoulli de paramètre ${}_{\beta_i - \alpha_i}q_{x+\alpha_i}$ valant 1 si l'assuré résilie dans l'année et 0 sinon
- $D_x = \sum_{i=1}^{n_x} X_i$ la variable aléatoire représentant le nombre de résiliations observées en $[x, x+1]$
- $d_x$ la réalisation de $D_x$

Comme pour l'estimateur binomial, on suppose que chaque résiliation est indépendante des autres (indépendance des $X_i$). Nous faisons une autre hypothèse fondamentale (hypothèse de Balducci) :

$${}_{\beta_i - \alpha_i}q_{x+\alpha_i} = (\beta_i - \alpha_i)q_x$$

Nous avons alors :

$$\mathbb{E}[D_x] = \sum_{i=1}^{n_x} \mathbb{E}[X_i] = \sum_{i=1}^{n_x} {}_{\beta_i-\alpha_i}q_{x+\alpha_i} = \sum_{i=1}^{n_x} (\beta_i - \alpha_i) q_x = q_x \times \sum_{i=1}^{n_x} (\beta_i - \alpha_i)$$

En notant $E_x = \sum_{i=1}^{n_x} (\beta_i - \alpha_i)$ l'exposition au risque à l'âge $x$, nous pouvons écrire :

$$q_x = \frac{\mathbb{E}[D_x]}{E_x}$$

Nous en déduisons alors l'estimateur de Hoem[NDI16] :

$$\hat{q}_x^{Hoem} = \frac{d_x}{E_x}$$

ce qui correspond au ratio entre le nombre de résiliations à l'âge $x$ et l'exposition au risque de résiliation à l'âge $x$.

L'estimateur de Hoem est un estimateur paramétrique, car il repose sur une hypothèse de distribution infra-annuelle des résiliations : l'hypothèse de Balducci. C'est un estimateur sans biais[BAL13], c'est-à-dire que $\mathbb{E}[\hat{q}_x^{Hoem}] = q_x$. C'est aussi un estimateur convergent : $\forall\, \epsilon > 0, \lim_{x \to +\infty} P(|\hat{q}_x^{Hoem} - q_x| > \epsilon) = 0$. La variance de cet estimateur est $\text{Var}(\hat{q}_x^{Hoem}) = \frac{q_x(1-q_x)}{E_x}$. Quand l'exposition $E_x$ et le nombre de résiliations $D_x$ sont assez grands, il est possible d'approximer $D_x$ par une loi normale. En utilisant le théorème de la limite centrale, on obtient un intervalle de confiance asymptotique de niveau $1 - \alpha$ pour l'estimateur de Hoem, en remplaçant $q_x$ par l'estimateur de $q_x$ :

$$IC_{1-\alpha}(q_x) = \left[ \hat{q}_x^{Hoem} \pm \phi_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{q}_x^{Hoem}(1 - \hat{q}_x^{Hoem})}{E_x}} \right]$$

où $\phi_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite $\mathcal{N}(0, 1)$. Ce quantile vaut environ 1,96 pour un intervalle de confiance à 95% ($\alpha = 0,05$). Nous pouvons remarquer que la variance de l'estimateur et donc la largeur de l'intervalle de confiance sont inversement proportionnelles à l'exposition au risque.

### 4.3.3 Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier[KAP58] est un estimateur non paramétrique qui permet d'estimer la fonction de survie $S$ définie par $S(t) = P(X > t)$ avec $t$ la variable temps. Soient les notations suivantes :

- $t_i$ les dates triées par ordre croissant pour lesquelles il y a au moins une résiliation

- $n_i$ le nombre de contrats en cours juste avant la date $t_i$
- $d_i$ le nombre de résiliations à la date $t_i$

L'estimateur de Kaplan-Meier de la fonction de survie $S$ est :

$$\hat{S}^{KM}(t) = \prod_{i\ tq\ t_i \leq t} (1 - P(X = t_i | X \geq t_i)) = \prod_{i\ tq\ t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

On en déduit les taux de résiliation :

$$\hat{q}_x^{KM} = 1 - \frac{\hat{S}^{KM}(x+1)}{\hat{S}^{KM}(x)}$$

La fonction de survie obtenue par Kaplan-Meier est constante par morceaux[PLA11][SAI21]. C'est une courbe en escalier où un saut n'apparaît qu'en présence d'une observation de résiliation. Cela signifie qu'il y a autant de sauts de marche dans la fonction de survie $\hat{S}^{KM}$ que de dates de résiliation uniques. Un point important est que cet estimateur prend en compte les censures. Sans phénomène de censure, l'estimateur de Kaplan-Meier est équivalent à la fonction de survie empirique $\hat{S}_n(t) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{t_k > t}$.

L'estimateur de Kaplan-Meier est convergent et biaisé positivement. Cela signifie que la résiliation peut être surestimée. L'estimateur de Greenwood permet d'estimer la variance de l'estimateur de Kaplan-Meier :

$$\widehat{\text{Var}}(\hat{S}^{KM}(t)) = \hat{S}^{KM}(t)^2 \times \sum_{t_i < t} \frac{d_i}{n_i(n_i - d_i)}$$

Il est alors possible d'obtenir un intervalle de confiance asymptotique pour l'estimateur de Kaplan-Meier :

$$IC_{1-\alpha}(\hat{S}(t)) = \left[\hat{S}^{KM}(t) \pm \phi_{1-\frac{\alpha}{2}} \times \widehat{\text{Var}}(\hat{S}^{KM}(t))\right]$$

où $\phi_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite $\mathcal{N}(0, 1)$.

## 4.4 Calcul des taux bruts

### 4.4.1 Définition de la période d'observation et de la censure

Afin d'estimer au mieux la résiliation, il est important de choisir judicieusement la période d'observation. Pour ce faire, il convient de s'assurer que la quasi-totalité des résiliations survenues au cours de la période d'observation sont bien visibles dans la base d'exposition. Pour rappel, dans le cas d'une résiliation pour non-paiement de la prime, le contrat est clôturé

40 jours après envoi d'une lettre recommandée de mise en demeure. Pour une résiliation à la demande du client, le contrat cesse à l'échéance de prime suivant la demande de résiliation.

En choisissant une date de fin d'observation trop récente, le risque serait de ne pas capter toutes les résiliations et donc sous-estimer les taux de résiliation. Il ne faut pas non plus choisir une date de début d'observation trop ancienne, car le portefeuille et les comportements de résiliation peuvent drastiquement changer au cours du temps. Observons les taux bruts de résiliation par année.

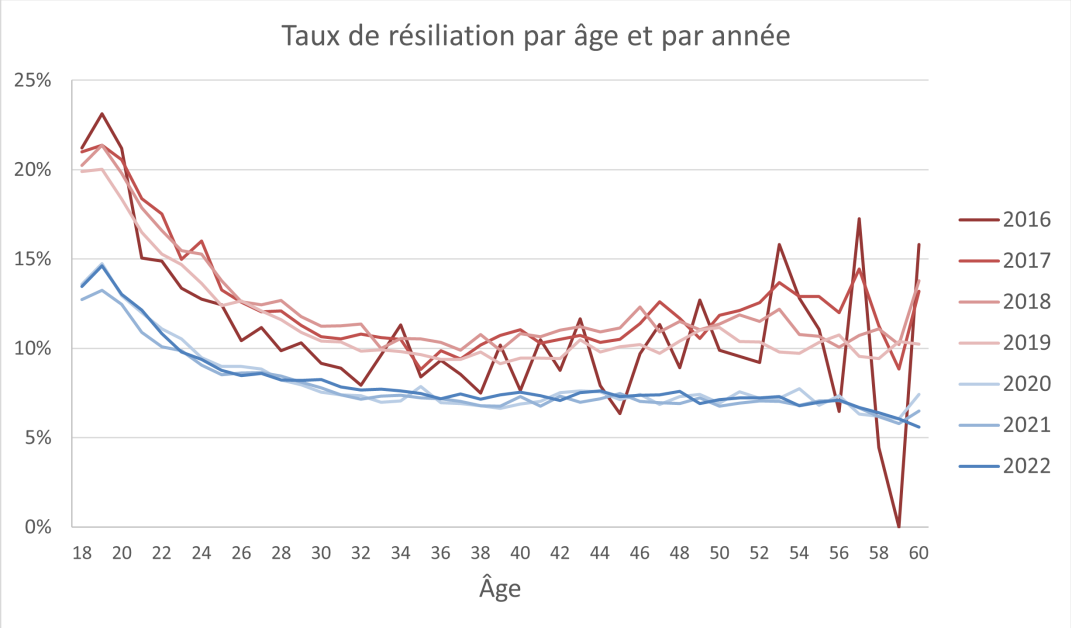


FIGURE 4.1 – Taux bruts de résiliation par âge et par année

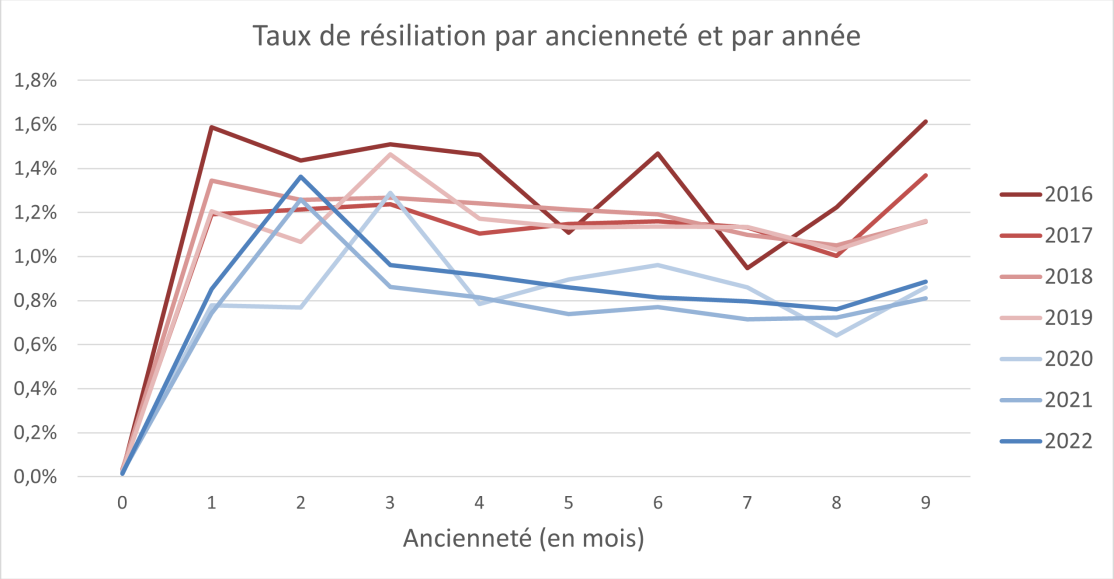


FIGURE 4.2 – Taux bruts de résiliation par ancienneté (en mois) et par année

Nous pouvons observer une nette différence avec les taux de résiliation avant et après 2020. Notre période d'observation sera donc de 2020 à 2022 pour la loi d'expérience par âge. Nous ne choisissons pas la même période d'observation pour la loi par ancienneté, car dans ce cas l'ancienneté de contrat la plus grande serait de 3 ans, avec une exposition relativement faible. Les taux de résiliation pour les dernières anciennetés ne seraient pas forcément pertinents. Pour l'usage que nous voulons faire de notre loi par ancienneté, à savoir les projections à moyen terme sur 3 ans, nous jugeons qu'il est préférable de choisir une grande période d'observation, quitte à surestimer les taux de résiliation. Il faut cependant noter que plus les années passent et plus le nombre de contrats augmentent, les contrats souscrits après 2020 auront donc plus de poids que les contrats souscris avant 2020. La loi par ancienneté sera donc construite avec les cohortes de 2016 à 2022.

La prise en compte de la censure est indispensable pour différencier les résiliations appartenant à la fenêtre d'observation et pour avoir un calcul plus précis des durées du portefeuille. Pour définir cette variable, nous considérons la date de clôture du contrat comme date de résiliation potentielle. Les contrats étant marqués comme résiliés dans la base de données et dont la date de clôture appartient à la période d'observation sont non censurés. Les contrats dont l'état est autre que la résiliation ou dont la date de résiliation n'appartient pas à la période d'observation sont censurés.

### 4.4.2 Exposition et nombre de résiliations

Il est important de calculer les expositions par âge et par ancienneté ainsi que le nombre de contrats résiliés afin d'avoir une meilleure connaissance du portefeuille et des profils de risques. Les figures suivantes montrent les différentes expositions des contrats Famille par âge et par ancienneté pendant les périodes d'observations définies précédemment, ainsi que le nombre de résiliations associés.

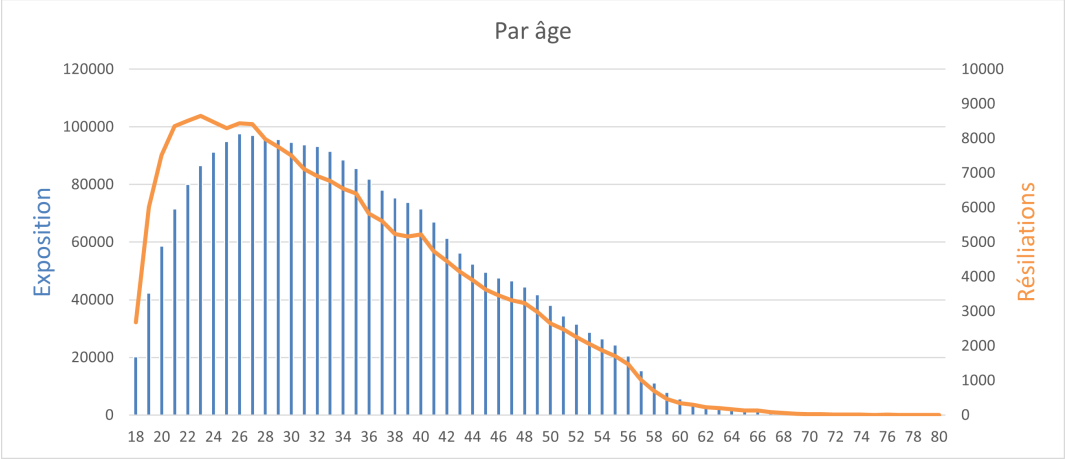


FIGURE 4.3 – Exposition et nombre de résiliations par âge

La figure 3.3 montre bien la forte résiliation dans les âges les plus jeunes du portefeuille.

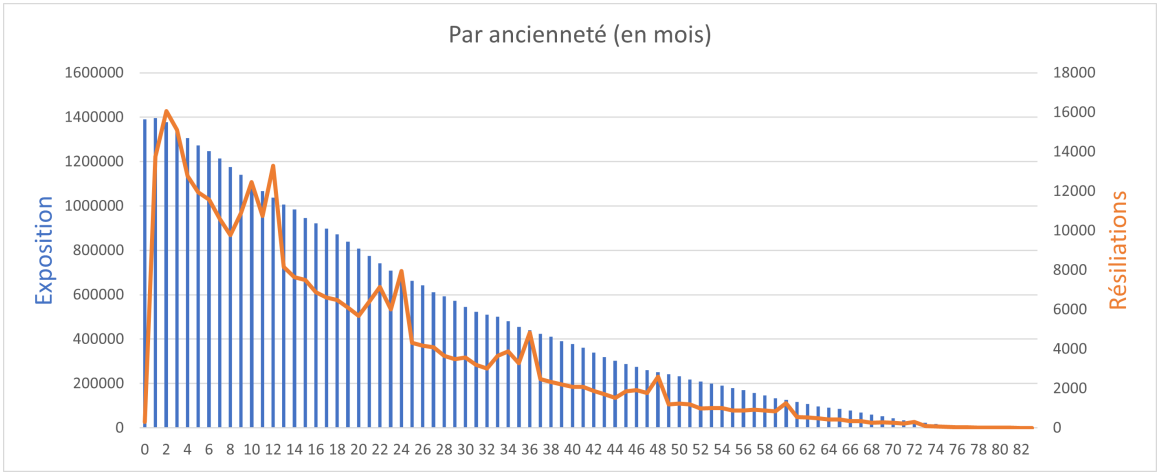


FIGURE 4.4 – Exposition et nombre de résiliations par ancienneté

Nous constatons qu'un très faible nombre de contrats sont résiliés pendant le premier mois d'ancienneté du contrat. Nous pouvons également voir très clairement que des pics

de résiliations ont lieu à 12, 24, 36, 48 et 60 mois d'ancienneté. Ce phénomène est sûrement dû au fait que les contrats Famille sont des contrats d'assurance temporaire décès d'une durée d'un an, avec une révision tarifaire annuelle. Notons toutefois que l'intensité des pics diminue avec l'ancienneté du contrat.

Calculons maintenant les taux bruts de résiliation grâce aux méthodes de Kaplan-Meier et de Hoem présentées précédemment.

### 4.4.3 Taux bruts par âge et par ancienneté

Les figures ci-dessous montrent que les taux estimés par la méthode de Kaplan-Meier sont très proches de ceux estimés par la méthode de Hoem et appartiennent bien aux intervalles de confiance de Hoem. Toutefois, les intervalles de confiance pour les derniers âges sont très larges, à cause de la faible exposition.

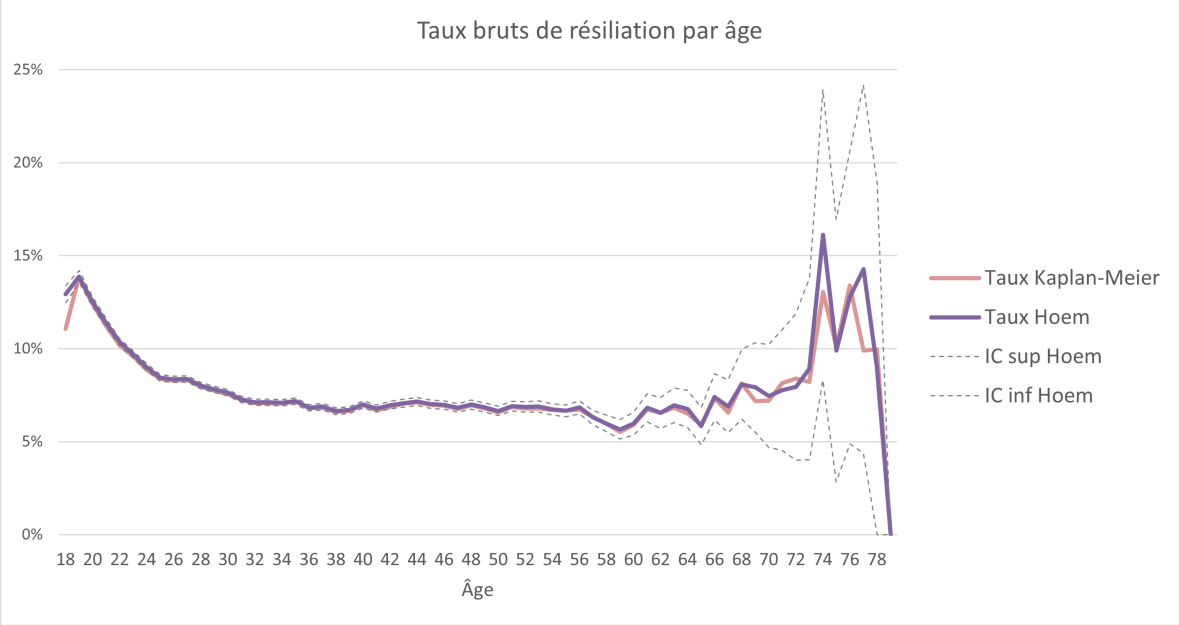


FIGURE 4.5 – Taux bruts de résiliation par âge

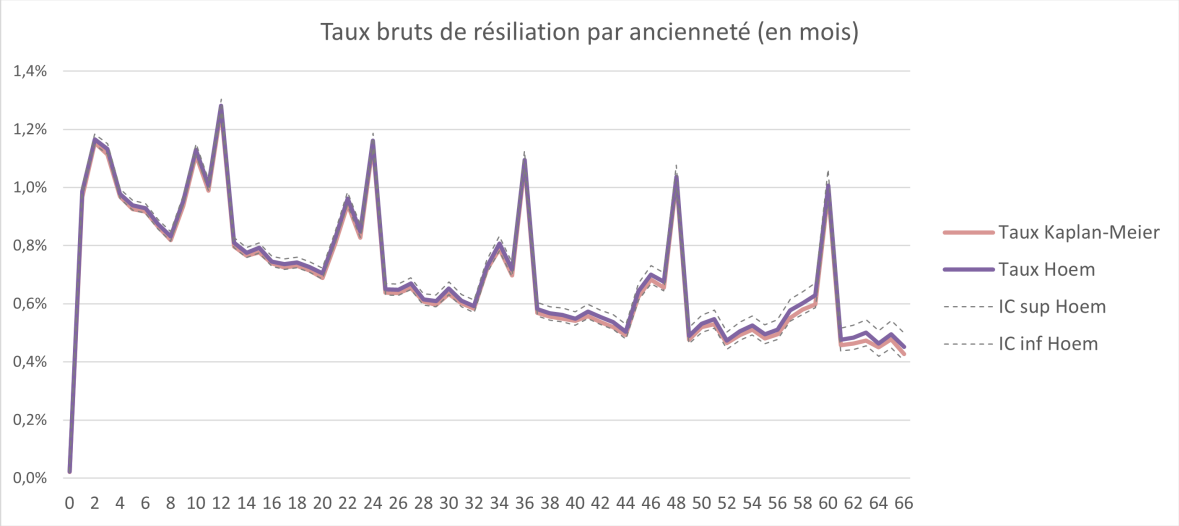


FIGURE 4.6 – Taux bruts de résiliation par ancienneté

Nous retrouvons les pics de résiliations pour les taux bruts par ancienneté. Le taux moyen de résiliation globale est de 8% pour les deux lois. Par la suite, les taux bruts désigneront les taux estimés par la méthode de Hoem, car bien qu'étant très proches des taux de Kaplan-Meier, leurs valeurs sont légèrement supérieures et ils sont donc un peu plus prudents.

Les taux bruts de résiliation ne sont pas réguliers. Une procédure d'ajustement des taux est donc nécessaire. Nous allons présenter deux méthodes de lissage non-paramétrique : la méthode de Whittaker Henderson et la méthode des noyaux discrets. Nous comparerons ces différentes méthodes de lissage et nous retiendrons celle qui ajuste au mieux la courbure des taux bruts selon des critères que nous définirons.

Les méthodes de lissage ne seront mises en place que pour la loi par âge. En effet, nous estimons qu'il n'est pas pertinent de lisser la loi de résiliation par ancienneté car nous voulons garder le phénomène de périodicité des taux bruts.

## 4.5 Lissage des taux bruts

### 4.5.1 Méthode de Whittaker-Henderson

La méthode de Whittaker-Henderson[PLA22] est une méthode de lissage non-paramétrique dont le principe consiste à rechercher les valeurs qui minimisent une combinaison linéaire des critères de fidélité et de régularité.

Le critère de fidélité se base sur le calcul de la distance entre les taux lissés et les taux bruts. L'objectif est que cette distance soit petite pour que les taux lissés ne soient pas trop éloignés des taux bruts. Soit $F$ tel que

$$F = \sum_{x=x_{min}}^{x_{max}} (q_x^{lissés} - q_x^{bruts})^2$$

Plus $F$ est proche de 0 et plus les taux lissés sont proches des taux bruts.

Le critère de régularité calcule quant à lui la somme des accroissements des taux par âge $x$ afin de s'assurer qu'il n'y a pas une amplitude trop élevée entre les taux pour des âges consécutifs. Cette mesure reflète à quel point les taux sont lissés. Soit $R$ tel que

$$R = \sum_{x=x_{min}}^{x_{max-1}} (q_x^{lissés} - q_{x+1}^{lissés})^2$$

Plus $R$ est proche de 0 et plus les taux lissés sont réguliers.

La méthode de Whittaker-Henderson consiste donc à déterminer les taux $q_x^{WH}$ qui minimisent une fonction de coût $M = F + h \times R$. Le paramètre $h$ permet de contrôler l'importance accordée à la fidélité et à la régularité dans la fonction de coût. En particulier, si $h = 0$, aucun lissage n'est effectué car seul le critère de fidélité est pris en compte, ce qui donne $q_x^{WH} = q_x^{bruts}$. L'expression de la fonction de coût est

$$M = F + h \times R = \sum_{x=x_{min}^{x_{max}}} w_x (q_x^{WH} - q_{x+1}^{bruts})^2 + h \times \sum_{x=x_{min}}^{x_{max}-z} \Delta^z (q_x^{WH})^2$$

$w_x$ est le poids attribué à l'âge $x$. Cela permet de ne pas donner la même importance à la fidélité selon l'âge. Dans ce contexte, les poids sont définis de la manière suivante :

$$w_x = \frac{E_x}{max_x(E_x) - min_x(E_x)}$$

$z$ représente le nombre de fois où est appliqué l'opérateur $\Delta$ des « différences avant ». Cet opérateur est la généralisation du critère de régularité défini précédemment.

$$\text{Pour } z = 1,\ \Delta(q_x^{WH}) = q_x^{WH} - q_{x+1}^{WH}$$

$$\text{Pour } z = 2,\ \Delta^2(q_x^{WH}) = \Delta \circ \Delta(q_x^{WH}) = (q_x^{WH} - q_{x+1}^{WH}) - (q_{x+1}^{WH} - q_{x+2}^{WH})$$

$$\text{Pour } z = n,\ \Delta^n(q_x^{WH}) = \sum_{j=0}^{n} \binom{n}{j} (-1)^{n-j} q_{x+n-j}^{WH})$$

### 4.5.2 Méthode des noyaux discrets

Cette méthode est une autre approche de lissage non-paramétrique et a été implémentée dans la librairie *DBKGrad*[MAZ14]. Le lissage s'effectue de façon discrète, ce qui est pertinent pour le lissage d'une loi de résiliation par âge entier. Il utilise un estimateur à noyau discret. Les fonctions du noyau sont choisies parmi une famille de densités bêta discrétisées et re-paramétrées.

Soit $q_x$ le taux de résiliation théorique à l'âge $x \in X = 18, 19, ..., w$, où $w$ correspond à l'âge maximal possible de résiliation, et $\hat{q}_x$ les taux bruts de résiliation. La forme générale de l'estimateur à noyau discret s'écrit de façon suivante :

$$\hat{q}_x = \sum_{j=18}^{w} q_j k_h(j; m = x),\ x \in X$$

où $k_h(.; m = x)$ est la fonction de noyaux discrets, $m$ le mode unique du noyau et $h > 0$ un paramètre de lissage . Comme l'âge est considéré comme une variable discrète, le lissage par noyaux discrets peut être vu comme un lissage pondéré des moyennes mobiles locales des noyaux. Une fenêtre de rayon $h$ est utilisée pour calculer pour chaque $x$ une moyenne pondérée de tous les points appartenant à la fenêtre.

### 4.5.3 Critères de validation

Nous avons déjà évoqué les critères de fidélité et de régularité pour juger de la qualité du lissage des taux bruts. Nous allons présenter deux autres critères que nous allons utiliser par la suite.

Le ratio observés sur attendus est une méthode simple et pragmatique qui permet de vérifier la fidélité de la table construite à la résiliation observée. Il est défini ainsi :

$$O/A = \frac{D_x}{\hat{q}_x E_x}$$

Les résiliations attendues correspondent à l'exposition multipliée par le taux brut de résiliation. Un ratio observés attendus proche de 1 montre une bonne capacité de la table à prédire la résiliation d'expérience. Un ratio supérieur à 1 indique une sous-estimation de la résiliation et à l'inverse, un ratio inférieur à 1 montre une surestimation.

Nous allons également calculer l'erreur relative pondérée. C'est la somme des erreurs relatives entre les résiliations attendues et les résiliations observées à l'âge $x$, pondérées par

l'exposition à cet âge :

$$Erreur\ pondérée\ relative = \sum_{x=x_{min}}^{x_{max}} \frac{D_x - \hat{q}_x E_x}{D_x} \times \frac{E_x}{\sum_{x=x_{min}}^{x_{max}} E_x}$$

### 4.5.4 Validation des méthodes de lissage

Sur la figure ci-dessous, nous observons le lissage des taux bruts avec la méthode des noyaux discrets entre les âges 18 et 77 ans. Les intervalles de confiance sont d'autant plus larges que l'exposition est petite. Nous remarquons que ce lissage sous-estime la résiliation aux plus petits âges.

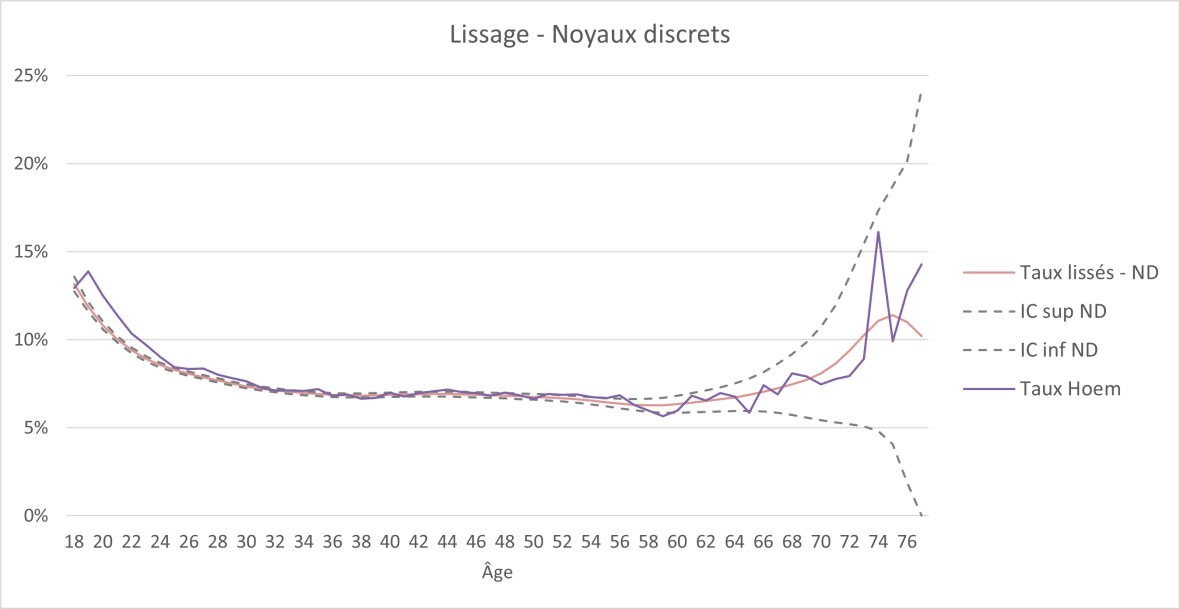


FIGURE 4.7 – Lissage par la méthode des noyaux discrets

Comparons visuellement les taux lissés avec la méthode de Whittaker-Henderson avec les taux bruts et les taux issus de la méthode des noyaux discrets.


FIGURE 4.8 – Comparaison graphique des différentes méthodes de lissage

Le lissage de Whittaker-Henderson sous-estime moins la résiliation des premiers âges par rapport au lissage par noyaux discrets, mais sous-estime relativement plus les taux bruts pour les derniers âges. Cela peut être moins problématique car le nombre de résiliations est beaucoup plus élevé pour les âges faibles. Il nous faut comparer les différentes méthodes de lissage plus finement. Le tableau ci-dessous contient les critères de validation présentés dans la section précédente.

| Méthode de lissage | Fidélité | Régularité | O/A | Erreur relative pondérée |
|---|---|---|---|---|
| Aucune (taux bruts) | 0 | 0,1267 | 0,9999 | 0 % |
| Noyaux discrets | 0,0691 | 0,0019 | 1,0374 | -0,39% |
| Whittaker-Henderson | 0,0609 | 0,0006 | 1,0103 | 0,58 % |

TABLE 4.2 – Critères de comparaison des lissages

Les taux lissés avec la méthode de Whittaker-Henderson sont ceux avec la meilleure fidélité et la meilleure régularité. Whittaker-Henderson est également le seul lissage qui fournit des taux de résiliation lissés sur tous les âges de 18 à 80 ans. C'est ces taux que nous retiendront pour la suite.

## 4.6 Validation des lois de résiliation

Après avoir construit nos lois de résiliation par âge et par ancienneté, nous devons tester leurs capacités de prédiction sur des données qui n'ont pas participé à la construction de ces lois. Pour rappel, la période d'observation de la loi par âge est comprise entre le 1er janvier 2020 et le 31 décembre 2022. La loi par ancienneté a quand à elle été construite sur tous les contrats souscrits avant le 1er janvier 2023.

Nous comparons le nombre de résiliations théoriques avec le nombre de résiliations observées entre le 1er janvier et le 31 mai 2023. Les taux de résiliation par âge étant des taux annuels, nous les mensualisons d'après la formule suivante :

$$q_x^{mensuel} = 1 - (1 - q_x^{annuel})^{\frac{1}{12}}$$

Le nombre de résiliations théoriques est obtenu en multipliant pour chaque âge et ancienneté l'exposition et le taux de résiliation associés. Les expositions ont aussi été calculées par mois.

Nous rappelons qu'en cas de défaut de paiement de prime, BPCE Vie a dix jours pour envoyer à l'adhérent une lettre recommandée de mise en demeure. L'adhérent a ensuite un délai de 40 jours à compter de l'envoi de cette lettre pour régulariser sa situation. Cela fait que nous nous attendons à surestimer légèrement le nombre de résiliations entre janvier et mai 2023.

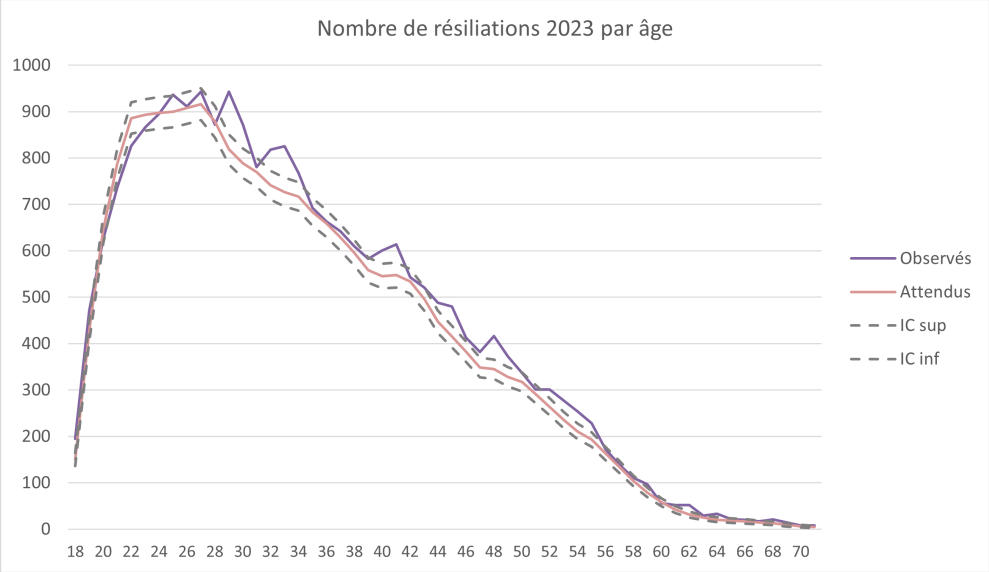


FIGURE 4.9 – Nombre de résiliations théoriques et observées - Par âge

Nous remarquons déjà que la loi par âge a tendance à sous-estimer le nombre de résiliations à partir de 28 ans. Le nombre de résiliations à partir de cet âge n'appartiennent pas à l'intervalle de confiance théorique à 95%. Nous pouvons émettre l'hypothèse que les résiliations ont augmenté en début 2023 par rapport aux années antérieures. Nous devrons suivre ce risque attentivement. Comparons maintenant avec la loi par ancienneté.

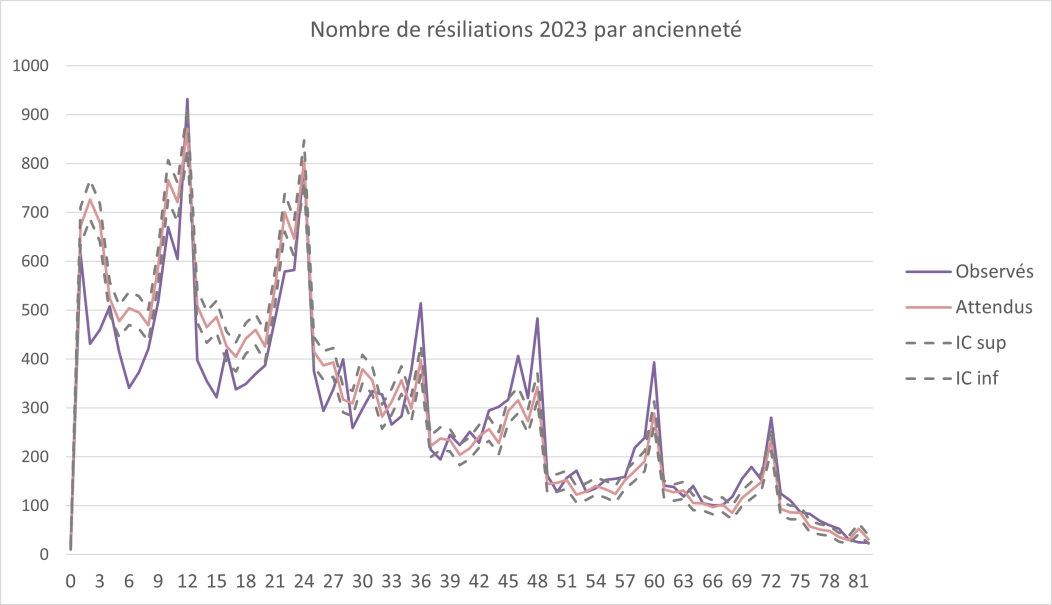


FIGURE 4.10 – Nombre de résiliations théoriques et observées - Par ancienneté

La loi par ancienneté semble mieux prédire le nombre de résiliations. Calculons les ratios observés sur attendus de nos deux lois.

| **Loi par âge** | **Loi par ancienneté** |
|---|---|
| 105,5% | 94,99 % |

TABLE 4.3 – Ratios observés sur attendus - Résiliations 2023

La loi par ancienneté s'avère bien adaptée. Cela confirme l'intérêt d'une loi de résiliation par ancienneté dans le cadre de projections à moyen terme. Nous allons désormais pouvoir mesurer l'impact de nos lois de résiliations sur le *Best Estimate* et sur les PMT.

# Chapitre 5

# Application

Dans ce chapitre, nous allons quantifier les impacts de nos nouvelles lois de résiliation construites à l'aide des données des contrats Famille, dans un premier temps sur le *Best Estimate* Solvabilité II, puis sur le calcul du Produit Net Bancaire projeté à court terme.

## 5.1 Impact Solvabilité II

### 5.1.1 Généralités

Entrée en application le 1er janvier 2016, Solvabilité II[ACP19] est une norme prudentielle fixant les règles de solvabilité applicables aux entreprises d'assurances dans l'Union Européenne. Ces règles sont réparties en trois piliers.

Le pilier I de Solvabilité II regroupe les exigences quantitatives, c'est-à-dire les règles de valorisation des actifs et des passifs, ainsi que les exigences de capital et leur mode de calcul. Les exigences de capital peuvent être calculées au moyen de la formule standard, ou au moyen d'un modèle interne complet ou partiel (plus rare).

Le pilier II regroupe d'une part les exigences qualitatives, notamment les règles de gouvernance et de gestion des risques, et d'autre part l'évaluation propre des risques de la solvabilité (*Own Risk and Solvency Assessment* ou ORSA).

Le pilier 3 de Solvabilité II concerne la communication d'informations au public et aux autorités de contrôle. Il vise à harmoniser au niveau européen les informations publiées par

les organismes d'assurance ainsi que celles remises aux superviseurs. Ces informations, à la fois quantitatives et qualitatives, sont à remettre à une fréquence annuelle et, pour certaines, trimestrielles.

Solvabilité II définit la valeur des provisions techniques comme étant égale à la somme du *Best Estimate* (« meilleure estimation ») et de la marge de risque. Selon la directive, « la meilleure estimation correspond à la moyenne pondérée par leur probabilité des flux de trésorerie futurs, compte tenu de la valeur temporelle de l'argent (valeur actuelle attendue des flux de trésorerie futurs), estimée sur la base de la courbe des taux sans risque. Le calcul de la meilleure estimation est fondé sur des informations actualisées et crédibles et des hypothèses réalistes et il fait appel à des méthodes actuarielles et statistiques adéquates, applicables et pertinentes ».

La marge de risque « est calculée de manière à garantir que la valeur des provisions techniques est équivalente au montant que les entreprises d'assurance et de réassurance demanderaient pour reprendre et honorer les engagements d'assurance et de réassurance ». A noter que « les entreprises d'assurance et de réassurance procèdent à une évaluation séparée de la meilleure estimation et de la marge de risque ». Le *Best Estimate* est donc un outil indispensable pour les professionnels de l'assurance et les régulateurs européens.

Conformément à la directive européenne, le *Best Estimate* doit être calculé brut de réassurance, ce qui signifie qu'il ne doit pas prendre en compte les effets de la réassurance lors de son calcul initial. Toutefois, il est important de préciser qu'un actif de réassurance est reconnu à l'actif de l'entreprise. Cette reconnaissance tient compte des probabilités de défaut du réassureur.

### 5.1.2 Impact sur le *Best Estimate*

Une formule générale du BE est la suivante :

$$\begin{aligned} BE &= VAP(Flux\ entrants - Flux\ sortants) \\ &= VAP(Primes\ TTC - Prestations - Commissions - Taxes \\ &\quad - Frais\ généraux\ sur\ les\ primes - Frais\ généraux\ sur\ les\ prestations\ futures) \end{aligned}$$

avec $VAP$ la valeur actuelle probable, c'est-à-dire la somme des flux actualisés pondérés par leur probabilité d'occurrence.

Le *Best Estimate* est calculé à l'aide d'un outil interne. Notons qu'à l'heure actuelle, une loi de résiliation globale construite en 2020 est utilisée pour tous les produits et tous les risques. Le modèle de projection des flux futurs ne prenant en entrée que des lois par âge, nous calculons le *Best Estimate* des produits Famille avec la loi de résiliation globale par âge puis avec la loi de résiliation construite précédemment par âge. Nous pouvons comparer graphiquement ces deux lois :

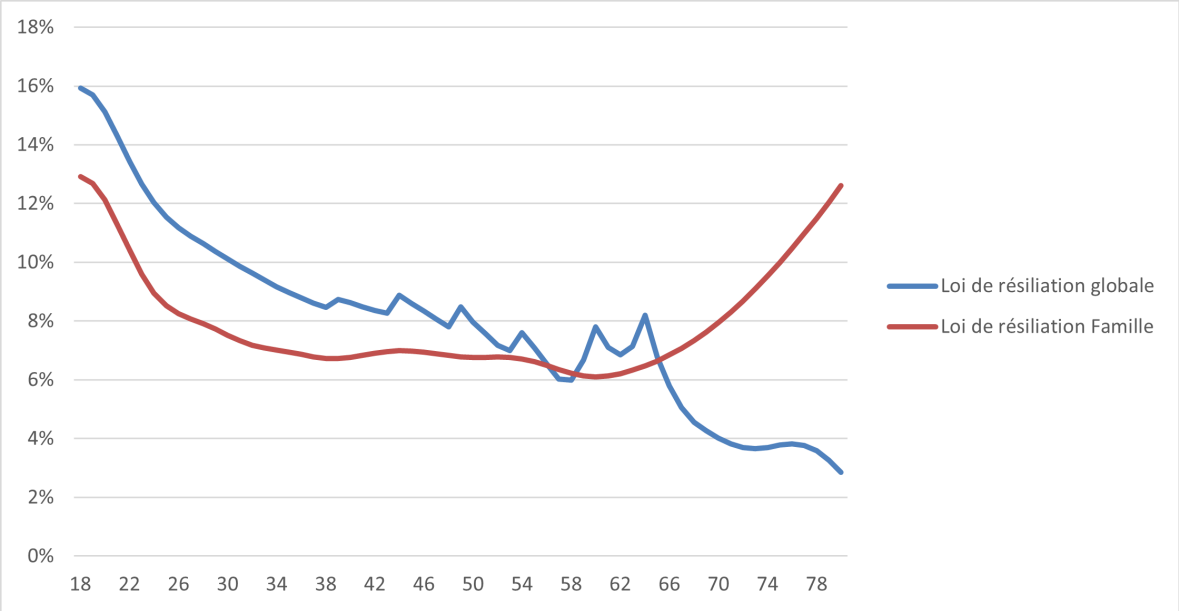


FIGURE 5.1 – Comparaison loi globale - loi Famille

Les taux de résiliation pour les produits Famille sont globalement inférieurs à ceux de tous les produits du portefeuille Prévoyance Individuelle confondus, excepté pour les grands âges. La loi globale n'est pas lisse, comme en témoigne la présence de pics de résiliation tous les cinq ans à partir de 40 ans, ce qui correspond aux âges de révision tarifaire d'autres produits du portefeuille.

Les calculs des *Best Estimates* sont présentés ci-dessous. Les chiffres ont été modifiés pour cause de confidentialité.

| **BE - Loi globale** | **BE - Loi Famille** | **BE - Écart relatif** |
|---|---|---|
| 19 799 117 | 19 719 920 | 0,4 % |

TABLE 5.1 – Sensibilité *Best Estimate*

L'écart entre les deux *Best Estimate* calculés est très faible (0,4 %). L'impact est donc négligeable, ce qui signifie que cela ne présente pas d'intérêt d'utiliser une loi de résiliation spécifique pour les produits Famille. Ce résultat peut s'expliquer car la projection n'est faite que sur un horizon d'un an. En effet, le calcul du *Best Estimate* tient compte de tous les flux futurs jusqu'au moment où l'assureur peut soit résilier le contrat, soit refuser une prime, soit modifier de façon illimitée les tarifs ou prestations prévus au contrat.

Nous avons mesuré l'impact de la nouvelle loi relativement au calcul des provisions techniques réglementaires. Voyons à présent l'impact du changement de loi de résiliation pour les apporteurs sur un horizon moyen terme.

## 5.2 Projection à moyen terme

### 5.2.1 Produit Net Bancaire

Pour rappel, les produits d'assurance de BPCE Assurances sont commercialisés par les Banques Populaires et les Caisses d'Epargne. En tant qu'apporteurs d'affaires, ces banques touchent une partie de la rentabilité des contrats détenus par leurs clients. Ainsi, pour chaque contrat, une commission exprimée en pourcentage de la prime est versée à la banque dont il dépend. Le Produit Net Bancaire représente pour un établissement la somme des commissions des contrats qu'il a vendus.

Le PNB est donc pour une banque un indicateur de rentabilité important, c'est pourquoi il est intéressant de le prédire. Toutefois, le PNB est très sensible au nombre de contrats souscrits, et donc à la stratégie marketing de la banque, qui est un facteur très changeant au cours du temps. Nous nous intéressons dans cette section au PNB des contrats Assurance Famille par banque.

La prédiction du PNB à court terme est la plupart du temps calculée avec un taux de résiliation annuel constant. Le calcul du PNB se distingue en deux étapes : la projection des affaires nouvelles attendues et la projection des contrats en stock. C'est ce dernier point qui nous intéresse particulièrement, car le taux de résiliation constant ne prend pas en compte la diversité d'ancienneté des contrats en stock.

La méthode actuellement utilisée pour estimer le PNB à court terme des contrats en stock est la suivante :

- Écoulement du nombre de contrats en stock par formule (diminué des résiliations et des décès). La distinction par formule est très importante car le montant de commission varie en fonction des risques couverts dans le contrat.
- Calcul de l'âge moyen et du capital moyen par formule, que l'on suppose fixes dans le temps. On suppose également que la répartition du nombre de contrats par formule est constante dans le temps (ce qui est une hypothèse très forte).
- Calcul du PNB en multipliant le nombre de contrats en stock par formule avec le montant de commission correspondant à l'âge moyen et au capital moyen par formule, auquel est ajouté une marge. Le montant de commission est donné par la grille tarifaire du produit Assurance Famille.

Un exemple de calcul de PNB sur un an est présent ci-dessous en prenant un taux de résiliation annuel de 7%. Les chiffres ont été modifiés en respectant la cohérence des résultats.

| **Formule** | **Stock n** | **Stock n+1** | **Proportion** | **Âge moy.** | **Capital moy.** | **PNB n+1** |
|---|---|---|---|---|---|---|
| Accident | 20 | 19 | 0,07 % | 55 ans | 32 000€ | 500€ |
| Confort | 10 000 | 9 300 | 35,7% | 34 ans | 23 000€ | 160 000€ |
| Essentiel | 10 000 | 9 300 | 35,7% | 34 ans | 22 000€ | 140 000€ |
| Premium | 8 000 | 7 440 | 28,6% | 36 ans | 26 000€ | 150 000€ |

TABLE 5.2 – Calcul du PNB - Exemple

### 5.2.2 Calculs au global et par apporteur

Nous comparons le nombre de contrats en stock, le chiffre d'affaires et le PNB des Banques Populaires pour les produits Assurance Famille obtenus avec un taux de résiliation constant égal à 7% (taux de résiliation global annuel des contrats Famille en 2022) et obtenus grâce à la loi de résiliation par ancienneté. Les axes des ordonnées ne sont pas affichés par souci de confidentialité.

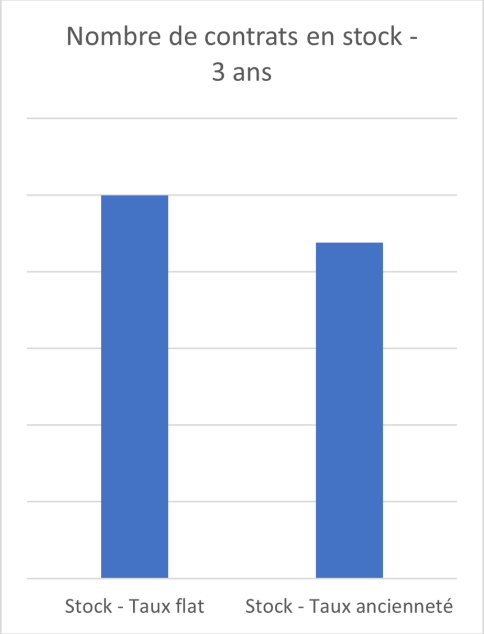


FIGURE 5.2 – Nombre de contrats en stock (toutes BP) - Projection sur 3 ans

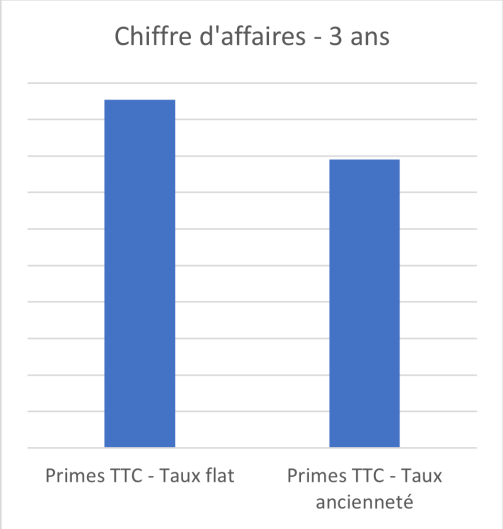


FIGURE 5.3 – Primes TTC (toutes BP) - Projection sur 3 ans

Les résultats obtenus grâce à la loi par ancienneté sont plus faibles par rapport à ceux obtenus par le taux flat. Regardons maintenant les écarts par apporteur.

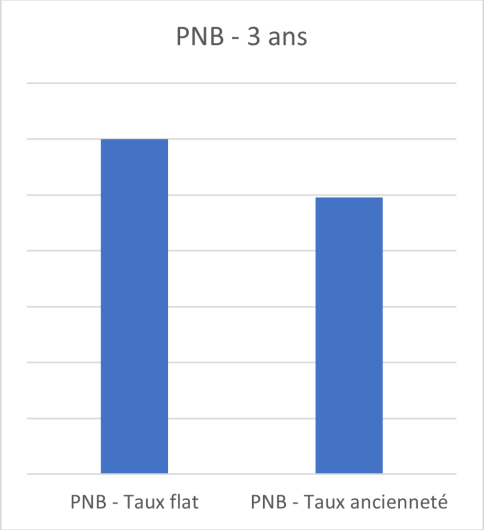


FIGURE 5.4 – PNB (toutes BP) - Projection sur 3 ans

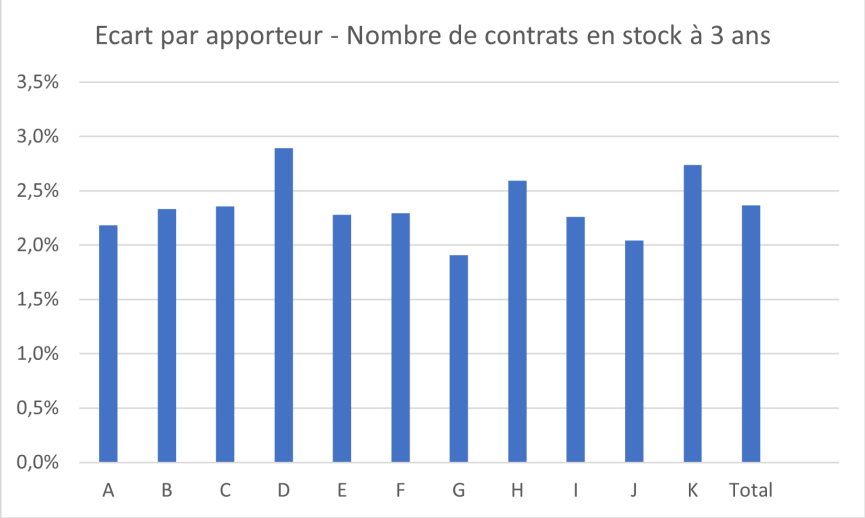


FIGURE 5.5 – Ecart du nombre de contrats en stock par BP - Projection sur 3 ans

Nous pouvons observer une hétérogénéité relative entre les différentes Banques Populaires. L'utilité de cette modélisation plus fine de la loi comportementale apparaît clairement au vu des graphiques précédents. Ainsi, un scénario central qui se base sur une hypothèse de taux flat égal au taux annuel des contrats Famille pour l'année 2022 surestimerait le nombre de contrats en stock et donc la production et le taux d'équipement à 3 ans, avec un risque d'être en dessous des objectifs commerciaux fixés. La figure 5.4 montre que la loi construite par ancienneté permettrait de fixer des objectifs à priori plus réalistes par apporteur, dans l'hypothèse que les profils d'attrition restent inchangés pour chaque banque.

De façon itérative, chaque apporteur pourrait donc se positionner sur la loi comportementale que l'actuaire lui propose par défaut afin de fixer ou non un autre seuil, par exemple par comparaison avec les profils des autres banques. En plus des indicateurs de production d'affaires nouvelles, le suivi du taux d'attrition infra-annuelle par cohorte année-mois par apporteur paraît intéressant à mettre en place.

# Conclusion

L'objectif de ce mémoire était d'étudier et mieux comprendre le comportement de résiliation des contrats Famille, ainsi que d'améliorer les lois de résiliation actuellement utilisées dans le modèle de projection des flux de trésorerie et dans les projections à moyen terme.

La phase exploratoire menée en premier lieu a permis de diagnostiquer un fort taux de résiliation des contrats dans les douze premiers mois de leur souscription, accompagné d'une forte variabilité entre les apporteurs. Dès lors, nous avons cherché à détecter les facteurs explicatifs qui pourraient être sources de plans d'action pour diminuer ces résiliations précoces et donc augmenter le taux de rétention de ces contrats.

L'autonomie dans l'extraction des données volumineuses bien qualifiées nous a incité à tester des modèles d'apprentissage automatiques supervisés. Ces modèles ont retourné des variables explicatives similaires, liées aux conditions de souscription des contrats. La régression logistique, bien qu'étant simple et frugale, s'est avérée être le modèle le plus adapté pour prédire la résiliation précoce. L'interprétabilité intuitive de ce modèle est également un atout lors d'échanges avec des responsables commerciaux.

D'un point de vue méthodologique, l'accès à l'entrepôt de données Prévoyance Individuelle a également permis d'étudier la résiliation des contrats Famille avec une analyse par cohorte année-mois, chose qu'il n'était pas possible de faire jusque là, ainsi que la mise en production d'un modèle de détection de la résiliation précoce à un an qu'il conviendra d'améliorer par la suite. Ces données pourront être utilisées dans le cadre d'études *data science* plus larges menées par le Groupe BPCE auxquelles sont souvent associés d'autres métiers, notamment les actuaires.

L'étude de la résiliation des contrats Famille s'est poursuivie par la construction de lois par âge et par ancienneté grâce à des modèles de survie classiques de type Kaplan-Meier et Hoem. Nous avons ensuite mesuré l'impact de la loi par âge sur les provisions *Best Estimate*, qui s'est révélé très faible. Nous en avons conclu qu'une loi de résiliation spécifique par produit n'est pas pertinent, le *Best Estimate* étant peu sensible à une variation des taux de résiliation.

Nous pouvons cependant souligner que l'utilisation d'une loi de résiliation par ancienneté pour les projections à moyen terme apporte un réel avantage et permet une modélisation plus fine de l'écoulement des contrats. Ainsi, des lois spécifiques par cohorte et apporteur sont très intéressantes pour la fixation d'objectifs dans le cadre des plans à moyen terme.

# Table des figures

# Liste des tableaux

# Bibliographie


[COR22] CORRIC J. *Quelle est la plus grande banque française ?* Juill. 2022. URL : https://www.agefi.fr/banque-assurance/actualites/quotidien/20220720/quelle-est-plus-grande-banque-francaise-347280.

[Jou17] Journal officiel. *Règlement délégué (UE) 2017/2358*. 21 sept. 2017.

[LE 18] LE CORRE A., RIAUX C. *La Directive sur la Distribution d'Assurance : Appréhender les enjeux de la réforme en matière de protection de la clientèle*. 17 mai 2018. URL : https://www.optimind.com/medias/documents/2612/la-directive-sur-la-distribution-d-assurance-dda.pdf.

[Ora23] Oracle. *Data Warehouse : Qu'est-ce que c'est ?* 2023. URL : https://www.oracle.com/fr/database/data-warehouse-definition/.

[CNI18] CNIL. *RGPD : de quoi parle-t-on ?* 18 avr. 2018. URL : https://www.cnil.fr/fr/rgpd-de-quoi-parle-t-on.

[Ent22] Entreprendre Service Public. *Obligations en matière de protection des données personnelles*. Avr. 2022. URL : https://entreprendre.service-public.fr/vosdroits/F24270.

[Cod19] Code de la Défense. *Article L4123-9-1*. Juin 2019. URL : https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000037825223.

[Com19] Communauté Datascience BPCE. *Relative Risks and Odds Ratio : How to interpret them*. Interne au Groupe BPCE. Juin 2019.

[CHE23] CHESNEAU C. *Introduction aux arbres de décision*. Mars 2023.

[ROK05] ROKACH L., MAIMON O. *Decision trees*. Jan. 2005.

[KRY18] KRYSHTOPENKO A. *Etude de l'attrition en prévoyance individuelle : de la construction de lois à l'interprétation des modèles d'apprentissage automatique*. Mémoire d'actuariat. 2018.

[BRE01] BREIMAN L. *Random Forests*. 2001.

[MAC22] MACHADO L., HOLMER D. *Credit risk modelling and prediction : Logistic regression versus machine learning boosting algorithms*. 2022.

[JOH19] JOHN-MATHEWS J. *Interprétabilité en Machine Learning, revue de littérature et perspectives*. Avr. 2019. URL : https://www.researchgate.net/publication/343134828_Interpretabilite_en_Machine_Learning_revue_de_litterature_et_perspectives.

[PEL21] PELTRE N., SALAUN T. *Les valeurs de Shapley en intelligibilité locale des modèles*. 11 jan. 2021. URL : https://www.quantmetry.com/blog/valeurs-de-shapley/.

[JEN13] JENI L., COHN J., DE LA TORRE F. *Facing Imbalanced Data : Recommandations for the Use of Performance Metrics*. Humaine Association Conference on Affective Computing and Intelligent Interaction. 11 jan. 2013.

[FAW04] FAWCETT T. *ROC Graphs : Notes and Practical Considerations for Researchers*. HP Laboratories. 16 mars 2004.

[DAV04] DAVIS J., GOADRICH M. *The Relationship Between Précision-Recall and ROC Curves*. 16 mars 2004.

[SAI15] SAITO T., REHMSMEIER M. *The Precision-Recall Plot is more informative than the ROC Plot when evaluating binary classifiers on imbalanced datasets*. 4 mars 2015.

[SER20] SERRE A. *Understanding Lift Curve : A brief introduction to lift curve usage in marketing and machine learning*. 16 fév. 2020. URL : https://medium.com/analytics-vidhya/understanding-lift-curve-b674d21e426.

[KUM20] KUMAR S. *7 Ways to Handle Missing Values in Machine Learning : Popular strategies to handle missing values in the dataset*. 24 juill. 2020. URL : https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e.

[GAR22] GARG S. *How to Deal with Categorical Data for Machine Learning*. 4 août 2022. URL : https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html.

[ALE17] ALENCAR A. *Resampling strategies for imbalanced datasets*. 15 nov. 2017. URL : https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets.

[Mil20] MILA, BCG GAMMA, HAVERFORD COLLEGE, COMET.ML. *CodeCarbon*. 2020. URL : https://mlco2.github.io/codecarbon/.

[sci23] SCIKIT LEARN. *Probability calibration*. 2023.

[HUB94] HUBER-CAROL C. *Durées de survie tronquées et censurées*. Journal de la Société de statistique de Paris. 1994. URL : http://www.%20numdam.org/item/JSFS_1994__135_4_3_0/.

[PLA11] PLANCHET F., THEROND P. *Modélisation statistique des phénomènes de durées*. Economica. 2011. URL : http://www.numdam.org/item/JSFS_1994__135_4_3_0/.

[HOE76] HOEM J. *The Statistical Theory of Demographic Rates : A Review of Current Developments*. Scandinavian Journal of Statistics. 1976.

[NDI16] NDIAYE E. *Construction d'une table de mortalité et lissage par positionnement*. 2016.

[BAL13] BALTESAR B. *Construction d'une table de mortalité sur un portefeuille de temporaire décès*. Mémoire d'actuariat, ISFA. 2013.

[KAP58] KAPLAN, E.L., MEIER, P. *Nonparametric Estimation from Incomplete Observations*. Journal of the American Statistical Association. 1958.

[SAI21] SAINT PIERRE P. *Introduction à l'analyse des données de survie*. 2021.

[PLA22] PLANCHET F. *Méthodes de lissage et d'ajustement*. Support de cours. 2022.

[MAZ14] MAZZA A., PUNZO A. *DBKGrad : An R Package for Mortality Rates Graduation by Discrete Beta Kernel Techniques*. Journal of Statistical Software. 2014.

[ACP19] ACPR. *Solvabilité II*. 3 mars 2019. URL : https://acpr.banque-france.fr/europe-et-international/assurances/reglementation-europeenne/solvabilite-ii.

# Annexes

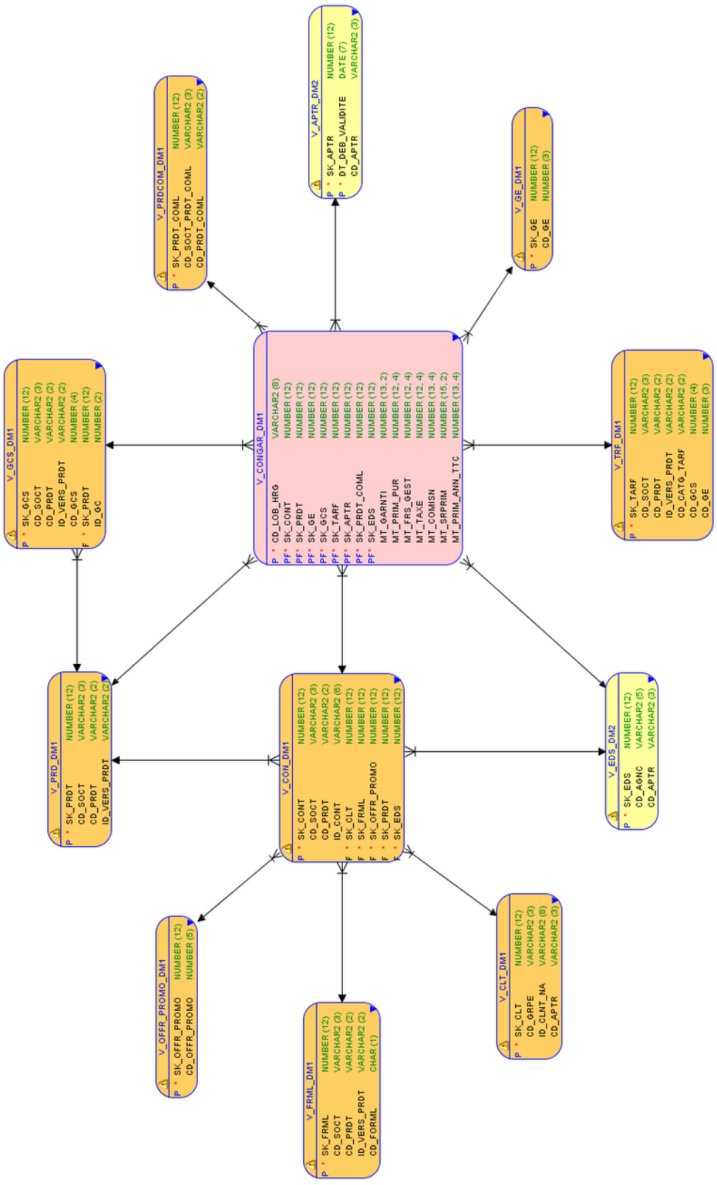

Figure 5.6 – Modèle de données du domaine Contrat Garantie du DWH PI

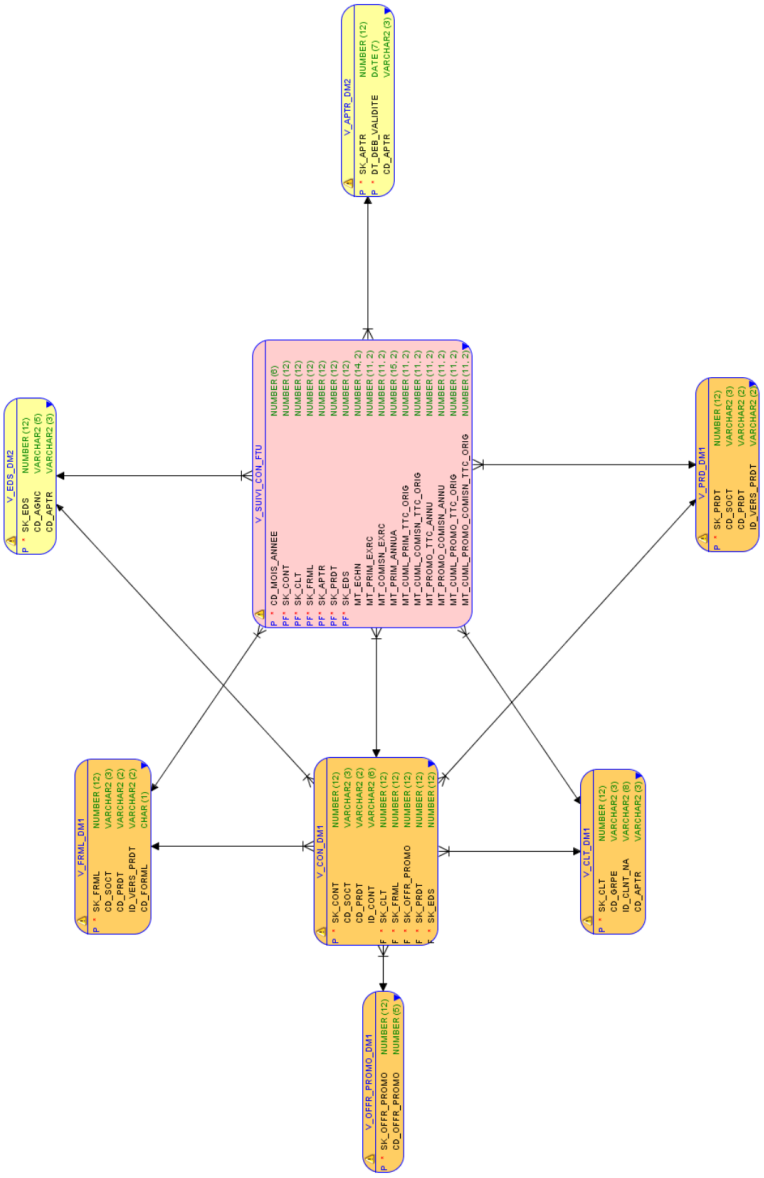


FIGURE 5.7 – Modèle de données du domaine Suivi Contrat du DWH PI

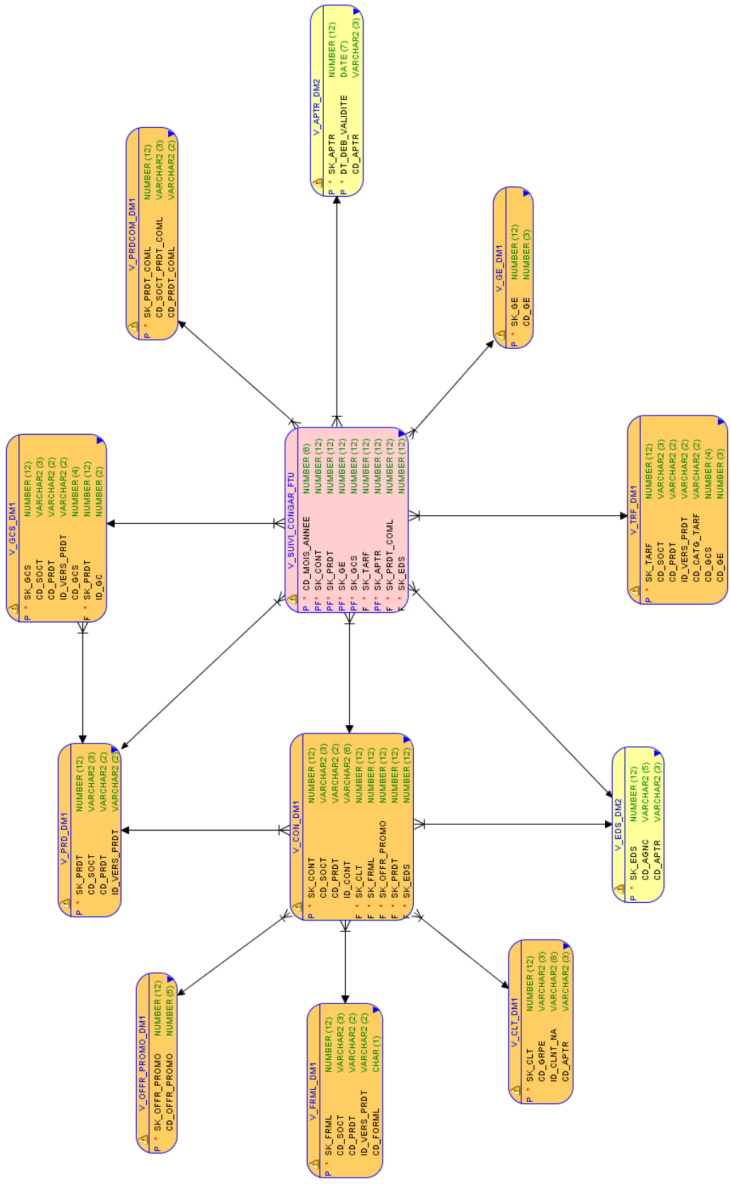


FIGURE 5.8 – Modèle de données du domaine Suivi Contrat Garantie du DWH PI

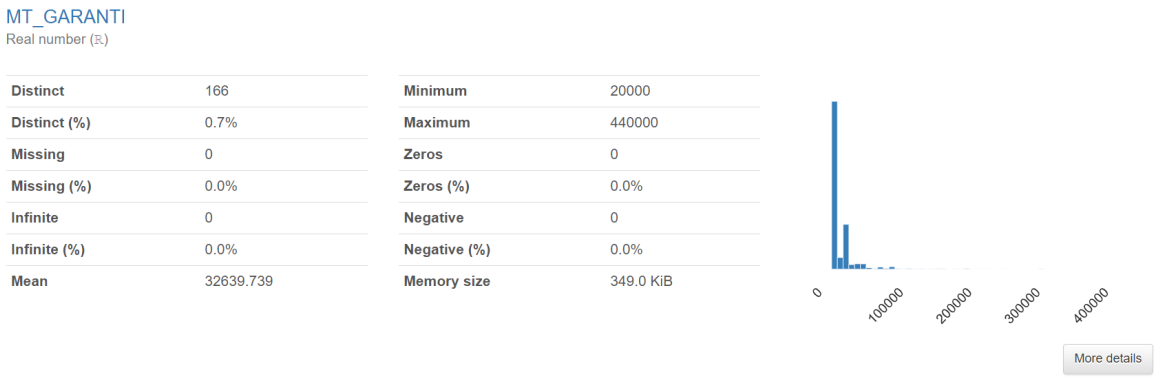


FIGURE 5.9 – Extrait d'un *Profile Report* Python

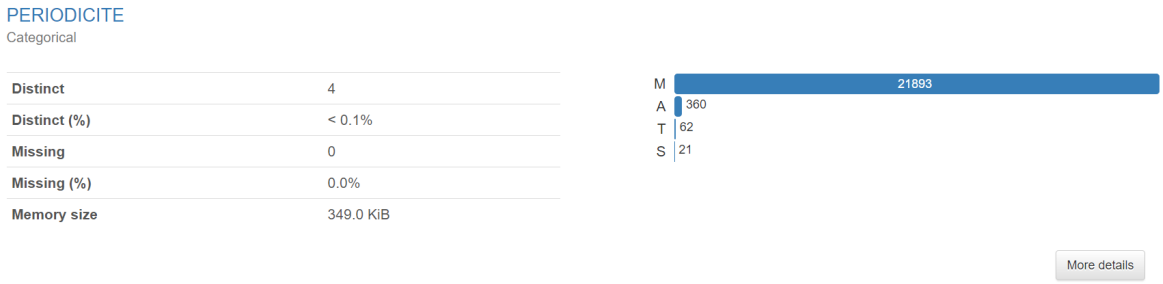


FIGURE 5.10 – Extrait d'un *Profile Report* Python

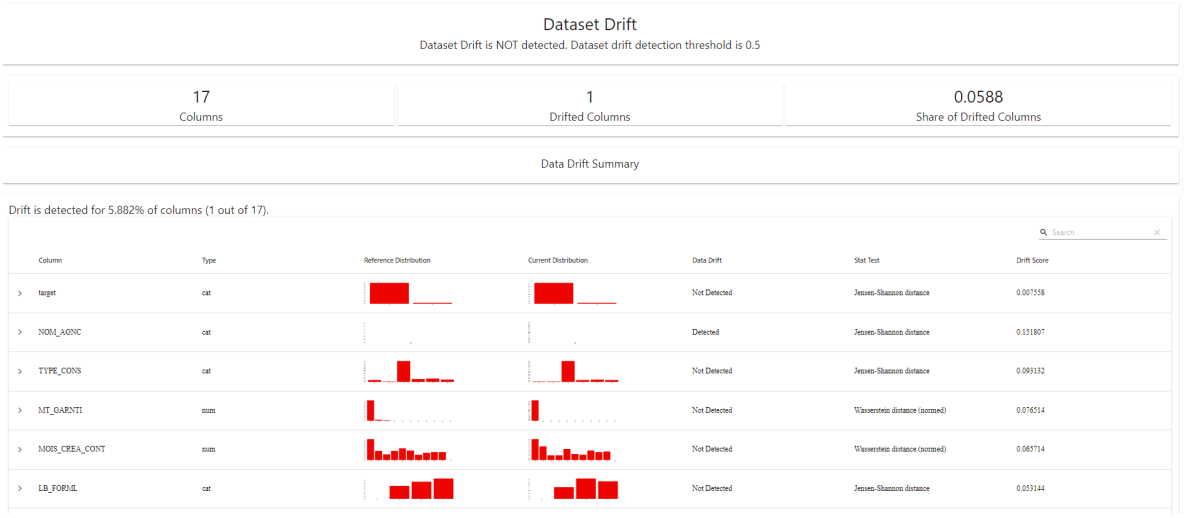


FIGURE 5.11 – Extrait d'un *data drift report* Python