IRM | institut du risk management

Mémoire présenté devant l'Institut du Risk Management pour la validation du cursus à la Formation d'Actuaire de l'Institut du Risk Management et l'admission à l'Institut des actuaires le

Par : Benoit BONENFANT

Titre : << MODELISATION DES VERSEMENTS LIBRES SOUS IFRS 17 >>

Confidentialité : ☐ NON ☐ OUI (Durée : ☐ 1an ☒ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

Membres présents du jury de l'Institut du Risk Management :

Secrétariat :

Bibliothèque :

Entreprise : AXA FRANCE

Nom : GUERRAULT Xavier

Signature et Cachet :

Directeur de mémoire en entreprise :

Nom : GUERRAULT Xavier

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature(s) du candidat(s)

SAS au capital de 96 550 euros - Déclaration d'activité enregistrée sous le n° 11 75 44451 75 auprès du Préfet de Région IDF
4, rue Chauveau-Lagarde - 75008 Paris - Tél : 01 44 51 72 79 - Fax : 01 44 51 72 73
Email : formation@institutdesactuaires.fr
Siret : 512 264 441 000 19 - Code NAF : 8559 A - TVA intracommunautaire : FR 25 512264441 - RCS Paris : 512 264 441

Table des matières

Remerciements

Avant toute chose, je tiens à remercier les personnes qui m'ont soutenu, aidé et appuyé tout au long du cycle de formation du Centre d'Etudes Actuarielles et pour l'élaboration de ce mémoire.

Je remercie ma femme Marie et mes quatre enfants Anselme, Louise, Mathilde et Charles de m'avoir accompagné dans ce projet.

Je souhaite remercier Mme Hélène LECOQ, M. Philippe METAIS et plus généralement la famille actuarielle d'AXA France d'avoir soutenu mon choix de suivre la formation du CEA.

Je remercie également M. Michel FONTEZ de m'avoir proposé ce sujet de mémoire au sein de la Direction du Risk Management d'AXA France.

Je remercie tout particulièrement M. Xavier GUERRAULT et son équipe Mme Leslie GAMET, Mme Carelle MERLO, Mme Corinne CHERKI et M. Yacov HAMOU d'avoir accepté de m'accompagner opérationnellement dans la réalisation de mon mémoire.

Je remercie enfin l'ensemble des professeurs du CEA qui m'ont beaucoup apporté durant les années du cycle de formation.

Résumé en français avec mots clés

Résumé

La norme IFRS 17 sur les contrats d'assurance qui remplacera la norme IFRS 4 a été publiée en mai 2017. Elle repose sur le principe d'une évaluation économique des passifs.
Pour les contrats d'épargne, la frontière des portefeuilles est différente sous IFRS 17, puisqu'elle implique désormais d'intégrer les versements libres futurs non programmés.

Ce mémoire a pour objectif de proposer une approche avancée de modélisation des versements libres qui puisse être déployée opérationnellement. Cette étude a été réalisée sur un portefeuille historique de contrats épargne, d'un volume significatif et toujours ouverts à la commercialisation.

La démarche retenue a été tout d'abord de classifier des variables explicatives à l'aide d'algorithmes de Machine Learning (forêts aléatoires, XGBoost) en vue de modéliser la fréquence de versement libre par contrat. Les travaux de modélisation ont ensuite été étendus au montant moyen de versement libre par contrat.

Plusieurs méthodes ont été appliquées pour modéliser les versements libres : de la plus simple utilisant des triangles de versement aux plus avancées basées sur le modèle linéaire généralisé et son extension le modèle additif généralisé.

Certains couples méthodes / variables ont permis d'obtenir des résultats significativement meilleurs que l'approche habituellement retenue.

Mots clés

IFRS 17, Versements libres, Machine Learning, Forêts aléatoires, XGBoost, Arbre de régression, Classification, Modèles linéaires généralisés, Modèles additifs généralisés, Splines de lissage.

Résumé en anglais avec mots clés

Abstract

The international standard for accounting for insurance contracts IFRS 17, which will replace IFRS 4, was published in May 2017. It is based on the principle of an economic valuation of liabilities.
For savings contracts, the portfolio frontier is different under IFRS 17, since it now requires the inclusion of the future unscheduled free payments.

This thesis aims to provide an advanced free payments modeling approach that can be deployed operationally. This study was carried out on a historical portfolio of savings contracts, of a significant volume and still available for sale.

First of all, the approach adopted was to classify explanatory variables using Machine Learning algorithms (random forests, XGBoost) in order to model the frequency of free payment by contract. The modeling work was then extended to the average amount of free payment per contract.

Several methods have been applied to model free payments: from the simplest using payment triangles to the most advanced based on the generalized linear model and its extension the generalized additive model.

Some combination of methods and variables provide significant better results than the usual approach.

Key words

IFRS 17, Free payments, Machine Learning, Random forests, XGBoost, Regression tree, Classification, Generalized linear models, Generalized additive models, Smoothing splines

Introduction

En mai 2017, la norme IFRS 17 sur les contrats d'assurance qui remplacera la norme IFRS 4 a été publiée. Reportée une seconde fois en mars 2020, la date d'entrée en vigueur est désormais prévue au 1er janvier 2023. Cette nouvelle norme traite de l'évaluation des engagements d'assurance et de la reconnaissance du résultat d'assurance.

Ce nouveau référentiel comptable repose sur le principe d'une évaluation économique des passifs d'assurance composés d'un « Best Estimate », d'une marge pour risque et de la valeur actuelle des profits futurs, comme dans le cadre de la directive Solvency II. Toutefois, elle introduit de nombreuses spécificités (maille du calcul, contrats onéreux, logique de présentation de la performance) occasionnant des défis d'implémentation opérationnelle et une modification des processus et de l'organisation.
En particulier, cette norme prévoit désormais pour le calcul des provisions techniques des contrats d'épargne, la prise en compte des flux futurs, ce qui sous-entend la nécessité de modéliser le comportement futur des assurés (loi de rachat, loi d'arbitrage fonds euros / unités de compte, loi de versement des primes futures et notamment des versements libres, etc.).
En effet, la nouvelle norme IFRS 17 introduit une modification de la frontière des contrats en Epargne en France qui se limitait jusqu'alors, dans le cadre de Solvabilité II, aux primes reçues jusqu'à la date de comptabilisation ainsi qu'aux versements futurs programmés. La frontière des contrats implique désormais d'intégrer également les versements libres futurs (non programmés) sur les contrats en portefeuille à la date de comptabilisation.

Ce mémoire a pour objectif d'étudier le comportement de versement libre des assurés d'un portefeuille donné et d'identifier des modèles de prédiction performants pour modéliser les versements libres. Il s'agit de construire des modèles qui puissent être déployés opérationnellement dans le modèle interne de projection des « cash flows » d'AXA France. De nombreux travaux de mémoire d'actuariat ont été réalisés sur la modélisation des comportements de rachat des assurés, mais assez peu sur les versements libres (fréquence et montant moyen de versement).

Dans le premier chapitre, sont présentés d'une part les grands principes de la nouvelle norme IFRS 17 (en particulier les principales évolutions sur le calcul des engagements) et d'autre part les travaux d'analyse statistique descriptive de la base de données considérée pour l'étude. La deuxième partie est consacrée à l'introduction des outils mathématiques utilisés dans le cadre des travaux de recherche (le modèle linéaire généralisé, le modèle additif généralisé, le Machine Learning et les algorithmes des forêts aléatoires et XGBoost). Dans la troisième partie, les travaux de classification des variables explicatives et de construction de modèles pour la prédiction des versements libres sont présentés.

1. Contexte de l'étude et présentation des données

1.1. Contexte de l'étude : IFRS 17 et versements libres en épargne individuelle

1.1.1. Origine des normes IFRS

En 1973, les institutions comptables d'Australie, du Canada, de France, d'Allemagne, du Japon, du Mexique, des Pays-Bas, du Royaume-Uni, d'Irlande et des Etats-Unis ont créé une nouvelle organisation en vue de faire converger les normes comptables locales vers un même référentiel : l'IASC (International Accounting Standards Committee).
L'IASC a mis en place les normes IAS (International Accounting Standards) dans le but d'inciter les sociétés internationales à fournir des résultats plus transparents et universels pour les investisseurs.
En 2000, l'IASC a été renommée l'IASB (International Accounting Standards Board) et les normes IAS sont appelées dorénavant les normes IFRS (International Financial Reporting Standards).

En 2020, les normes IFRS sont requises pour les sociétés cotées dans plus de 150 juridictions dans le monde.

FIGURE 1 : Carte du monde présentant en bleu les pays où les normes IFRS sont requises pour les sociétés cotées (source : www.IFRS.org)

L'application de ces normes peut s'avérer obligatoire selon que l'entreprise est cotée ou non.

Au sein de l'Union Européenne, les entreprises cotées sont tenues d'établir leurs comptes consolidés conformément aux normes comptables internationales IFRS.
En effet, l'application des normes IFRS en Europe a été précisée par le Règlement (CE) n°1606/2002, dit « règlement IAS », relatif à l'application des normes comptables internationales.
L'article 4 de ce règlement précise que depuis le 1er janvier 2005 les sociétés cotées européennes sont tenues de préparer leurs comptes consolidés conformément aux normes IFRS.
L'utilisation du référentiel IFRS pour les comptes consolidés des sociétés non cotées est optionnelle en France.

Le tableau suivant reprend les obligations actuelles en France :

	Comptes sociaux	Comptes consolidés
Sociétés cotées	Normes françaises	Normes IFRS
Sociétés non cotées	Normes françaises	Normes IFRS sur option

TABLEAU 1 : Obligations IFRS en France selon que la société est cotée ou non (source : ACPR[1])

1.1.2. Les principales normes IAS/IFRS ayant un impact sur les comptes des assureurs

- **IAS 1 - Présentation des états financiers :** L'IAS 1 énonce les dispositions générales relatives à la présentation des états financiers (bilan, compte de résultat et tableau des flux de trésorerie, etc.) et des lignes directrices concernant leur structure et les dispositions minimales en matière de contenu.
- **IAS 39 → IFRS 9 (à partir du 1er janvier 2023) - Comptabilisation et évaluation des instruments financiers (à l'actif du bilan) :**
  La norme IFRS 9, qui remplacera la norme IAS 39, entrera en application en France le 1er janvier 2023. Cette nouvelle norme apporte un certain nombre de changements quant au classement et à la valorisation des instruments financiers et introduit notamment une nouvelle méthodologie de dépréciation des créances sur la base des pertes attendues (dites Expected Credit Loss - ECL) et non plus des pertes avérées.
- **IFRS 7 - Informations à communiquer sur les instruments financiers et leur degré d'importance pour la société.**
- **IFRS 4 → IFRS 17 (à partir du 1er janvier 2023) - Comptabilisation et évaluation des contrats d'assurance (au passif du bilan).**

1.1.3. Comptabilisation et évaluation des contrats d'assurance : d'IFRS 4 à IFRS 17

La norme IFRS 4, entrée en vigueur le 1er janvier 2005, a pour objectif de spécifier l'information financière pour les contrats d'assurance émis et les traités de réassurance détenus par la compagnie d'assurance (Bibliographie : 1. et 2.).

Les points clés d'IFRS 4 sont :

- **La définition de la notion de contrat d'assurance :**
  « Un contrat d'assurance est un contrat selon lequel une partie (l'assureur) accepte un risque d'assurance significatif d'une autre partie (le titulaire de la police) en convenant d'indemniser le titulaire de la police si un événement futur incertain spécifié (l'événement assuré) affecte de façon défavorable le titulaire de la police ».
  « Le risque d'assurance est un risque, autre que le risque financier, transféré du titulaire du contrat à l'émetteur ».
  « Le risque est significatif si et seulement si un événement assuré peut obliger un assureur à payer des prestations complémentaires significatives dans n'importe quel scénario à l'exclusion des scénarios qui manquent de substance commerciale ».

[1] Autorité de Contrôle Prudentiel et de Résolution (ACPR)

- **La notion de « shadow PB » / « shadow accounting » :**
  Dans la plupart des normes locales, les actifs sont comptabilisés en valeur historique (valeur fondée sur les coûts d'acquisition amortis).
  La « shadow PB » enregistrée au passif a pour objectif de réduire le décalage entre l'actif et le passif, puisqu'à l'actif, les instruments financiers dont comptabilisés quant à eux en valeur de marché (IAS 39 / IFRS 9).
  L'idée est de comptabiliser les plus ou moins-values latentes pour les actifs valorisés en valeur de marché impactant le passif sous la forme de « provisions pour participation aux bénéfices différée ».
- **La nécessité de mettre en place un test sur la suffisance des passifs d'assurance : « Liability Adequacy Test » (LAT).**
  Le test du LAT repose sur la comparaison des passifs d'assurance (les Provisions Mathématiques et la Provision pour Participation aux Excédents) avec les Flux futurs de Trésorerie : si PM + PPE > FT, le résultat de l'entreprise serait alors positif.

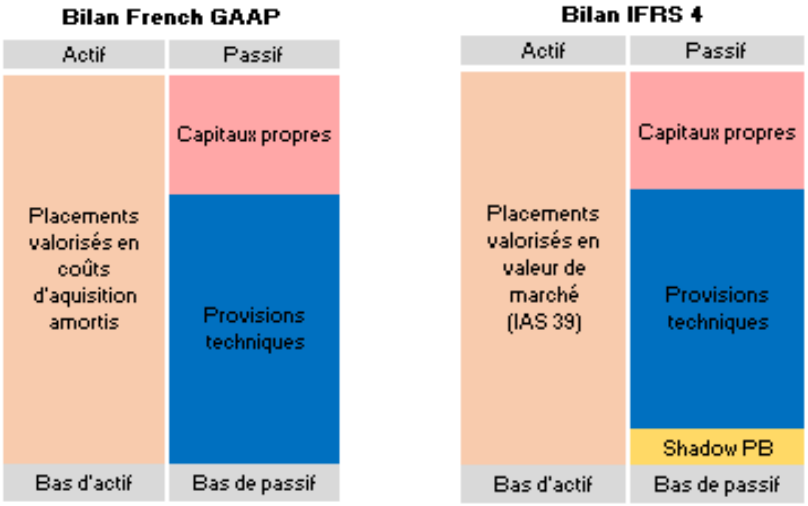


FIGURE 2 : Présentation des grands postes des bilans sociaux français (French GAAP) et IFRS 4

Les normes comptables pour un même type de contrat émis dans des différents pays peuvent différer de manière très significative. Le fait qu'IFRS 4 repose sur des normes locales va à l'encontre du principe fondamental d'IFRS visant à faire converger les normes comptables locales vers un référentiel commun. Ainsi, en vue d'uniformiser et d'optimiser la comptabilisation et l'évaluation des contrats d'assurance, la norme IFRS 17 remplacera la norme IFRS 4 à partir du 1er janvier 2023.
Les principaux objectifs de cette norme sont :

- réduire les différences d'évaluation des contrats d'assurance entre pays (plus de recours au local GAAP),
- valoriser les options et les garanties des contrats d'assurance,
- et favoriser la cohérence avec les autres normes IFRS (valorisation en juste valeur / « fair value »).

Ainsi les provisions techniques au passif du bilan seront décomposées en trois parties :

- **La partie « Best Estimate » (BE**) qui correspond à l'estimation des cash-flows futurs des contrats d'assurance (primes, sinistres, intérêts et participations aux bénéfices versées, frais) pondérés par leur probabilité d'occurrence et actualisés de manière « market-consistent » et incluant la valeur temps des options et garanties (calcul stochastique). Ils doivent refléter de la manière la plus fiable et la plus neutre possible les engagements de l'assureur.

- **La partie « Risk Adjustment » (RA)** qui correspond à la marge que demanderait un autre assureur pour reprendre le risque assuré, compte tenu de l'incertitude sur le montant de la partie « Best Estimate ».
  A noter : aucune méthode de calcul n'est imposée (par exemple pour estimer le coût du capital).
- **La partie « Contractual Service Margin » (CSM)** qui correspond à la marge calibrée à la souscription du contrat en considérant qu'aucun profit comptable n'est dégagé à la souscription.
  Cette marge résiduelle est ensuite reprise au fur et à mesure de la vie du contrat, permettant de dégager progressivement un profit comptable.
  La CSM est amortie et les bénéfices sont comptabilisés au fur et à mesure que les prestations d'assurance sont réglées pendant la période de couverture du contrat.
  Néanmoins, les pertes résultantes de contrats onéreux sont comptabilisées quant à elles en résultat courant (immédiatement).

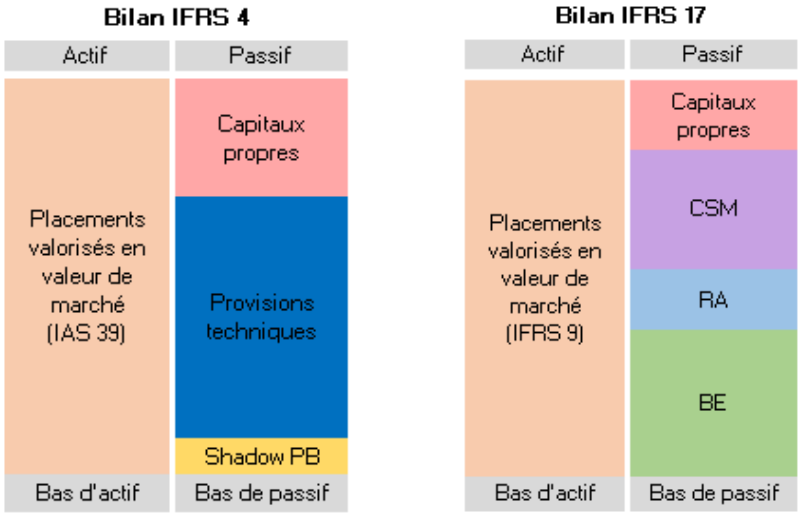


FIGURE 3 : Présentation des grands postes des bilans IFRS 4 et IFRS 17

Les principales évolutions sur le calcul des provisions techniques sont :

- En assurance non-vie, l'actualisation des flux futurs en prenant en compte l'impact des branches longues (RC, construction) ;
- En assurance vie, les paramètres de calcul des provisions ne sont plus figés à la souscription (taux d'actualisation) ;
- La possibilité d'intégrer le comportement futur du top management de la compagnie d'assurance : « management actions » (modification de l'allocation d'actifs ou de la politique de couverture, modification de la politique de PB, etc.) ;
- La prise en compte des primes futures possible dans certains cas : nécessité de modéliser le comportement futur des assurés (loi de rachat, loi d'arbitrage fonds euros/unités de compte, loi de versement des primes futurs, etc.).

1.1.4. Rapprochement entre IFRS 17 et directive Solvabilité II

La directive européenne Solvabilité II, entrée en vigueur au 1er janvier 2016, a pour principal objectif de mieux adapter les fonds propres exigés des compagnies d'assurance et de réassurance aux risques que celles-ci encourent dans leur activité et ainsi mieux protéger les assurés.
Les assureurs européens ont beaucoup investi dans le développement d'outils de modélisation et de reporting pour répondre aux exigences de cette directive. Ainsi, les assureurs français peuvent opter pour la « formule standard » ou bien ils peuvent développer un modèle interne qui devra être validé par l'ACPR.
Certains de ces assureurs, dont AXA France, ont choisi de développer des modèles en interne leur permettant de modéliser leur bilan en vision prudentiel et de piloter leur ratio de solvabilité (calcul des provisions techniques en vision « Best Estimate » et des exigences en fonds propres en tenant compte des spécificités de chaque assureur). Ces modèles modélisent les interactions entre l'actif et le passif en tenant compte de la valeur actuelle des garanties et des engagements.
Pour optimiser la performance de calcul de ces modèles, les contrats de même catégorie (garanties similaires) et de risque homogène (niveau des cotisations, niveau des prestations…) sont classifiés et agrégés pour former des groupes de contrats : les « Model Points ». Un « Model Point » est un groupe de contrats ayant des caractéristiques similaires. Ce qui réduit ainsi considérablement le nombre de calculs à réaliser par le modèle.
Pour la mise en œuvre de la norme IFRS 17, et notamment pour modéliser les flux futurs des contrats d'assurance ou de réassurance, la plupart des assureurs souhaitent capitaliser sur les modèles développés dans le cadre de Solvabilité II.

Toutefois, les conditions de modélisation des contrats ne sont pas les mêmes entre IFRS 17 et Solvabilité II (Bibliographie 28.).

	IFRS 17	**Solvabilité II**
Date du début du contrat	Au début de la couverture	A la signature du contrat
Valeur des profits futurs	La CSM fait partie des provisions techniques et non des capitaux propres comptables	La Value In Force (VIF) est incluse dans les capitaux propres règlementaires
Granularité de calcul des provisions	Risques similaires x Année d'émission x Classification (Profitables / Potentiellement onéreux / Onéreux)	Entité x Line Of Business (LOB)
Ajustement pour le risque	Risk Adjustment (RA) selon méthode à définir par chaque groupe d'assurance	Risk Margin (RM) selon méthode du coût du capital imposé par le Régulateur européen (European Insurance and Occupational Pensions Authority - EIOPA)
Courbe des taux d'actualisation	Taux cohérents avec les flux du passif de l'assureur	Courbe des taux sans risque

TABLEAU 2 : Principales différences entre IFRS 17 et Solvabilité II (modélisation des contrats)

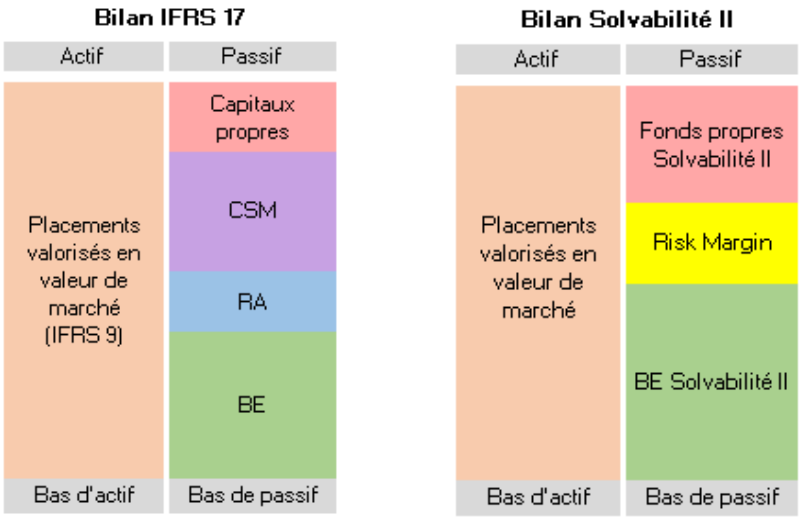


FIGURE 4 : Présentation des grands postes des bilans IFRS 17 et prudentiel Solvabilité II.

1.1.5. IFRS 17 : prise en compte des versements libres futurs

La nouvelle norme IFRS 17 introduit une modification de la frontière des contrats en Epargne en France. En effet, jusqu'alors, dans le cadre de Solvabilité II, la frontière des contrats en Epargne se limitait aux primes reçues jusqu'à la date de comptabilisation ainsi qu'aux versements futurs programmés.
Dans le cadre d'IFRS 17, la définition de la frontière des contrats est revue et implique désormais d'intégrer également les versements libres futurs (non programmés) sur les contrats en portefeuille à la date de comptabilisation.

Dans le cadre d'un contrat d'épargne, l'épargnant verse un montant initial à l'ouverture du contrat et peut effectuer des versements supplémentaires tout au long de la vie du contrat.
Parmi les versements, nous pouvons différencier :

- **Les versements programmés**, qui sont prédéfinis à l'avance, en termes de date de versement et de montant à verser ;
- **Les versements libres programmés**, qui sont prédéfinis à l'avance, en termes de date de versement mais le montant à verser reste au libre choix de l'épargnant ;
- **Les Versements Libres (VL)**, quant à eux sont non programmés, au libre choix de l'épargnant. Ils se caractérisent par leur probabilité d'occurrence (fréquence de versement) et le montant moyen en euros qui est versé.

1.2. Périmètre de l'étude

Le portefeuille de contrats d'assurance vie choisi pour réaliser les travaux de classification des variables explicatives et de modélisation des versements libres est un des portefeuilles historiques et de référence d'AXA France.

Les principaux points forts pour l'étude :

- la taille du portefeuille (> 100 000 contrats),
- un historique long,
- le contrat est toujours ouvert à la commercialisation.

1.3. Création de la base de données

1.3.1. Récupération des données

L'objectif de cette partie est de présenter les différentes étapes du processus de construction de la base de données qui a été exploitée dans le cadre des analyses actuarielles présentées dans ce mémoire.

La base de données contient notamment :

- L'ensemble des **flux en vision annuelle** (les versements initiaux, les versements libres programmés ou non programmés, les rachats totaux ou partiels) survenus sur les contrats entre le début de l'année 2010 et la fin de l'année 2017.
- Pour chaque flux annuel,
    - les caractéristiques clés du contrat sont les suivants :
        - le numéro de contrat (champ « NUMCONT » dans la base),
        - l'année d'effet du contrat (champ « ANEFFET »),
        - l'année de survenance du mouvement sur le contrat (champ « ANEFFETMVT »),
        - la fiscalité du contrat (champ « FISC »)
        - la provision mathématique d'ouverture du contrat (champ « PMOUV »),
    - les informations clés sur l'assuré sont les suivants :
        - l'âge de l'assuré (champ « AGE »),
        - le sexe de l'assuré (champ « SEXE_Vision_à_fin_2019 »).
    - les indices macro-économiques intégrés sont les suivants :
        - le Taux Moyen des emprunts d'Etat en vision annuelle (champ « TME_ANNUEL »),
        - l'Indice des Prix des Logements en vision annuelle (champ « IPL_ANNUEL »),
        - le taux de chômage en vision annuelle (champ « TAUX_CHOMAGE_ANNUEL »).
        - l'indice CAC 40 (principal indice de la bourse de Paris)
          A noter : l'indice Cotation Assistée en Continu 40 (CAC 40) est déterminé à partir des cours de 40 actions cotées en continu sur le premier marché parmi les cent sociétés dont les échanges sont les plus abondants sur Euronext Paris, qui fait partie d'Euronext, la première bourse européenne.

Source des données :

- La base de données par année de survenance a été construite à partir des bases « *flux* » mensuelles extraites des systèmes de gestion en récupérant pour chaque flux les caractéristiques clés du contrat et de l'assuré.
  Remarque : la donnée « sexe » de l'assuré a été rajouté manuellement dans un second temps.
- Les valeurs des indices macro-économiques ont été récupérés sur internet et ont été associés manuellement à chaque flux.
  Sources des indices :
  - TME : https://www.banque-france.fr/statistiques/
  - IPL et taux de chômage : https://www.insee.fr/fr/statistiques/
  - CAC 40 : https://investir.lesechos.fr/cours/

Conversion des flux mensuels en flux annuels

Pour convertir la base « flux » mensuels en vision annuelle par année de survenance (de 2010 à 2017), certains retraitements ont été réalisés :

Pour les versements libres :

- **Pour le nombre de versements libres annuel sur chaque contrat** (champ « NBVL ») : il a été décidé de ne pas tenir compte du nombre de versements infra-annuels et de ne considérer qu'un seul versement libre par an pour chaque contrat (NBVL = 1, en cas de versement libre, même s'il y a eu plusieurs versements au cours de l'année).
  Point d'attention : cette décision se justifie par le fait que les travaux de modélisation des versements libres ont vocation à être intégrés dans un modèle interne de projection des « cash flows » uniquement EN VISION ANNUELLE.
  A titre indicatif, sur les deux dernières années d'historique (2016 et 2017), le nombre de versements libres infra-annuels sur chaque contrat est compris entre 1 et 10 inclus pour plus de 50% des contrats concernés (73% pour l'année 2016 et 53% pour l'année 2017).
- **Pour le montant annuel de versements libres sur chaque contrat** (champ « VL ») : il a été décidé de cumuler les montants mensuels observés.

Pour les autres types de flux (par exemple : rachat partiel, rachat total, versement libre programmé), le nombre et le montant restent tous deux cumulés sur 12 mois (par exemple : pour un flux annuel de rachat partiel, NBRP pourra être supérieur à 1 s'il y a plusieurs rachats partiels au cours de l'année).

Pour la provision mathématique d'ouverture du contrat : un nouveau champ est créé dans la cadre de l'extraction et renseigné à partir de la valeur connue de la provision mathématique de clôture de l'année précédente.

1.3.2. Contrôles de cohérence

Le principal objectif de cette partie est de vérifier la cohérence des montants annuels en euros de versements libres renseignés dans la base de données avec les données validées par la Direction du contrôle de gestion de la Direction Financière d'AXA France.

De plus, pour contrôler la qualité de l'extraction, il a été opportun de vérifier également la cohérence des montants en euros de PM d'ouverture et de rachats.

Point d'attention : pour les contrôles suivants, les écarts observés sont principalement liés au fait que les chiffres de la Direction financière résultent d'une estimation au moment de l'arrêté de fin d'année et non d'une extraction a posteriori des données des systèmes de gestion.

Contrôle des montants de versements libres par année de mouvement.
A chaque arrêté annuel, la Direction Financière consolide sa vision des primes encaissées par exercice (APE : « Annualized Premium Equivalent »).

Les montants de primes encaissées sont déclinés à la maille produit x réseau avec la répartition entre :

- les primes uniques (« Single ») et les primes périodiques (« Regular »),
- les primes relatives aux affaires nouvelles et les primes relatives aux affaires déjà en portefeuille en début d'exercice.

Pour récupérer le montant annuel en euros de versement libres sur l'année d'exercice, il convient de :

- récupérer le total des primes uniques sur l'exercice (colonne « single ») et de le retraiter du montant total d'affaires nouvelles (colonne « Affaires nouvelles »),
- et ensuite multiplier le résultat par 10.
  En effet, la durée moyenne observée des contrats d'assurance vie épargne AXA France est d'environ 10 ans. La Direction financière d'AXA France a donc formulé l'hypothèse suivante : → APE = Montant total des primes versées divisé par 10.

Résultats du contrôle :

Année	Ecarts en euros	Ecarts en %
2010	-2 552 806	**-2%**
2011	657 794	**0%**
2012	-2 745 599	**-2%**
2013	-3 024 814	**-2%**
2014	476 133	**0%**
2015	-1 591 481	**-1%**
2016	-1 150 878	**-1%**
2017	-1 215 327	**-1%**

TABLEAU 3 : Résultats du contrôle de cohérence des montants en euros de versements libres par année de mouvement avec les données de la Direction financière d'AXA France

L'écart maximum est de 2%, ce qui est considéré comme acceptable pour réaliser cette étude.

Contrôle des montants de PM d'ouverture par année de mouvement
Résultats du contrôle :

Année	Ecarts en euros	Ecarts en %
2010	-34 827 103	**-1%**
2011	-40 752 470	**-1%**
2012	-25 226 952	**-1%**
2013	-83 640 297	**-3%**
2014	-18 506 650	**-1%**
2015	-74 444 740	**-2%**
2016	-68 680 304	**-2%**
2017	-66 074 447	**-2%**

TABLEAU 4 : Résultats du contrôle de cohérence des montants en euros de PM d'ouverture des contrats par année de mouvement avec les données de la Direction financière

L'écart maximum observé est de 3%, ce qui est considéré comme acceptable pour réaliser cette étude.

Contrôle des montants en euros de rachat par année de mouvement (rachats totaux et partiels)
Résultats du contrôle : l'écart maximum observé est de 2% ce qui est considéré comme acceptable pour réaliser les travaux d'étude.

1.4. Analyse descriptive des données

L'objectif de cette partie est de commenter certains éléments de statistique descriptive sur les versements libres de la base de données et les éventuelles relations qui pourraient être observés avec les variables explicatives (données assuré, contrat et macro-économiques).

1.4.1. Montant et fréquence de versement libre

Entre 2010 et 2017, le montant total annuel en euros de versements libres évolue à la hausse ou à la baisse, avec un « pic » à la hausse en 2015 et à la baisse en 2012 et 2016. Ces extremums peuvent être liés à des événements « intrinsèques » à AXA (à titre d'exemple : un changement de politique de souscription), ou bien aux conditions de marché de l'assurance, ou alors au contexte macro-économique.

Année	Montant de VL (€)
2010	136 809 651
2011	140 168 004
2012	**129 873 606**
2013	140 554 042
2014	152 326 548
2015	**159 718 231**
2016	**136 737 867**
2017	144 503 225

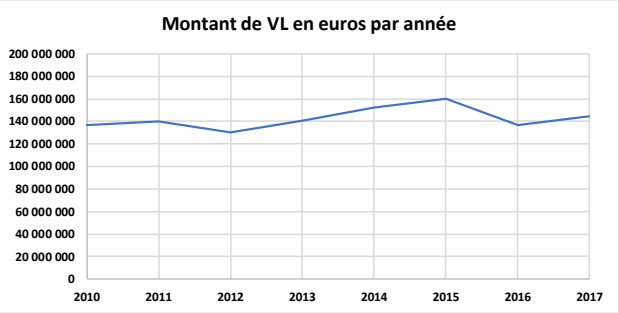


TABLEAU 5 et FIGURE 5 : Evolution du montant total annuel en euros de versements libres de 2010 à 2017

Les statistiques usuelles des montants en euros de versements libres entre 2010 et 2017 mettent en évidence :

- un versement médian à hauteur de 6 000 € (cumul 2010 à 2017), qui évolue selon l'année, avec un écart maximum de 47% entre l'année 2016 (5 010 €) et les années 2012 & 2017 (7 350 €),
- des montants de versements libres globalement plus faibles en 2016,
- une dispersion importante des maximums selon l'année (à titre d'exemple : de 1,3 M€ en 2012 et 2016 à 2,8 M€ en 2014).

Années	Minimum	1er quartile	Médiane	Moyenne	Ecartype	3ème quartile	Maximum
2010	0,01	1 684	5 463	19 897	3 438	17 470	1 750 000
2011	0,01	1 455	5 233	21 296	3 291	19 400	2 370 000
2012	0,01	1 519	7 350	23 993	2 707	24 563	1 295 864
2013	0,01	1 523	5 940	20 876	3 367	19 600	1 787 627
2014	0,01	1 823	6 000	20 300	3 752	19 716	2 850 000
2015	0,01	1 930	6 313	20 673	3 863	19 600	1 542 085
2016	0,01	1 473	5 010	19 349	3 534	19 700	1 300 000
2017	0,01	1 960	7 350	22 194	3 256	22 500	2 134 892
Cumul 2010 à 2017	**0,01**	**1 627**	**6 000**	**20 964**	**27 206**	**19 800**	**2 850 000**

TABLEAU 6 : Statistiques descriptives usuelles des versements libres en euros entre 2010 et 2017

Compte tenu du montant minimum de versements libres observé chaque année (0,01€), il a été nécessaire d'évaluer la significativité des versements libres inférieurs à 100 euros par an (seuil qui a été considéré comme significatif pour ce type de contrat).
L'analyse montre que ce type de versement ne représente que 2% du nombre total de versements observés sur la période 2010 à 2017 (entre 1,9% et 2,4% selon l'année).

Fréquence et montant moyen en euros de versements libres par année
Pour estimer le montant de versements libres par contrat, la formule de calcul qui peut être considérée est la suivante :

Montant de VL par contrat = Fréquence de VL (%) x Montant moyen de VL (€)

L'analyse de l'évolution des fréquences et des montants moyens en euros de versements libres met en évidence des variations quasi-opposés. Une des explications pourrait être que plus les versements sont fréquents et plus l'impact des valeurs extrêmes est minoré.

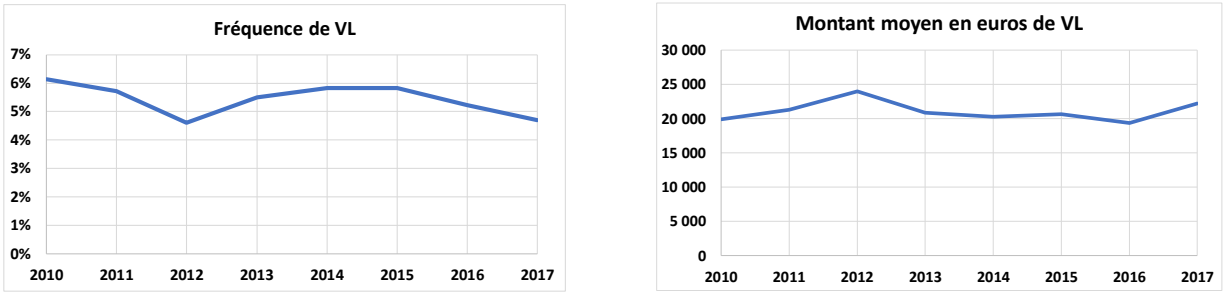


FIGURE 6 : Evolution de la fréquence et du montant moyen en euros des VL entre 2010 et 2017

1.4.2. Versements libres et données assuré

L'âge de l'assuré
Le nombre de versements libres (cumul 2010-2017) évolue en fonction de l'âge de l'assuré avec un « pic » de versements sur la période d'environ 50 à 70 ans qui correspond à une période plus active de l'épargnant notamment en vue de préparer sa retraite et plus avantageuse fiscalement qu'après 70 ans (les versements sont exonérés de droits de succession jusqu'à 152 500 € avant 70 ans versus 30 500 € après 70 ans).

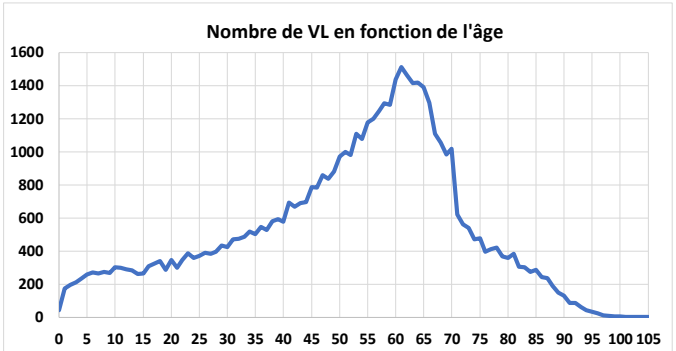


FIGURE 7 : Evolution du nombre de versements libres en fonction de l'âge de l'assuré (cumul 2010-2017)

L'observation de l'évolution de la fréquence et des montants moyens en euros de versements libres en fonction de l'âge met en évidence :

- Un « pic » de fréquence importante pour les âges inférieurs à 5 ans. En effet, les autres types de flux (comme les rachats) ne sont pas encore assez nombreux ;
- Un « pic » de fréquence entre environ 50 et 70 ans (précédemment cité) ;
- Des montants moyens de versements libres qui évoluent quasi linéairement en fonction de l'âge entre 2 et 68 ans ;
- Des montants plus importants de versements libres de manière discrétionnaire à partir de 85 ans, qui peuvent correspondre à des contextes particuliers de transmission de patrimoine.

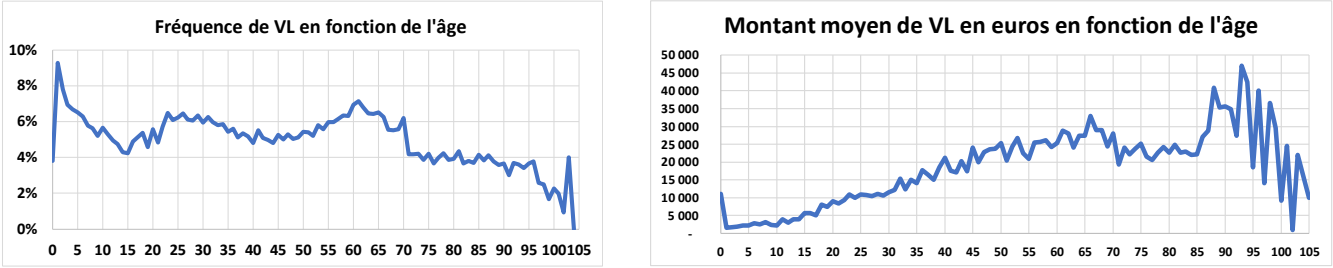


FIGURE 8 : Evolution de la fréquence et du montant moyen en euros des versements libres en fonction de l'âge de l'assuré (cumul 2010-2017)

Le Sexe de l'assuré

L'observation de la répartition du nombre et du montant moyen en euros de versements libres entre les hommes et les femmes montre que la variable « Sexe » n'est pas discriminante. Cette variable ne sera par conséquent pas pris en compte dans le cadre des travaux de classification et de modélisation.

Année	Homme	Femme	Non défini
2 010	41%	43%	15%
2 011	42%	45%	13%
2 012	43%	45%	12%
2 013	44%	47%	9%
2 014	45%	48%	7%
2 015	45%	49%	6%
2 016	45%	50%	5%
2 017	47%	50%	3%

Année	Homme	Femme	Non défini
2010	22 076	18 691	17 408
2011	22 232	20 100	22 483
2012	24 821	23 432	23 090
2013	21 173	21 014	18 694
2014	20 713	19 318	24 206
2015	20 846	19 985	25 029
2016	19 629	18 921	21 277
2017	23 377	21 255	18 861

TABLEAU 7 : Répartition homme / femme du nombre et du montant moyen en euros de versements libres par année de mouvement

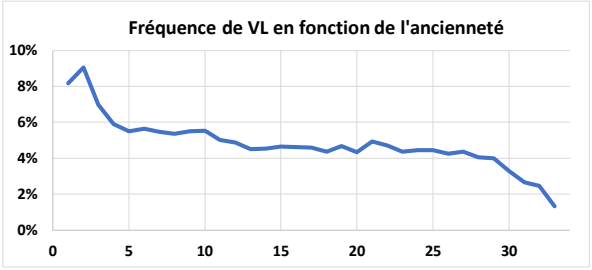
1.4.3. Versements libres et données contrat

L'ancienneté du contrat

L'observation de l'évolution de la fréquence et des montants moyens en euros de versements libres en fonction de l'ancienneté met en évidence :

- Une fréquence de versements libres plus importante avec un « pic » juste après la souscription du contrat (autour de 9% en 2ème année), qui diminue ensuite de moitié entre la 2ème et la 5ème année (autour de 4,5%). La fréquence se stabilise ensuite jusqu'à la 27ème année.

- Des montants moyens de versements libres qui évoluent quasi linéairement en fonction de l'ancienneté entre 2 et 27 ans.
- Des montants plus importants de versements libres de manière discrétionnaire à partir de 30 ans d'ancienneté, qui peuvent correspondre à des contextes particuliers de transmission de patrimoine (plus le contrat est ancien et plus l'assuré est potentiellement âgé).



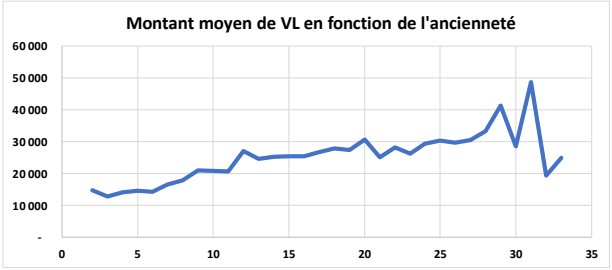


FIGURE 9 : Evolution de la fréquence et du montant moyen en euros des versements libres en euros en fonction de l'ancienneté du contrat (cumul 2010-2017)

La Provision Mathématique (PM) d'ouverture du contrat (« richesse » du contrat en début d'année)

Les quartiles des montants en euros de PM d'ouverture (cumul 2010-2017) se présentent de la manière suivante :

Quartiles du montant de PM d'ouverture	Montant de PM
1er quartile	661
2ème quartile	5 087
3ème quartile	22 071
4ème quartile	20 081 192

TABLEAU 8 : Quartiles des montants en euros de Provisions Mathématiques d'ouverture des contrats de la base de données (cumul 2010-2017)

L'observation de la fréquence et des montants moyens en euros des versements libres en fonction de la PM d'ouverture montre :

- Une fréquence plus importante pour les contrats ayant plus de richesse (3ème et 4ème quartiles) ;
- Des montants moyens de versements libres qui apparaissent significativement plus importants pour le 4ème quartile

Quartiles du montant de PM d'ouverture	Fréquence de VL
1er quartile	4,4%
2ème quartile	4,4%
3ème quartile	6,0%
4ème quartile	6,9%

Quartiles du montant de PM d'ouverture	Montant moyen de VL
1er quartile	19 826
2ème quartile	15 323
3ème quartile	16 370
4ème quartile	29 246

TABLEAU 9 : Fréquences et montants moyens en euros de versements libres par quartile de PM d'ouverture (cumul 2010-2017)

Par ailleurs, l'observation des niveaux de PM d'ouverture en euros en fonction de l'ancienneté et de l'âge montre une augmentation quasi linéaire (notamment sur les premières années). Il est donc indispensable d'analyser le niveau de corrélation et de colinéarité qui peut exister entre ces trois variables (pour maîtriser la variance du modèle, et de fait le risque d'erreur de prédiction)

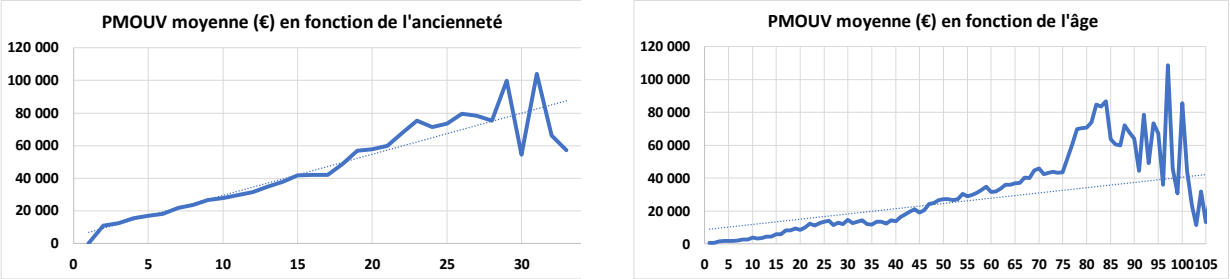


FIGURE 10 : Evolution de la PM d'ouverture en euros en fonction de l'ancienneté et de l'âge (cumul 2010-2017)

Les rachats partiels sur le contrat (qui pourraient refléter une certaine « activité » sur le contrat)

La relation entre la fréquence de rachats partiels et la fréquence de versements libres n'est pas du tout évidente. En effet, seuls environ 9% des contrats ayant fait l'objet d'un moins un versement libre au cours de l'année ont également fait l'objet d'au moins un rachat partiel (pourcentage stable entre 2010 et 2017). Cette variable ne sera par conséquent pas pris en compte dans le cadre de l'étude.

La fiscalité du contrat

Dans le cadre de cette étude, le régime fiscal des contrats concernés par les versements libres est principalement celui de l'assurance vie classique (88% du nombre de versements libres et 89% des montants en euros de versements libres). Le reste de l'effectif et des montants en euros de versements libres se compose d'environ 3% de régime fiscal Plan Epargne Populaire et de moins d'un pourcent de régime fiscal DSK.

Remarque : pour 8% des VL de la base de données, la nature du régime fiscal n'est pas renseignée (qualité des données).

Cette variable n'est pas clairement discriminante et ne sera donc pas pris en compte dans le cadre de l'étude.

Le Taux Minimum Garanti (TMG) du contrat

Dans les conditions générales du produit considéré pour l'étude, il n'existe aucun engagement de taux fixe garanti (au-delà de ce qui est fixé par la réglementation). Cette variable n'a donc pas été retenue.

1.4.4. Versements libres et données macro-économiques

L'objectif de cette partie est d'observer les liens qui peuvent exister entre les versements libres et certaines variables macro-économiques sélectionnées dans le cadre de l'étude.

Année	Montant de VL (€)	Taux de chômage annuel*	IPL annuel*	TME annuel*	CAC 40 au 31/12
2010	136 809 651	8,88	99,90	3,18	3 805
2011	140 168 004	8,83	105,70	3,38	3 160
2012	129 873 606	9,38	105,15	2,58	3 641
2013	140 554 042	9,93	103,13	2,26	4 296
2014	152 326 548	9,93	101,50	1,69	4 273
2015	159 718 231	10,05	99,98	0,88	4 637
2016	136 737 867	9,75	101,03	0,51	4 862
2017	144 503 225	9,10	104,15	0,85	5 313

**Moyenne des taux mensuels sur 12 mois*

TABLEAU 10 : Evolution par année des montants en euros de versements libres et de différents indices macro-économiques sélectionnés dans le cadre de l'étude

L'observation des courbes d'évolution par année des montants en euros de versements libres et des différents indices macroéconomiques met en évidence un lien apparent entre les indices macro-économiques. Il faudra donc analyser avec attention le niveau de corrélation et de colinéarité entre ces variables pour maîtriser la variance du modèle, et de fait le risque d'erreur de prédiction.

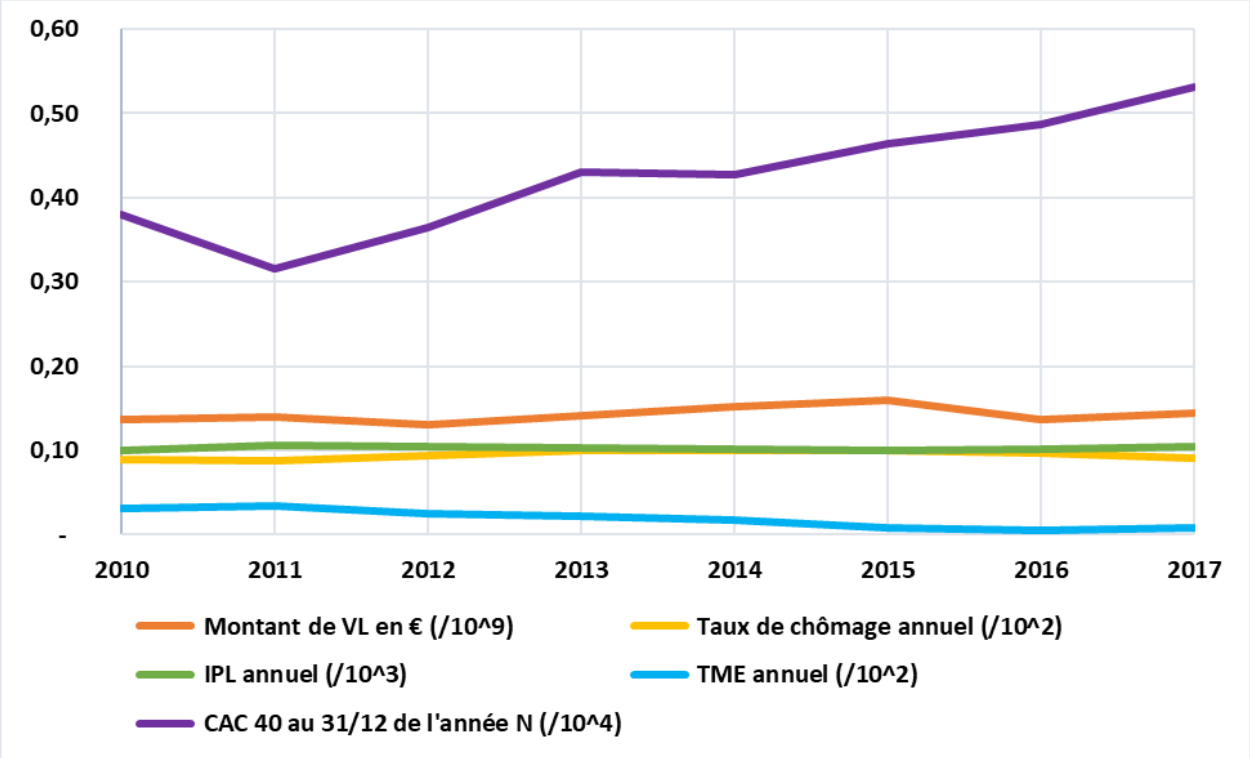


FIGURE 11 : Evolution des montants en euros de versements libres et des indices macro-économique entre 2010 et 2017

1.4.5. Etude de corrélation et de colinéarité

L'objectif de cette partie est d'analyser le niveau de corrélation et de colinéarité qui peut exister entre les différentes variables explicatives retenues dans le cadre de l'analyse descriptive.

Etude de corrélation

Le coefficient de corrélation de Pearson $r(X, Y)$ de deux variables aléatoires X et Y s'écrit de la manière suivante :

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

avec :

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

et :

$$\sigma_X = \sqrt{Cov(X,X)}$$

Ce coefficient mesure l'intensité de liaison linéaire entre les 2 variables X et Y.

Soit une variable booléenne « VL_BOOL » qui vaut "1" en cas de versement libre et "0" en l'absence.

La matrice de corrélation obtenue met en évidence une certaine corrélation entre :

- la variable ancienneté et la variable âge,
- la variable PM d'ouverture du contrat et la variable ancienneté ou la variable âge,
- l'ensemble des variables macro-économiques et plus fortement entre la variable TME et la variable CAC 40 avec un coefficient de corrélation à hauteur de -91,51%.

	Ancienneté	Age	PM d'ouverture du contrat (en euros)	Taux de chômage (annuel)	IPL (annuel)	TME (annuel)	CAC 40 (au 31/12 de l'année N)	VL_BOOL
Ancienneté	100,00%	**49,03%**	**14,24%**	4,71%	-1,02%	-10,56%	10,17%	-4,46%
Age	**49,03%**	100,00%	**14,67%**	-0,67%	0,24%	0,81%	-0,80%	-1,37%
PM d'ouverture du contrat (en euros)	**14,24%**	**14,67%**	100,00%	-0,24%	0,04%	0,27%	-0,27%	1,56%
Taux de chômage (annuel)	4,71%	-0,67%	-0,24%	100,00%	**-46,38%**	**-56,73%**	**39,12%**	0,27%
IPL (annuel)	-1,02%	0,24%	0,04%	**-46,38%**	100,00%	**37,60%**	**-34,06%**	-1,39%
TME (annuel)	-10,56%	0,81%	0,27%	**-56,73%**	**37,60%**	100,00%	**-91,51%**	0,73%
CAC 40 (au 31/12 de l'année N)	-10,17%	-0,80%	-0,27%	**39,12%**	**-34,06%**	**-91,51%**	100,00%	-0,79%
VL_BOOL	-4,46%	-1,37%	1,56%	0,27%	-1,39%	0,73%	-0,79%	100,00%

TABLEAU 11 : Matrice de corrélation Pearson des variables explicatives

D'autres mesures de la corrélation entre 2 variables X et Y existent, comme notamment **le coefficient de Spearman ou le tau de Kendall.**
Il est donc intéressant d'observer si ces niveaux de corrélation se confirment en déterminant les deux autres types de mesure.

Le coefficient de Spearman est un cas particulier du coefficient de Pearson et présente l'intérêt d'être non paramétrique : plus d'hypothèse de normalité.
Le principe est de substituer aux valeurs observées leurs rangs. Deux nouvelles variables sont ainsi créées : $Ri = Rang(xi)$ et $Si = Rang(yi)$.

Le coefficient ρ de Spearman correspond au coefficient de Pearson qui est calculé sur les rangs :

$$\rho = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \overline{S})^2}}$$

Le tau de Kendall est défini pour mesurer l'association entre variables ordinales. Il repose sur la notion de paires discordantes et concordantes :

- On dit que les paires d'observations i et j sont concordantes si et seulement si $(x_i > x_j)$ alors $(y_i > y_j)$ ou $(x_i < x_j)$ alors $(y_i < y_j)$. Nous pouvons simplifier l'écriture avec $(x_i - x_j)x(y_i - y_j)>0$.
- On dit que les paires d'observations i et j sont discordantes lorsque $(x_i > x_j)$ alors $(y_i < y_j)$ ou $(x_i < x_j)$ alors $(y_i > y_j)$, en d'autres termes $(xi - x_j)x(y_i - y_j) < 0$.

Pour un échantillon de taille n, on note P le nombre de paires concordantes et Q le nombre de paires discordantes. Le tau de Kendall τ est défini de la manière suivante :

$$\tau = \frac{P - Q}{\frac{1}{2}n(n-1)}$$

Pour des bases de données de taille importante (supérieure à environ 1 000 lignes, chiffre considéré comme significatif), il est conseillé d'utiliser une fonction de calcul optimisé (notamment avec le logiciel R : la fonction « Fast estimation of Kendall's tau rank correlation coefficient » ou « Cor.fk() » de la librairie « pcaPP »).
Les résultats des trois mesures de corrélation montrent que :

- Le niveau de corrélation entre les variables âge et ancienneté se confirme (à hauteur de 49%) ;
- **Le niveau de corrélation entre la variable « PM d'ouverture » et la variable « âge » ou la variable « ancienneté » observé avec la mesure de Pearson augmente significativement avec la mesure Spearman et la mesure de Kendall ;**
- Le niveau de corrélation entre les variables macro-économiques se confirme notamment la forte anticorrélation qui existe entre la variable TME et la variable CAC 40 (entre -78% et -93%).

Corrélation entre :	Pearson	Spearman	Kendall
- les variables "ancienneté" et "âge"	49,03%	49,11%	34,96%
- les variables "PM d'ouverture" et "ancienneté"	**14,24%**	**46,90%**	**33,92%**
- les variables "PM d'ouverture" et "âge"	**14,67%**	**46,81%**	**32,30%**
- les variables "TME" et "CAC 40"	-91,51%	-92,66%	-78,27%

TABLEAU 12 : Comparatif des mesures de corrélation Pearson, Spearman et Kendall entre certaines variables explicatives

Analyse de la colinéarité

On parle de « *multicolinéarité parfaite* » lorsqu'une des variables explicatives d'un modèle est une combinaison linéaire d'une ou plusieurs autres variables explicatives introduites dans le même modèle. L'absence de « multicolinéarité parfaite » est une des conditions requises pour pouvoir mettre en place un modèle linéaire et, par extension, un modèle linéaire généralisé.

En pratique, la « multicolinéarité » n'est jamais parfaite.

L'approche la plus classique consiste à examiner les *Facteurs d'Inflation de la Variance* (*FIV*) ou *Variance Inflation Factor* (*VIF*) en anglais :

- Les VIF estiment de combien la variance d'un coefficient est « augmentée » en raison d'une relation linéaire qui existe avec d'autres prédicteurs. Ainsi, un VIF de valeur 1,8 indique que la variance de ce coefficient particulier est supérieure de 80% à la variance qui aurait dû être observée si ce facteur n'était absolument pas corrélé aux autres prédicteurs ;
- Si tous les VIF sont égaux à 1, il n'existe pas de « multicolinéarité ».

Pour un modèle linéaire, le VIF peut être calculée à partir de la formule suivante :

$$\text{VIF} = \frac{1}{1\text{-}R^2}$$ avec R^2 qui est le coefficient de détermination (uniquement pour le modèle linéaire)

Dans le cas d'un modèle GLM, le VIF peut être approchée en remplaçant le coefficient de détermination par le Pseudo-R^2 (défini en partie 2.1).

MESURE DE LA COLINEARITE (VIF)	Régression GLM_Logit avec toutes les variables potentiellement explicatives	Régression GLM_Logit avec uniquement ancienneté, âge et Niveau_PM	Régression GLM_Logit avec uniquement ancienneté et PM d'ouverture du contrat	Régression GLM_Logit avec uniquement ancienneté, âge	Régression GLM_Logit avec uniquement ancienneté, âge et TME
Ancienneté	1,38	1,36	1,01	1,27	1,29
Age	1,43	1,43		1,27	1,28
PM d'ouverture du contrat	1,38	1,37	1,01		1,01
TME (annuel)	8,99				
IPL (annuel)	1,27				
Taux de chômage (annuel)	2,00				
CAC 40 au 31/12 de l'année N	7,19				

TABLEAU 13 : Mesure de la colinéarité (VIF) entre les variables explicatives

Les résultats présentés dans le tableau précédent montrent notamment :

- Une très forte colinéarité entre les variables macro-économiques TME et CAC 40 ;
- Une colinéarité plus forte entre les variables âge et ancienneté qu'entre les variables PM d'ouverture et ancienneté.

1.4.6. Liste des variables explicatives retenues pour l'étude

Suite à l'analyse descriptive des données de la base, il a été choisi de ne retenir que les variables explicatives suivantes pour l'étude :

Variables explicatives analysées	Variables explicatives retenues
Variables assuré : - Age - Sexe	- Age
Variables contrat : - Ancienneté - Fiscalité - PM d'ouverture - Nombre de rachats partiels - Taux Minimum Garanti	- Ancienneté - PM d'ouverture
Variables macro-économiques : - TME annuel - IPL annuel - Taux de chômage annuel - CAC 40 au 31/12	- TME annuel - IPL annuel - Taux de chômage annuel - CAC 40 au 31/12

TABLEAU 14 : Variables explicatives retenues pour l'étude

Remarque : même si les variables TME et CAC 40 sont fortement anticorrélées, elles ont été retenues toutes les deux dans le cadre des travaux de classification pour pouvoir comparer leur significativité.

2. Classification et modélisation : introduction des outils mathématiques utilisés

L'objectif de cette partie est de présenter succinctement les différents outils mathématiques utilisés en application dans le cadre des travaux d'analyse et de modélisation de ce mémoire (partie 3.).
Pour cette partie, i, j, k, p et n sont des entiers naturels.

2.1. Le modèle linéaire généralisé (GLM)

Le Modèle Linéaire Généralisé (noté « GLM ») est un outil mathématique fréquemment utilisé par les compagnies d'assurance pour modéliser leurs risques et les comportements des assurés à partir de différentes variables dites « explicatives » (comme par exemple : l'âge de l'assuré). Il s'agit alors de modèles dits « paramétriques » (Bibliographie : 8. et 20.).

Ce modèle a été développé initialement en 1972 par Nelder et Wedderburn et présenté en détails dans les exposés de Nelder et Mc Cullagh (1983), Agresti (1990) ou Antoniadis et al. (1992).

Le modèle GLM permet notamment d'expliquer une variable Y en fonction d'une combinaison linéaire de variables explicatives $X = (X_1, X_2, \ldots, X_p)$.
On considère alors l'expression conditionnelle : $E(Y/X = x)$.

Un point important à signaler est que la relation entre la variable à expliquer Y et les variables explicatives X n'est pas forcément linéaire (Y ne suit pas toujours une loi normale).
L'observation de l'évolution de Y en fonction de X peut mettre en évidence le caractère non-linéaire de la relation qui lie Y avec X. Dans ce cas, l'utilisation d'un modèle linaire simple serait alors imprudente et pourrait conduire à des prédictions erronées.

Les modèles linéaires généralisés sont caractérisés par trois composantes :

- **La composante aléatoire**
- **Le prédicteur linéaire**
- **La fonction de lien**

La composante aléatoire
La composante aléatoire identifie la distribution de probabilités de la variable à expliquer. On suppose que l'échantillon statistique est constitué de n variables aléatoires Yi (avec i allant de 1 à n) indépendantes admettant des distributions issues d'une structure exponentielle.

Cela signifie que les lois de ces variables sont dominées par une même mesure dite de référence et que la famille de leurs densités par rapport à cette mesure se met sous la forme :

$$f_{\theta,\phi}(y) = \exp\left[\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right], \qquad (\bullet)$$

Cette formulation de f inclut la plupart des lois exponentielles usuelles comportant un ou deux paramètres : Gaussienne, Gamma, Poisson, Binomiale, etc.
Le paramètre θ est appelé paramètre naturel de famille exponentielle (appartenant à R), et le paramètre Φ est appelé « paramètre de dispersion » (appartenant à R^*_+) qui sert à ajuster la variance du modèle par rapport à l'observation.
$b(.)$ est une fonction définie sur R deux fois dérivable, $c(.)$ fonction définie R^2.
A titre d'exemple, la densité d'une loi de Poisson satisfait ces critères, elle est donnée par :

$$\begin{aligned} f(y) &= \frac{e^{-\mu}\mu^y}{y!} \\ &= \exp\{y\ln\mu - \mu - \ln(y!)\} \end{aligned}$$

En considérant $\Phi=1$; $\theta=ln(\mu)$; $b(\theta)=\mu$ et $c(y, \Phi) = -ln(y!)$, on retrouve l'expression précédente de f (•).

Le prédicteur linéaire

Les observations planifiées des variables explicatives sont organisées dans la matrice X de planification d'expérience. Soit η un vecteur de p paramètres, le prédicteur linéaire, composante déterministe du modèle, est le vecteur à n composantes :

$$\eta = X\beta = \beta_0 + \sum_{i=1}^{p} X_i\beta_i$$

Où $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ est le vecteur des paramètres / coefficients inconnus à estimer, β_0 est une constante et les variables Xi peuvent être considérées comme les variables quantitatives explicatives du modèle.

La fonction de lien

La troisième composante exprime une relation fonctionnelle entre la composante aléatoire et le prédicteur linéaire.

$g(E[Y]) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, avec p le nombre de variables explicatives

où g, appelée fonction lien, est supposée monotone et différentiable.
Quelques exemples de fonction de lien pour quelques lois usuelles :

Loi	Nom du lien	Fonction de lien
Bernoulli/Binomiale	lien logit	$g(\mu) = \text{logit}\ (\mu) = \log(\mu/(1-\mu))$
Poisson	lien log	$g(\mu) = \log(\mu)$
Normale	lien identité	$g(\mu) = \mu$
Gamma	lien réciproque	$g(\mu) = -1/\mu$

Cas particulier de la régression logistique

La fonction inverse de la fonction lien **Logit** s'écrit sous la forme :

$$g^{-1}(x) = \frac{e^x}{1+e^x}.$$

Pour tout i entier, la loi de probabilité du modèle Logit est définie par :

$$\mathbb{P}[y_i = 1|X_i] = \frac{1}{1+e^{-X_i\beta}}.$$

La décomposition en une composante déterministe et une composante aléatoire permet de s'assurer d'une estimation moyenne même si pour les mêmes valeurs des variables X, la variable Y observée peut différer selon les cas.

Cas particulier de la régression Log-Gamma

La fonction densité de la loi Gamma $\Gamma(\mu,\nu)$ s'écrit sous la forme :

$$f(y) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} exp\left\{-\frac{\nu}{\mu}y\right\}, y \in \mathbb{R}_+$$

Considérant la fonction logarithme comme fonction lien (régression Log-Gamma), l'expression de $E(Yi / Xi)$, pour toute variable Yi à expliquer peut s'écrire :

$$E(Yi / Xi) = \mu_i, \quad avec \quad \mu_i = exp\left\{\beta_0 + \sum_j^p \beta_j \times x_{ij}\right\}.$$

Remarque : pour une régression Log-Normale, la fonction lien est de la même forme, à savoir $E(Yi / Xi) = \exp(\beta Xi)$.

Rappel : une variable aléatoire Y suit une loi Log-Normale $LN(\mu, \sigma^2)$ si et seulement si log(Y) suit une loi normale $N(\mu, \sigma^2)$ avec $E(Y) = \exp(\mu + \sigma^2/2)$ et $Var(Y) = \exp(2\mu + \sigma^2)x(\exp(\sigma^2) - 1)$.

Estimation des paramètres β_j du modèle GLM

L'estimation des paramètres β_j est calculée en maximisant la log-vraisemblance du modèle linéaire généralisé. La vraisemblance en y du modèle s'écrit :

$$\mathcal{L}(y;\theta,\phi) = \prod_{i=1}^{n} f(y_i;\omega_i,\phi)$$

La log-vraisemblance en y du modèle s'écrit alors :

$$\ell(y_i,\theta_i,\phi) = lnf(y_i,\theta_i,\phi) = \frac{y_i\theta_i - \upsilon(\theta_i)}{u(\phi)} + \omega(y_i,\phi)$$

L'estimation par maximum de vraisemblance conduit à résoudre les équations normales suivantes, obtenues en dérivant deux fois la log-vraisemblance par rapport au paramètre naturel :

$$\frac{\partial \ell}{\partial \theta_i} = \frac{y_i - \upsilon'(\theta_i)}{u(\phi)}$$

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = -\frac{\upsilon''(\theta_i)}{u(\phi)}$$

Après quelques calculs, ces équations peuvent se mettre sous la forme :

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i|X=x)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \qquad \forall j = [\![1,p]\!]$$

La résolution de ces équations non-linéaires en β requiert des méthodes numériques itératives dont les plus répandues sont les algorithmes de Newton-Raphson et de Fisher. En pratique, seul l'algorithme de Fisher est utilisé (Bibliographie 34.).

Précision des estimateurs des paramètres du modèle GLM

Intervalle de confiance : pour tout j entier, l'estimation de la variance de l'estimateur de β_j notée $\hat{\sigma}_j^2$ est obtenu à partir de la matrice de matrice d'information de Fisher au point β avec :

$$\frac{(\hat{\beta}_j - \beta_j)^2}{\hat{\sigma}_j^2} \xrightarrow{\mathcal{L}} \chi_1^2 \quad \text{ou encore} \quad \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Un intervalle de confiance (asymptotique) de niveau $1-\alpha$ pour β_j est donc donné par :

$$IC_{1-\alpha}(\beta_j) = \left[\hat{\beta}_j - u_{1-\alpha/2}\hat{\sigma}_j ; \hat{\beta}_j + u_{1-\alpha/2}\hat{\sigma}_j\right]$$

où $u_{1-\alpha/2}$ représente le quantile de niveau $(1-\alpha/2)$ de la loi normale $N(0, 1)$.

Test de nullité ou de significativité des paramètres / coefficients du modèle GLM

Ce test cherche à valider si l'on accepte ou on rejette l'hypothèse nulle H_0 suivante :

$$H_0 : \beta_j = 0 \quad \text{et} \quad \hat{\beta}_j/\hat{\sigma}_j \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Où β_j est le paramètre associé à la variable explicative X_j.

$H0$ sera rejeté si la valeur observée de $\hat{\beta}_j/\hat{\sigma}_j$ dépasse en valeur absolue le quantile d'ordre $1-\alpha/2$ de la loi $N(0, 1)$.

Notion de « valeur p » (ou « p-value » en anglais)

La notation de probabilité type notée p en appelant x le résultat observé et H_0 l'hypothèse nulle définit précédemment, la « p-value » peut s'écrire sous la forme :

$$p = \mathbb{P}(x|H_0).$$

Autrement dit, la « p-value » correspond à la probabilité que l'hypothèse H_0 ne soit pas rejeté et par conséquent que le paramètre β_j associé à la variable explicative X_j soit considéré comme non significatif (l'utilisation de cette variable ne serait alors pas pertinente).

Pour le modèle GLM, **la significativité des variables explicatives** est généralement testée par un des deux différents tests suivants :

- **Le test du rapport de vraisemblance**
- **Le test de Wald**

Le test du rapport de vraisemblance

Ce test utilise la statistique :

$$T = -2ln\left(\frac{Vraisemblance\ sans\ la\ variable\ X_k}{Vraisemblance\ avec\ la\ variable\ X_k}\right)$$

Sous l'hypothèse H_0, T suit une loi $KHI\text{-}2$ à 1 degré de liberté.

La variable X_k est jugée significative si H_0 est rejetée, ce qui se traduit par :

$$P = Prob(\chi^2(1) \geq T) \leq \alpha$$

α représente le seuil de significativité (« marge d'erreur »).

Le test de Wald

Le test de Wald se base sur la statistique suivante :

$$W = \frac{\hat{\beta}_x}{\sigma(\hat{\beta}_x)}$$

Sous l'hypothèse H_0, W suit une loi normale $N(0,1)$.

La variable X_k est jugée significative si H_0 est rejetée, ce qui se traduit par :

$$P = Prob(|N(0,1)| \geq W) \leq \alpha$$

Ce test compare l'écart entre le coefficient théorique et sa valeur estimée.

α représente le seuil de significativité (« marge d'erreur »).

Validation d'un modèle GLM – Déviance

Pour mesurer la qualité de l'ajustement d'un modèle GLM on utilise souvent la « déviance ».

Le modèle estimé est ainsi comparé avec le modèle dit saturé, c'est-à-dire le modèle possédant autant de paramètres que d'observations et par conséquent qui estime « de manière exacte » la valeur à expliquer.

L'expression de la déviance pour une estimation de β est définie par :

$$D(\hat{\beta}) = 2\{\log \mathcal{L}^{\text{sat}} - \log \mathcal{L}(\hat{\beta})\}$$

Plus D est faible et plus le modèle est de « qualité ».

Cet indicateur est assez « macro » et est en pratique complété par une analyse précise des résidus (du modèle de prédiction). On montre qu'asymptotiquement, D suit une loi du KHI-2 à n-p degrés de liberté, ce qui permet de construire un test de rejet ou d'acceptation du modèle selon que la déviance est jugée significative ou non significative. A partir de la mesure de la déviance, il est possible de calculer le Pseudo-R2 introduit par Mc Fadden (1973) qui est analogue au coefficient de détermination du modèle linéaire (mesure du pouvoir prédictif du modèle). Si le Pseudo-R2 est faible alors le modèle est très peu prédictif.

$$\text{Pseudo-}R^2 = \frac{1 - \text{déviance}}{\text{déviance nulle}}$$

où la déviance nulle est la déviance d'un modèle dit « parfait » (par exemple : dans la cas d'un modèle de régression logistique, le modèle « parfait » prédirait toujours la valeur 1).

Choix d'un modèle GLM – Critères *AIC* et *BIC*

La comparaison entre deux modèles nécessite de tenir compte de la complexité de chaque modèle. Les critères AIC et BIC pénalisent la log-vraisemblance du modèle avec le nombre de paramètres :

- **Le critère d'*AIC* (« Akaïke Informative Criterion »)** pour un modèle à k paramètres est défini par :

  $$AIC = 2k - 2\ln(L)$$

  où L est la vraisemblance du modèle.
- **Le critère de *BIC* (« Bayesian Informative Criterion »)** pour un modèle à k paramètres estimés sur n observations est défini par :

  $$BIC = -2\ln(L) + \ln(n)k$$

Pour chaque modèle concurrent, le modèle qui présente l'AIC ou le BIC le plus faible est sélectionné.
A noter : par construction le critère AIC est le plus adapté pour les modèles complexes ; en effet, le critère BIC pénalise davantage les modèles construits à partir d'une grande quantité de données observées.

A partir des critères AIC et BIC, il a été défini des processus de classification de variables explicatives à prendre en compte dans un modèle de prédiction :
Il s'agit d'une approche itérative (« step by step ») :

- Soit en démarrant avec un modèle utilisant uniquement la constante et en ajoutant chaque variable une par une en vue d'obtenir le critère AIC/BIC le plus faible : **la méthode ascendante (« forward stepwise selection »)** ;
- Soit en démarrant avec un modèle utilisant l'ensemble des variables disponibles et en retirant chaque variable une par une en vue d'obtenir le critère AIC/BIC le plus faible : **la méthode descendante (« backward stepwise selection »).**

Cette approche permet donc de classifier les variables explicatives à prendre en compte dans le cadre d'un modèle de prédiction.

2.2. Le modèle additif généralisé (GAM, une extension du modèle GLM)

Le modèle additif généralisé (« Generalized Additive Models » en anglais – GAM) a été introduit par Trevor Hastie et Rob Tibshirani en 1990.
Le modèle GAM est une extension du modèle GLM et peut dans certains cas se révéler plus performant notamment lorsque la relation entre la variable explicative et la variable à expliquer est très loin d'être linéaire. Ce qui les différencie c'est le prédicteur. Ce dernier est linéaire dans le modèle GLM, alors qu'il est dit additif dans le modèle GAM.
Le prédicteur est composé d'une somme de fonctions qui ne sont pas forcément paramétriques.
Le prédicteur d'un modèle GAM sera de la forme :

$$\eta_i = \alpha + \sum_{j=1}^{k} f_j(X_{ij})$$

Avec k un entier naturel.

Rappel : Le prédicteur d'un modèle GLM est de la forme

$$\eta = X\beta = \beta_0 + \sum_{i=1}^{p} X_i \beta_i$$

où les Xi sont les variables explicatives.
Les fonctions f_j sont des fonctions qui peuvent être paramétriques (fonctions polynomiales, trigonométriques…) ou non paramétriques (fonctions de lissage dites « splines »).
Remarque : le modèle GLM est un cas particulier des modèles GAM avec des fonctions $f_j(x) = \beta_j x$.

Ces fonctions non paramétriques dites « splines » représentent une façon de traiter la relation non linéaire qui peut exister entre certaines variables explicatives et la variable à expliquer.
<u>La principale difficulté réside dans leur détermination.</u>

L'estimation d'une fonction non paramétrique de n variables Xi (X_1, X_2, ..., X_n) peut être approchée par une combinaison linéaire de fonctions b_i non paramétriques qui peut s'écrire sous la forme :

$$f(x) = \sum_{i=1}^{q} b_i(x)\beta_i.$$

Avec β_i les paramètres du modèle.

Une des fonctions non paramétriques la plus fréquemment utilisée dans le cadre d'un modèle GAM est la fonction « spline de lissage cubique ».

La fonction spline cubique : de l'interpolation au lissage (Bibliographie 11.)
La fonction spline d'interpolation : il s'agit d'une fonction définie par morceaux par des polynômes passant par tous les points d'interpolation.

Définition d'une fonction spline
Le mot anglais « spline » désigne une latte flexible utilisée par les dessinateurs pour matérialiser des lignes à courbure variable et passant par des points fixés à "proximité" de ceux-ci. Le tracé ainsi réalisé minimise l'énergie de déformation de la latte.
Par analogie, ce mot désigne également des familles de fonctions d'interpolation ou de lissage présentant des propriétés « optimales » de régularité.
L'idée originale est attribuée à Whittaker (1923), puis reformulée par Schoenberg (1964), Atteia (1965) et Reinsch (1967). Elle connaît ses premières applications en Statistique avec Kimeldorf et Wahba (1970). Une liste complète des articles parus jusqu'en 1973 sur les splines a été publiée par Van Rooij et Schurer (1974) et Wegman et Wright (1983) en résument les applications statistiques.
Une fonction spline $S : [a, b] \in R \rightarrow R$ est une fonction polynomiale par morceaux définie sur un intervalle $[a;b]$ subdivisé en partition $[xi, xi+1]$ tels que :

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$$

Pour tout polynôme $Pi : [xi, xi+1] \rightarrow R$, la fonction spline à n intervalles s'écrit sur chaque intervalle $[xi, xi+1]$:

$$\begin{aligned} S(x) &= P_1(x), x_0 \leq x < x_1 \\ S(x) &= P_2(x), x_1 \leq x < x_2 \\ &\vdots \\ S(x) &= P_n(x), x_{n-1} \leq x < x_n \end{aligned}$$

Le degré de la spline est défini comme le maximum des degrés des polynômes P_i (si tous les polynômes ont le même degré, la spline est dite « uniforme »).

La continuité de la spline définit les caractéristiques de la jonction entre chaque intervalle.
Sachant que la dérivabilité d'un polynôme est infinie, la dérivabilité d'une spline dépend de la continuité au niveau de la jointure des fonctions polynômiales.

La spline linéaire et la spline cubique
Si pour tout i et j, tels que $0 < i < k$ et $0 < j < n$ l'égalité suivante : $P^{(j)}{}_i(xi) = P^{(j)}{}_{i+1}(xi)$ est vérifiée, alors la spline est de continuité n, notée S_n.

Nous regardons les cas particuliers $n = 0$ et $n = 2$:
S_0 est la spline de continuité minimum telle que $P^0_i(x_i) = P^0_{i+1}(x_i)$: Il s'agit de la spline d'interpolation linéaire.
Graphiquement, ce sont des segments qui relieront les points $(x_i ; y_i = S(x_i))$:

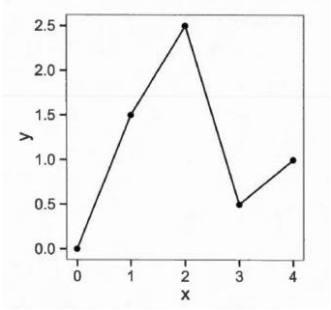


FIGURE 12 : Représentation graphique des segments reliant les points d'observation

S_2 est la continuité avec courbure telle que $P^2_i(x_i) = P^2_{i+1}(x_i)$: les polynômes successifs ont des dérivées secondes égales aux points jonction.

Graphiquement, ce sont des courbes qui relieront les points $(x_i ; y_i = S(x_i))$:

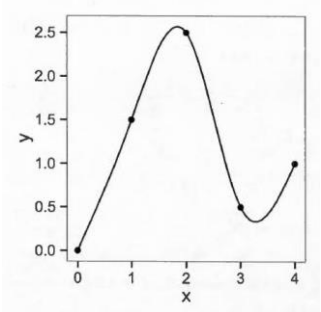


FIGURE 13 : Représentation graphique des courbes reliant les points d'observation

Dans le cas où $n = 2$, la spline est dite « cubique d'interpolation ». Elle est uniforme et définie par des polynômes de degré 3 s'écrivant sous la forme $P(x) = a+bx+cx2+dx3$ et nécessitant quatre contraintes $(a, b, c \text{ et } d)$ pour être défini.
Une spline cubique est dite « naturelle » si et seulement si $S''(a) = S''(b) = 0$ et « périodique » si et seulement si $S''(a) = S''(b)$ et $S'(a) = S'(b)$.

La fonction spline cubique de lissage
Dans le cadre d'un modèle GAM, l'utilisation d'une fonction spline cubique de lissage est privilégiée à la fonction spline cubique d'interpolation.
En effet, une fonction de lissage va « passer au travers des points » avec une courbe de tendance générale qui minimise l'impact des points isolés (variations très locales).
Ce qui est cohérent dans le cadre d'une construction de loi pour un modèle GAM.

L'objectif est double : approcher au plus près des points de la base de données tout en priorisant la tendance générale.

En considérant la fonction de lissage est g, l'approche de lissage proposée est la suivante :
Pour le premier objectif : minimiser la somme des carrés des différences entre les points observés et la fonction g qui est la fonction spline de lissage (définie pour tout $x_j \in R \rightarrow R$ avec j entier naturel).

Pour le deuxième objectif : instaurer un second terme de « pénalité pour irrégularité de variation » à minimiser pour s'assurer d'une tendance générale (avec λ paramètre de lissage à définir pour une pénalité fixée).

$$\sum_{j=1}^{n} \{y_j - g(x_j)\}^2 + \lambda \cdot \text{pénalité}$$

A noter : si $g(x_j) = y_j$, alors la fonction g est la fonction spline d'interpolation.
D'autres fonctions splines de lissage peuvent être utilisées dans le cadre d'un modèle GAM comme par exemple **la fonction spline de lissage « à plaques minces »** (Bibliographie 13., 25. et 33.) introduite inégalement par Duchon (1976), puis formalisée notamment par Meinguet (1979) et Wahba (1990). Cette dernière est obtenue, tout comme la fonction spline de lissage cubique, en approchant au plus près des points de la base de données tout en priorisant la tendance générale. La principale différence réside dans le fait que la fonction spline de lissage « à plaques minces » se caractérise par une surface lissée entre les points de l'échantillon observé alors que la spline de lissage cubique se caractérise par une courbe lissée entre ces mêmes points.

Une fonction spline de lissage « à plaques minces » se construit en minimisant le critère suivant qui combine une mesure de pénalité pour l'ajustement (la somme des résidus au carré) et une mesure pour le lissage :

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - f(t_i))^2 + \lambda J_m^d(f)$$

Où

- i est le nombre de données de la base (allant de 1 à n),
- z_i représente la $i^{\text{ème}}$ donnée,
- $t_i = (x_1(i), \ldots, x_d(i))$ est la base de données à d dimensions,
- f est la fonction spline d'interpolation,
- J_m^d est la fonction « pénalité de lissage » dérivables m fois dans d dimensions,
- λ est un paramètre de lissage.

En faisant varier le paramètre de lissage λ, on peut faire évoluer le niveau d'approximation au regard du niveau de précision attendu. Comme pour la spline de lissage cubique présenté précédemment, l'idée est d'identifier le meilleur compromis entre la vision lissée (tendance) et la vision « point par point » (plus précise).
Si $\lambda = 0$, alors il s'agit uniquement d'une interpolation (aucun lissage).

En utilisant le « Laplacien », la fonction « pénalité de lissage » peut s'écrire de la manière suivante :

$$J_m^d(f) = \sum_{\alpha_1+\ldots+\alpha_d=m} \frac{m!}{\alpha_1!\ldots\alpha_d!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\partial^m f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}\right)^2 \prod_j dx_j$$

où m correspond au nombre de dérivées continues désirées.

A titre d'exemple, pour $d = 3$ et $m = 2$, la fonction « pénalité de lissage » s'écrit :

$$J_2(f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(f_{x_1x_1}^2 + f_{x_2x_2}^2 + f_{x_3x_3}^2 + 2[f_{x_1x_2}^2 + f_{x_2x_3}^2 + f_{x_3x_1}^2])dx_1dx_2dx_3$$

Dans le cas général, en considérant l'expression suivante :

$$\langle f, g\rangle = \sum_{\alpha_1+\ldots+\alpha_d=m} \frac{m!}{\alpha_1!\ldots\alpha_d!}\mathbf{I}$$

Où $\mathbf{I}$ correspond à l'expression suivante :

$$\mathbf{I} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\frac{\partial^m f}{\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d}}\right)\left(\frac{\partial^m g}{\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d}}\right)\prod_j dx_j$$

En utilisant la méthode d'intégration par parties :

$$\langle f, g\rangle = (-1)^m\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f\cdot\Delta^m g + c_\infty$$

Où c_∞ correspond aux valeurs limites et *f* et *g* sont dérivables m fois appartenant à un espace de fonctions qui s'assimile à un espace de Hilbert à noyau si est seulement si $2m - d > 0$.

Le « vecteur nul » / « espace nul » M de la fonction « pénalité de lissage » peut s'écrire :

$$M = \binom{d+m-1}{d}$$

M correspond donc à un espace dimensionnel, décrit par un polynôme de degré inférieur ou égal à $m - 1$ avec d variables.
Si $t_1, \ldots, t_n$ sont choisis de telle manière que la régression des moindres carrés sur M notés $\Phi_1, \ldots, \Phi_M$ est unique alors le critère de minimisation a une solution unique *f*, donnée par :

$$f_\lambda(t) = \sum_{i=1}^{M} d_i\phi_i(t) + \sum_{j=1}^{n} c_j E_m(t, t_j)$$

Où

- t est une variable à d dimensions,
- d_i (i allant de 1 à M) et c_j (j allant de 1 à n) sont des constantes.

E_m est une fonction de Green (solution d'une équation aux m dérivées partielles du Laplacien), qui peut s'écrire de la façon suivante :

$$E_m(t,s) = \begin{cases} \alpha_{m,d}|t-s|^{2m-d}\ln|t-s|^2, & \text{Si } 2m - d \text{ est pair.} \\ \beta_{m,d}|t-s|^{2m-d}, & \text{Sinon.} \end{cases}$$

Avec

$$\alpha_{m,d} = \frac{(-1)^{d/2+1+m}}{2^{2m}\pi^{d/2}(m-1)!(m-d/2)!}$$

Et

$$\beta_{m,d} = \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!}$$

Remarque : au vu des expressions formulées précédemment, il est important de considérer un degré de dérivation suffisamment important pour s'assurer de la performance de la fonction de lissage (pour éviter de trop « approximer »).

Détermination des splines (grâce à des logiciels de statistiques)
Tout comme les autres paramètres des modèles GLM ou GAM, les splines peuvent **être** déterminés grâce à des logiciels de statistiques (par exemple, le logiciel R) en fixant en amont notamment un certain degré de libertés afin d'améliorer la précision du calcul.
Les résultats obtenus en sortie du logiciel peuvent s'écrire $S_k(X)$ avec X la variable explicative considérée, et k le degré de liberté.

Le choix du degré de liberté k doit être défini de manière raisonnée (« jugement d'expert ») :

- k doit être assez grand pour suffisamment prendre en compte la complexité de la distribution des points de l'échantillon observé ;
- k ne doit pas être trop grand pour éviter le surajustement aux points de l'échantillon observé (et optimiser le temps de calcul).

La pertinence de l'utilisation de chaque $S_k(X)$ obtenus en sortie du logiciel pourra être analysée à partir des résultats du « backtesting » en sortie du modèle GAM (analyse des écarts entre les valeurs estimées par le modèle et les valeurs observées sur l'historique de données disponibles).

2.3. Le Machine Learning

2.3.1. Le principe du Machine Learning

Le « Machine Learning » ou "Apprentissage Automatique" en français est un ensemble de méthodes (ou algorithmes) qui utilisent la capacité de calcul et de mémoire des ordinateurs pour apprendre des données explorées, permettant ainsi d'orienter une prise de décision. Les algorithmes accumulent de la connaissance et de l'intelligence à partir de l'exploration de l'historique des données.
L'approche « Machine Learning » est donc opérationnelle et diffère ainsi de celle de la statistique traditionnelle qui utilise des objets abstraits (comme une loi de probabilité par exemple).
Historiquement, cette théorie a été développé avec les travaux des mathématiciens Vapnik et Chervonenkis (1960).

On regroupe les algorithmes de Machine Learning en trois grandes classes, qui correspondent à différents types d'apprentissage :

- **L'apprentissage supervisé :** dans ce cas, l'utilisateur de l'algorithme définit en amont des résultats attendus en sortie (« standardisés » / « labelisés ») correspondant à un certain type de données en entrée. L'algorithme doit déterminer la loi qui permet de trouver le résultat en sortie en fonction de la nature des données d'entrée.
  Exemple d'utilisation : classification automatique des messages indésirables dans les boîtes emails (« Spams »).
  Exemples de modèle : régression linéaire, régression logistique.
- **L'apprentissage non-supervisé :** ici, aucun label n'est associé à l'algorithme. Ce dernier doit découvrir sans assistance humaine la structure « caractéristique » des données en entrée.
  Exemple d'utilisation : la classification d'information dans le cadre de la reconnaissance d'images.
  Exemples de modèle :
  - **La Classification Ascendante Hiérarchique (CAH).** L'objectif est de créer une arborescence permettant :
    - la mise en évidence de lien hiérarchique entre individus et groupes d'individus,
    - la détection d'un nombre de classe au sein d'une population.

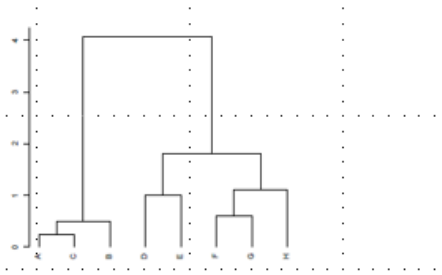


FIGURE 14 : Représentation graphique des résultats d'une CAH

La ressemblance entre deux individus peut être définie par exemple par la distance euclidienne qui les sépare.
Une classification peut être considérée comme proche si et seulement si :
- les individus d'une même classe sont proches (variabilité intra-classe faible),
- les individus de deux classes différentes sont éloignés (variabilité inter-classe importante).

  - **La classification par moyenne mobile (« K-Means ») :** algorithme d'agrégation des individus d'une population autour de centres mobiles (calcul des centres de gravités de chaque classe d'individus).
- **L'apprentissage par renforcement :** approche intermédiaire entre l'apprentissage supervisé et non supervisé.

En 1984, Leo Breiman, Jerome Friedman, Richard Olshen et Charles Stones introduisent les bases théoriques de l'apprentissage moderne des arbres de décision, utilisant des techniques innovantes et des algorithmes avancés permettant de traiter de grandes quantités de données (Bibliographie : 19.).
En 2001, Leo Breiman et Adele Cutler introduisent les **« forêts aléatoires »** (ou « Random forests » en anglais), un algorithme de Machine Learning aujourd'hui très populaire et très puissant.
L'algorithme de Machine Learning **« Gradient Boosting »** est également très populaire aujourd'hui et est issu des travaux de recherche de Breiman, Friedman, Mason, Baxter, Bartlett et Frean (1999).

Il existe d'autres algorithmes de Machine Learning très populaires, comme par exemple les réseaux de neurones. Un neurone est un objet mathématique qui fut conçu à l'origine pour modéliser le fonctionnement du cerveau humain dans le cadre d'études de la cognition. L'algorithme de chaque neurone donne une

réponse réelle entre 0 et 1 (via une fonction mathématique dite « fonction d'activation ») qui diffère selon la donnée en entrée qu'on lui fournit. La principale difficulté réside dans l'interprétation du résultat global avec le modèle de réseau de neurones et notamment des liens complexes (« pistes d'audit ») qui peuvent exister entre certaines variables explicatives du modèle et la variable à expliquer (les résultats en sortie du modèle sont difficilement interprétables).

2.3.2. Les arbres de régression

Le principe d'arbre de décision

Le principe d'arbre de décision repose sur les méthodes dites de partitionnement récursif ou de segmentation formalisées sous l'acronyme de CART : Classification and Regression Tree, algorithme développé par Breiman, Friedman, Olshen et Stone (1984).

Cet algorithme se présente sous la forme classification ou sous la forme de régression selon le type de variable à expliquer (respectivement discrète ou continue). L'intérêt de cet algorithme est que les liens entre les variables explicatives et la variable à expliquer peuvent-être facilement retrouvés (« pistes d'audit »), puisque les prédictions obtenues sont présentées sous une forme de « branches d'arbre ».

A titre d'exemple : le lancement d'une pièce de monnaie à trois reprises. Pour obtenir Pile au troisième coup, il y a quatre situations possibles (« branches d'arbre) : (Pile, Pile, Pile) ; (Pile, Face, Pile) ; (Face ; Pile ; Pile) et (Face, Face, Pile).

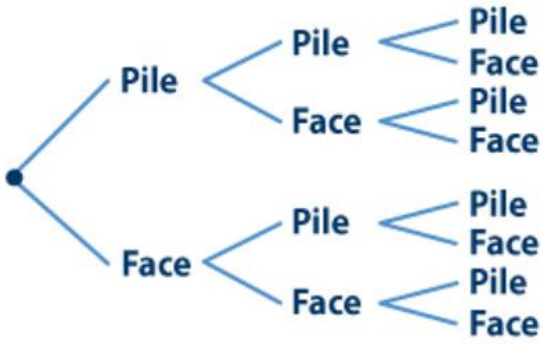


FIGURE 15 : Exemple d'arborescence (lancement d'une pièce de monnaie)

Un arbre de décision (Bibliographie : 3. et 15.) se construit à l'aide d'une séquence récursive de règles de division exécutées pour chaque nœud de l'arbre en partant de la racine (nœud initial), qui regroupe tous les individus de la base, jusqu'à ce qu'il ne reste qu'un individu par branche ou qu'un critère d'arrêt prédéfinie soit atteint.

À chaque étape, une variable aléatoire explicative sur laquelle se fait la segmentation et un critère de segmentation sont désignés pour optimiser l'homogénéité des deux régions issues de la division (les deux nouvelles régions sont disjointes).

Une fois l'arbre de décisions créé, K régions ou nœuds $R_1, R_2, \ldots, R_K$ sont obtenus. La fonction de régression peut s'écrire sous la forme :

$$f(x) = \sum_{j=1}^{K} c_j \mathbb{1}_{R_j}(x)$$

Où les c_j $(j = 1, \ldots, K)$ sont des constantes à estimer qui présentent la valeur attribuée par l'algorithme à un individu appartenant au nœud R_j.

La minimisation de la somme des carrés des erreurs (erreur quadratique) permet d'obtenir un estimateur de la constante c :

$$\min_{c} \sum_{i=1}^{n} (y_i - f(x))^2$$

La valeur optimale de l'estimateur c_j est la moyenne des y_i dans le nœud R_j :

$$\forall j \in [\![1\ ;\ K]\!] \quad c_j = \frac{1}{N_j} \sum_{x_i \in R_j} y_i$$

Avec N_j le cardinal du noeud R_j.

En raison du temps de calcul important pour trouver la meilleure séparation binaire du point de vue de l'erreur quadratique, il est utilisé en pratique un critère de division afin de minimiser pour chaque nœud parent, l'hétérogénéité des deux nœuds fils (récupération de la division optimale).

Le critère de division

Le critère de division choisi pour un nœud donné se base sur la fonction d'hétérogénéité. Dans le cas de la régression, l'hétérogénéité d'un nœud R_j s'écrit sous la forme :

$$D_{R_j} = \frac{1}{N_j} \sum_{i \in R_j} (y_i - \hat{c}_j)^2$$

Une division est définie par un couple de découpage (X_j, s) avec :

- X_j la variable aléatoire utilisée pour le découpage,
- s représente le critère de séparation du nœud.

L'objectif est de sélectionner le couple de découpage optimal qui maximise l'hétérogénéité des deux nœuds fils :

$$\min_{c_1} \sum_{x_i \in R_1} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2} (y_i - c_2)^2$$

Les deux constantes c_1 et c_2 qui minimisent les deux quantités ci-dessus sont :

$$\hat{c}_1 = \frac{1}{N_1} \sum_{x_i \in R_1} y_i \qquad\qquad \hat{c}_2 = \frac{1}{N_2} \sum_{x_i \in R_2} y_i$$

Avec N_1 et N_2, respectivement les cardinaux des deux nœuds R_1 et R_2.

Ce processus de segmentation est répété sur les deux nœuds obtenus jusqu'à ce qu'un critère d'arrêt soit atteint. L'enjeu est d'avoir une profondeur suffisamment importante pour obtenir un arbre de décision de qualité tout en évitant le phénomène de surapprentissage. Pour faire face à cet enjeu, le principe d'élagage est utilisé.

L'élagage

Dans ce cas, l'algorithme CART utilise une base de validation, différente de la base d'apprentissage.
L'élagage s'effectue en retirant les nœuds de l'arbre qui n'apporteraient pas grand-chose à la prédiction. Pour sélectionner l'arbre optimal, on définit un critère nommé coût de la complexité C afin de vérifier que l'arbre élagué T capte bien les schémas de décisions mais pas le bruit.

Pour l'algorithme CART, ce critère s'exprime à partir de l'erreur de classification $E(T)$ à laquelle on ajoute une pénalisation pour la taille de l'arbre (en considérant $|L(T)|$, le nombre de nœuds dans l'arbre).
Pour tout $\alpha \geq 0$, $C(T)$ s'écrit sous la forme :

$$C(T) = E(T) + \alpha |L(T)|$$

α est un paramètre de pénalisation qui représente le « compromis » entre la taille de l'arbre et l'adéquation de l'ajustement aux données. Pour $\alpha = 0$, la solution optimale est l'arbre initial comprenant le maximum de nœuds.

L'indice de Gini (pour le classement des variables explicatives)
Le critère de Gini, également appelé « critère d'impureté », permet d'évaluer la subdivision optimale de chaque arborescence en vue d'atteindre la variable cible.
Il est défini comme suit :

$$I(t) = \sum_{k \neq l} p(k|t)p(l|t) = 1 - \sum_{k=1}^{K} p(k|t)^2$$

où $p(k|t)$ désigne la proportion d'éléments de la classe k affectée au nœud t et K le nombre de classes de la variable cible. Un nœud est complètement pur s'il ne contient que des observations relatives à une seule classe de la variable cible. Dans ce cas $I(t) = 0$.
L'importance d'une variable est alors évaluée en sommant, dans chaque arbre, la baisse d'impureté entraînée par une coupure selon cette variable et en moyennant cette valeur sur la totalité des arbres (Bibliographie 3.).

2.3.3. L'algorithme des « forêts aléatoires »

L'algorithme des forêts aléatoires
Cette méthode des forêts aléatoires a été développée par Breiman (2001) et vise à rendre les arbres de l'agrégation plus indépendants en rajoutant de l'aléa dans le choix des variables explicatives intervenant dans les différents sous-modèles.
Le but est de rendre les arbres décorrélés les uns des autres au cours de l'agrégation en ajoutant du hasard dans le choix des variables explicatives intervenant dans les modèles.

L'algorithme des forêts aléatoires se base principalement sur deux principes :

- Le principe d'arbre de décision (CART) avec lequel chaque variable est tirée aléatoirement à chaque nœud de l'arbre ;
- Le principe du **« Bagging »** pour calculer la moyenne des prévisions de modèles indépendants qui vise à minimiser la variance globale et donc l'erreur de prédiction.

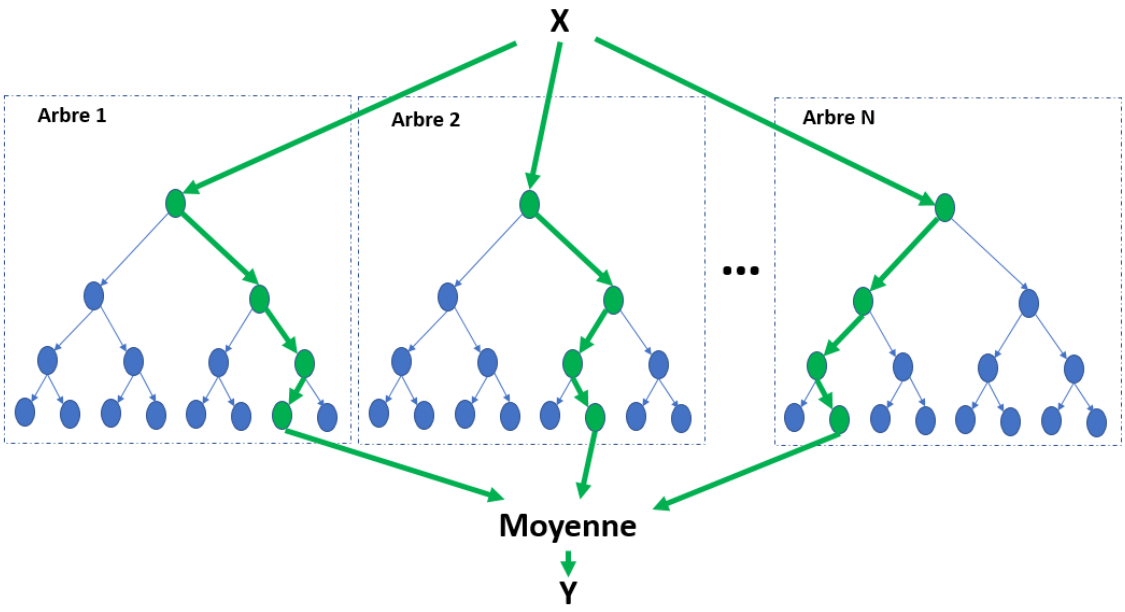


FIGURE 16 : Schéma illustrant le principe de l'algorithme des forêts aléatoires

Le Bagging

L'algorithme du Bagging (Bootstrap aggregating) introduit par Breiman en 1994 (Bibliographie : 3. et 15.) permet de construire un grand nombre d'arbre à partir de plusieurs échantillons de données qui sont choisies aléatoirement (appelés « Bootstraps »).

Il s'agit d'un tirage équiprobable avec remise des observations de la base pour créer un nouvel échantillon de la même taille que l'échantillon initial. Le Boostrap est une manière de reproduire artificiellement plusieurs bases d'apprentissage différentes pour s'assurer de l'indépendance des estimateurs construits.

De plus, à chaque nœud, un sous-ensemble de variables explicatives est sélectionné aléatoirement parmi l'ensemble des variables disponibles afin de diminuer la corrélation entre les arbres.

Voici les principales étapes de la construction d'une forêt aléatoire :

Soient b et B, deux entiers naturels.

En considérant un nombre b = 1, …, B de Bootstraps (B est ne nombre total de Bootstraps tirés).

Pour b = 1 jusqu'à B :

1. Sélectionner aléatoirement un échantillon d'observations parmi l'ensemble des données disponibles (base initiale).
2. Créer un arbre de décision aléatoire $\mathcal{T}_b$ à partir de l'échantillon sélectionné (Bootstrap) et d'un sous-ensemble de variables explicatives tirées au hasard.

Une forêt d'arbre est ainsi créée et son estimateur s'écrit de la façon suivante :

$$\hat{f}_B(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{\mathcal{T}}_b(x)$$

avec $\hat{\mathcal{T}}_b$ l'estimateur de l'arbre $\mathcal{T}_b$.

Cette méthode d'agrégation (Bagging) permet de réduire la variance du modèle. En considérant σ^2 la variance des arbres créées et ρ leur corrélation deux à deux, l'expression suivante montre que plus le nombre total d'arbre B est grand et plus la variance diminue :

$$\begin{aligned} Var(\hat{f}_B(x)) &= Var\Big(\frac{1}{B}\sum_{b=1}^{B}\hat{\mathcal{T}}_b(x)\Big) \\ &= \frac{1}{B^2}\sum_{b=1}^{B}\Big(\sigma^2 + \sum_{b'=1,b'\neq b}^{B}\rho\sigma^2\Big) \\ &= \frac{\sigma^2}{B^2}\sum_{b=1}^{B}(1+(B-1)\rho) \\ &= \rho\sigma^2 + \frac{(1-\rho)\sigma}{B} \end{aligned}$$

L'erreur Out Of Bag (OOB)

L'OOB est une mesure de l'erreur de prédiction de l'algorithme des forêts aléatoires. En effet, chaque arbre est construit en considérant une fraction de données (« Bag » ou « Bootstrap »), les prédictions hors de cette fraction seront erronées. L'objectif est par conséquent d'obtenir l'OOB le plus petit possible.

Les principaux avantages de l'algorithme des forêts aléatoires :

- Il est très adapté à des problématiques où le nombre de variables explicatives est très important ;
- Il permet de mettre clairement en évidence l'importance des liens qui existent entre les variables explicatives et la variable à expliquer. Il permet de fournir une assez bonne première vision de l'importance de chaque variable explicative dans la détermination de la variable à expliquer (classification des variables explicatives).

Les principales limites :

- La lenteur d'exécution de l'algorithme (apprentissage lent, donc peu adapté lorsque le nombre d'observations est important) ;
- La prédiction des variables extrêmes peut être souvent instable.

2.3.4. L'algorithme « eXtreme Gradient Boosting » (XGBoost)

Actuellement, l'XGBoost est la version la plus performante de l'algorithme « Gradient Boosting » (notamment en termes de précision et d'optimisation du temps de calcul).

L'algorithme « Gradient Boosting »

Cet algorithme calcule des prédictions en combinant les résultats de plusieurs CART. Il diffère des forêts aléatoires puisque l'« arbre de régression » ne résulte pas d'une moyenne d'arborescences de prédiction, mais d'une optimisation « pas à pas » vers un arbre de régression cible « optimal » qui présente la variance minimum et donc minimise au maximum l'erreur de prédiction (Bibliographie 19.).

Il s'agit de combiner l'information de plusieurs arbres (appelés également « classifieurs ») pour obtenir un modèle de prédiction meilleur. Cette agrégation est faite itérativement de telle sorte que chaque nouvel arbre créé améliore le modèle agrégé.

L'idée générale consiste à calculer une série d'arbres de décision simples, où chaque arbre consécutif est construit pour prévoir les résidus de la prévision de l'arbre précédent.
En généralisant, les résidus peuvent être estimés par le gradient négatif de l'erreur des moindres carrés $0{,}5x(y-g(x))^2$ où les y_i correspondent aux valeurs observées et les $g(x_i)$ aux valeurs prédites.
La notion de gradient intervient puisque pour rappel, l'algorithme de descente du gradient désigne un algorithme d'optimisation différentiable destiné à minimiser une fonction.
Il est itératif et procède par des améliorations successives. On se donne un point initial $x_0 \in E$ (où E est un espace Hilbertien) et un seuil de tolérance $\varepsilon \geq 0$. L'algorithme définit une suite de points $x_1, x_2, \ldots, x_n \in E$ (avec n entier naturel) jusqu'à ce qu'un test d'arrêt soit satisfait.
En introduisant les classifieurs f dits « faibles », le classifieur optimal est obtenu itérativement par l'algorithme en passant de x_i à x_{i+1} avec les actions suivantes :

Etapes	Actions de l'algorithme
Etape 1	Calcul du gradient de f en x : $\nabla f(x)$
Etape 2	Test d'arrêt : si $\|\|\nabla f(x)\|\| \leq \varepsilon$, alors arrêt
Etape 3	Calcul du pas ηi par une règle de recherche linéaire sur f en x le long de la direction $-\nabla f(x)$
Etape 4	Calcul de la nouvelle itération : $xi+1=xi-\eta i*\nabla f(xi)$

TABLEAU 15 : Etapes clés de l'exécution de l'algorithme Gradient Boosting

L'algorithme XGBoost
L'algorithme XGBoost repose sur l'approche originale du Gradient Boosting. Il a été initialement introduit par Tianqi Chen et Carlos Guestrin (2014) avant la contribution de nombreux développeurs.

Principe de l'algorithme XGBoost
Une **fonction objectif** Obj **est définie en complétant la fonction perte** L du Gradient Boosting classique (erreur quadratique moyenne) qui est convexe et différentiable au minimum deux fois dans le cas de l'algorithme XGBoost **par un terme de régularisation** K.

$$Obj(\cdot) = L(\cdot) + K(\cdot)$$

Cette régularisation limite l'ajustement de l'arbre ajouté à chaque étape de l'exécution de l'algorithme et contribue à éviter un surajustement du modèle.
En effet, lorsque des observations atypiques sont prises en compte par l'algorithme, augmenter le nombre d'itérations peut provoquer une dégradation de la performance.

Soit $\mathcal{T}_b$ un arbre construit à l'étape b et fb le modèle de prédiction à l'étape b :

$$f_b = \sum_{k=1}^{b} \mathcal{T}_k$$

L'algorithme XGBoost construit des arbres de manière séquentielle et si le modèle final contient B arbres, la prédiction finale de Y pour une observation x_i peut s'écrire:

$$\hat{y}_i = \sum_{b=1}^{B} \mathcal{T}_k(x_i)$$

Au lancement de l'algorithme, cette prédiction s'écrit de la façon suivante :

$$\hat{y}_i^0 = 0$$

Puis, à l'étape 1, elle s'écrit :

$$\hat{y}_i^1 = \hat{y}_i^0 + f_1(x_i)$$

Et ensuite, à l'étape b, elle s'écrit :

$$\hat{y}_i^b = \hat{y}_i^{b-1} + \mathcal{T}_b(x_i) = \sum_{k=1}^{b} \mathcal{T}_k(x_i)$$

A chaque étape, l'arbre choisi est celui qui minimise la fonction objectif Obj.

Classement des variables explicatives (importance d'une variable)

L'importance d'une variable correspond au gain en performance de l'algorithme lorsqu'il décide d'utiliser cette variable au moment de la construction des classifieurs. Ainsi, plus une variable est utilisée et plus son gain global sera élevé.

Les principaux avantages de l'algorithme XGBoost

Cette approche dispose d'un avantage majeur : comme tout algorithme construit à partir d'arbres de décisions, il est capable de restituer les variables les plus discriminantes du modèle qu'il a construit.
Un autre argument motivant le choix de cet algorithme est sa capacité à prendre en compte l'interaction entre les différentes variables explicatives mises à sa disposition.
Enfin, un autre avantage de cette technique est la possibilité de traiter efficacement des problèmes lorsque le lien entre la variable à prédire et les variables explicatives n'est pas linéaire.

L'une des principales limites : l'algorithme XGBoost est assez complexe et est par conséquent souvent qualifié de « boîte noire », en effet les liens entre les variables explicatives et la variable à expliquer ne sont pas faciles à mettre en évidence (« pistes d'audit »).

Machine Learning : schéma récapitulatif des méthodes présentés

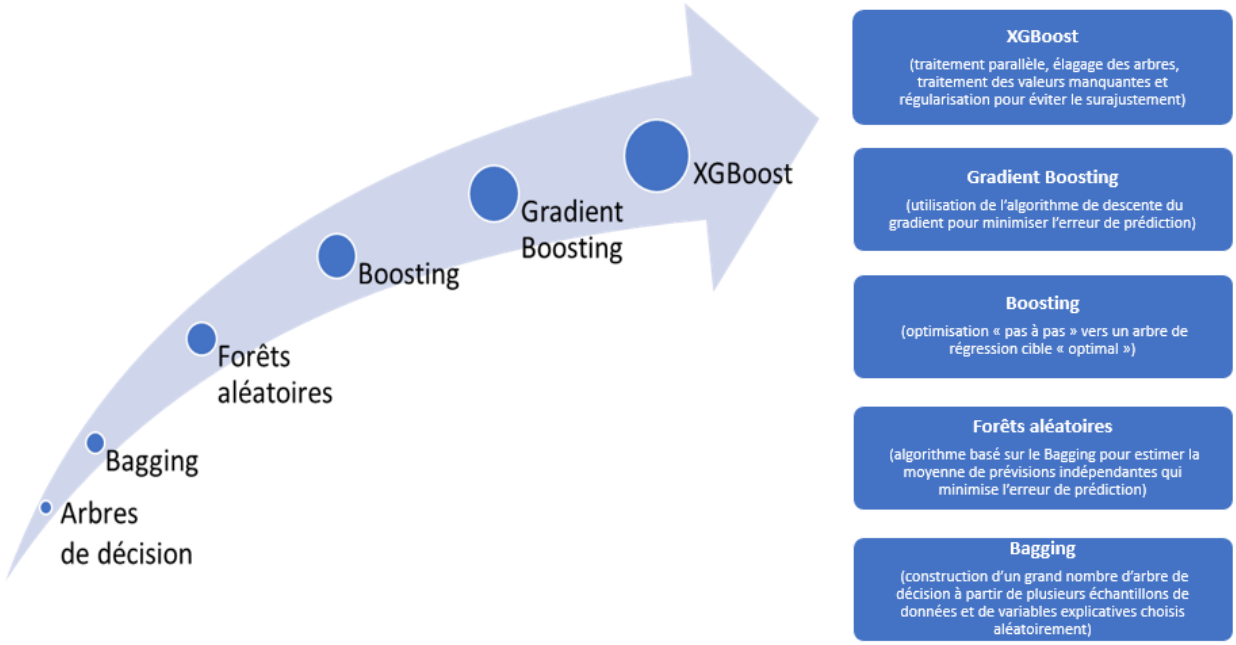


FIGURE 17 : Algorithmes de Machine Learning présentés

2.4. Evaluation des modèles de prédiction

L'objectif de cette partie est de présenter différents indicateurs qui sont fréquemment utilisés pour évaluer la performance d'un modèle de prédiction. Ces indicateurs sont utilisés notamment pour comparer la performance de différents modèles construits.

L'échantillon apprentissage / test et la validation croisée

La méthode de validation croisée (ou « cross-validation » en anglais) d'un modèle de prédiction (Bibliographie : 3. et 7.) permet d'utiliser l'intégralité de la base d'observations disponibles pour la construction et la validation du modèle. La base initiale est ainsi séparée en deux parties distinctes : la première partie pour l'apprentissage (et donc la création du modèle) et la seconde pour évaluer la performance du modèle. La méthode de validation croisée la plus fréquemment utilisée est la « cross-validation k-fold ». Elle consiste à diviser la base initiale en k sous bases et d'utiliser une à une chacune des bases en base de validation et le reste comme base d'apprentissage. Chaque sous base n'est utilisée qu'une seule fois comme base de validation. La moyenne des k erreurs quadratiques moyennes est calculée pour estimer l'erreur de prédiction.

Taux d'erreur MAE et RMSE

En considérant M un prédicteur construit à partir d'une base d'historique (apprentissage) par un algorithme de Machine Learning (par exemple de type forêts aléatoires ou bien Gradient Boosting).
Soit un échantillon de test de n observations. M est une matrice de tailles n x p (avec p variables explicatives X) et un vecteur Y de variables à expliquer de taille n, tel que $Y = (y_1 ; y_2 ; \dots ; y_n)$.

$$\hat{Y} = M(X)$$

Option 1 : l'erreur de prédiction peut être mesurée par l'**Erreur Absolue Moyenne / Mean Absolute Error (MAE)** :

$$MAE = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n}$$

Option 2 : l'erreur de prédiction peut également être mesurée par **l'erreur quadratique moyenne dite « Root Mean Squared Error » (RMSE)** qui peut être estimée par la racine carrée de l'estimateur de variance des résidus :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}}$$

La courbe ROC et l'indicateur AUC

Une courbe ROC (« Receiver Operating Characteristic ») est un graphique représentant les performances d'un modèle de prédiction. Cette courbe permet notamment d'évaluer les performances d'un modèle de classification pour tous les seuils de classification (Bibliographie 7.).

Soient T_n le nombre de vrais négatifs, F_p est le nombre de faux positifs, F_n est le nombre de faux négatifs et T_p le nombre de vrais positifs.

Pour construire la courbe ROC, Il est nécessaire de définir au préalable deux taux :

- La **Sensibilité** qui est le rapport entre T_p et la somme de T_p et de F_n.
- La **Spécificité** qui est le rapport entre T_n et la somme de T_n et de F_p.

La courbe ROC est la courbe représentant la Sensibilité en fonction de « 1 - la Spécificité » en faisant évoluer le seuil de probabilité à partir duquel il est considéré qu'une observation peut porter le « label » positif. Le choix des seuils utilise les probabilités « prédites » et les différentes valeurs qu'elles prennent en les utilisant tour à tour comme seuil.

L'AUC (« Area Under the Curve ») représente l'aire sous la courbe ROC, plus l'AUC est proche de 1, plus l'algorithme ou la méthode testé est performant. A l'inverse, lorsque l'AUC est proche de 0,5, le modèle testé n'est pas considéré comme performant (Bibliographie 19.).

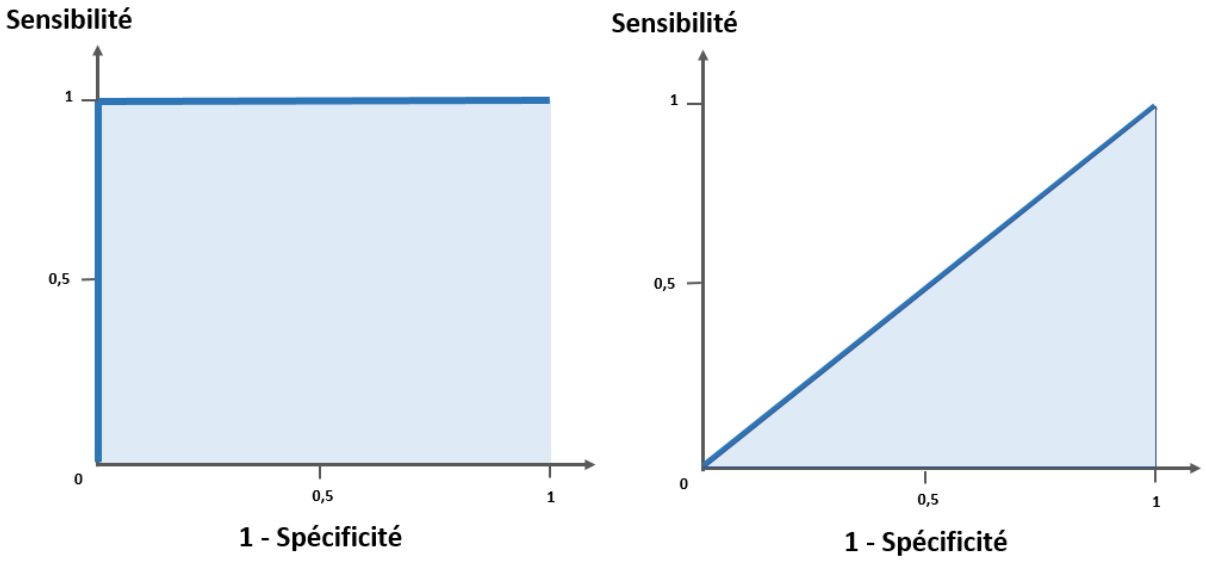


FIGURE 18 : Courbes ROC pour un modèle optimal (à gauche) et pour un modèle aléatoire (à droite)

3. Classification et modélisation : mise en application des outils

L'objectif de ce chapitre est de mettre en application les différents outils présentés en partie 2. pour classifier les variables explicatives et construire différents modèles de prédiction des versements libres.

3.1. Introduction des packages R utilisés (traitements informatiques)

La mise en application des outils de classification et de modélisation a été réalisée via le logiciel de statistiques R (RStudio), en utilisant les packages suivants :

Packages R utilisés	Objectifs
"doParallel"	Optimisation du temps de calcul avec R : La taille de la base de données : 1 002 460 lignes d'observation. Afin d'optimiser le temps de calcul, l'utilisation des 7 cœurs du microprocesseur de l'ordinateur a été optimisée en utilisant la librairie R « doParallel » : cette librairie permet de concentrer les 7 cœurs en parallèle à l'exécution d'un programme R.
"caret"	Utilisation de la fonction "train()" pour calibrer les paramètres d'exécution des algorithmes de Machine Learning.
"randomForest"	Application de l'algorithme des forêts aléatoires (classification).
"xgboost"	Application de l'algorithme XGBoost (classification).
"pROC"	Calcul de l'indicateur AUC (Aera Under the Curve) pour évaluer la performance d'un modèle de prédiction Machine Learning.
"aod"	Réalisation d'un test de significativité de Wald (construction des modèles GLM).
"mgcv"	Construction des modèles GAM.

TABLEAU 16 : Liste des packages R utilisés en partie 3.

3.2. Classification des variables explicatives

L'objectif de cette partie est d'établir une classification des variables explicatives à considérer dans le cadre des travaux de modélisation des versements libres en mettant en application les outils suivants :

1. La méthode ascendante (« forward stepwise selection ») basée sur le modèle GLM (Logit) avec le critère de sélection d'Akaike (AIC) ;
2. L'algorithme des forêts aléatoires ;
3. L'algorithme XGBoost.

La liste retenue pour les travaux de classification des variables explicatives est la suivante :

	Assuré	Contrat	Macro-économiques
Variables potentiellement explicatives	L'âge	- L'ancienneté - Le niveau de la PM d'ouverture : - "1" pour 1er quartile - "2" pour le 2ème quartile - "3 pour le 3ème quartile - "4" pour le 4ème quartile	- Le TME (annuel) - L'IPL (annuel) - Le taux de chômage (annuel) - Le CAC 40 au 31/12 de l'année N

TABLEAU 17 : Liste des variables explicatives retenues pour les travaux de classification

Le niveau de PM d'ouverture en euros par quartile a été privilégié au montant en euros de PM d'ouverture pour optimiser le temps de calcul et la consommation de mémoire vive de l'ordinateur utilisé lors de l'application des algorithmes de Machine Learning.

Principale limite de cette démarche
Les spécificités de la distribution intra-quartile des montants de PM d'ouverture ne sont pas prises en compte.
Remarque : le classement obtenu avec la méthode ascendante reste inchangé selon qu'il s'agisse du montant de la PM d'ouverture ou du niveau de la PM d'ouverture par quartile. Ce qui peut nuancer la limite de cette approche.

3.2.1. Application de la méthode ascendante (« forward stepwise selection » AIC)

Le processus d'exécution se déroule en 2 étapes :

1ère étape - Fixation des limites m_o et m_f d'exécution du programme d'ajout de variable par itération :

- m_o correspond au modèle GLM de base sans variables explicatives (« VL BOOL = 1 »),
- m_f correspond au modèle GLM intégrant l'ensemble des variables explicatives (« VL_BOOL = Ancienneté + NIVEAU_PM + … »).

2nd étape - Lancement de la classification des variables avec la méthode ascendante :
A chaque itération, les variables sont ajoutées au modèle une par une par ordre d'importance (les plus « explicatives » en premier).

Résultats de la classification (méthode ascendante) :

Rang	Variables explicatives	AIC à chaque ajout de variable au modèle
1	Niveau de la PM d'ouverture (quartiles)	420 788,90
2	Ancienneté	415 608,80
3	IPL (annuel)	415 387,70
4	Age	415 170,60
5	CAC 40 (au 31/12 de l'année N)	415 144,30
6	Taux de chômage (annuel)	415 139,10
7	TME (annuel)	415 139,00

TABLEAU 18 : Classification des variables explicatives obtenue avec la méthode ascendante (AIC)

Point d'attention : les montants d'AIC étant assez proches, la discrimination des variables explicatives n'est donc pas forte.

3.2.2. Application de l'algorithme des forêts aléatoires

Limites de capacités de l'ordinateur utilisé

Les capacités de l'ordinateur à disposition étant limitées, l'exécution de l'algorithme des forêts aléatoires sur l'ensemble de la base de données n'est pas possible (base de 1 002 460 lignes d'observation).
Au bout d'un certain temps, un message d'erreur apparait sous R précisant que la limite de consommation de la mémoire vive est dépassée. Le programme est alors stoppé.

Démarche retenue pour pallier ces limites

Pour ne pas réduire le périmètre de l'étude, il a été choisi de découper la base de données en plusieurs bases de données distinctes de telle manière que l'ordinateur puisse exécuter l'algorithme.
Le critère de découpage choisi est l'âge, puisque cette variable apparait secondaire par rapport à l'ancienneté ou le niveau de la PM d'ouverture dans les résultats de la classification avec la méthode ascendante (les indices macro-économiques n'ont pas été considérés, car ils évoluent en fonction de l'année de mouvement des flux observés (2010 à 2017), ce qui reviendrait à réduire l'historique 2010 à 2017).
Ainsi, la taille limite d'exécution de l'algorithme « Forêts aléatoires » a été évaluée par itération à environ 195 000 lignes.

La base de données (1 002 460 lignes d'observation) a par conséquent été découpé en 6 bases de données :

- 1ère base pour un âge de 0 à 30 ans (164 251 lignes),
- 2ème base pour un âge de 31 à 45 ans (165 242 lignes),
- 3ème base pour un âge de 46 à 55 ans (179 046 lignes),
- 4ème base pour un âge de 56 à 64 ans (188 574 lignes),
- 5ème base pour un âge de 65 à 76 ans (190 776 lignes),
- 6ème base pour un âge de 77 à 117 ans (114 571 lignes).

Principale limite de cette démarche

Cette démarche a comme principale limite d'apporter un biais à la vision globale (d'autant plus que les variables « ancienneté » et « PM d'ouverture du contrat » sont corrélées à la variable « âge »). Selon la tranche d'âge considérée, l'importance d'une de ces trois variables peut donc être surestimée par rapport à une autre.
La solution retenue consiste à confronter les résultats des classements obtenus par tranche d'âge et d'identifier les tendances communes sur l'ensemble des tranches en créant un classement global avec des groupes de variables.

1ère approche : utilisation de la fonction « randomForest() » pour exécuter l'algorithme des forêts aléatoires.

Dans le cadre de l'utilisation de la fonction « randomForest », deux paramètres clés définis par défaut sont :

- **« ntree »** : le nombre d'arbres créés qui est fixé à 500 par défaut,
- **« mtry »** : le nombre de variables testées à chaque division qui est fixé par défaut à la partie entière inférieure de la racine carré du nombre de variables explicatives considérées (ici égal à 2 pour 7 variables explicatives).

Compte tenu des limites de capacités informatiques de l'ordinateur utilisé, il a été retenu de faire varier uniquement le paramètre « mtry » pour réduire au maximum l'erreur de prédiction (en augmentant progressivement ce paramètre au-delà de 2).

Résultats de la classification :

Remarque : en utilisant la fonction « randomForest », le fait d'augmenter le paramètre « mtry » dans les limites de capacités de l'ordinateur utilisé n'a pas permis de réduire l'erreur de prédiction.

	0 à 30 ans	31 à 45 ans	46 à 55 ans	56 à 64 ans	65 à 76 ans	77 à 117 ans
Taux d'erreur Out Of Bag	**5,65%**	**5,33%**	**5,40%**	**6,51%**	**5,21%**	**3,864%**
Rang 1	**Niveau de la PM d'ouverture (quartiles)**	**Niveau de la PM d'ouverture (quartiles)**	**Ancienneté**	**Ancienneté**	**Ancienneté**	**Ancienneté**
Rang 2	**Age**	**Ancienneté**	**Niveau de la PM d'ouverture (quartiles)**	**Niveau de la PM d'ouverture (quartiles)**	**Age**	**Age**
Rang 3	**Ancienneté**	**Age**	**Age**	**Age**	**Niveau de la PM d'ouverture (quartiles)**	**Niveau de la PM d'ouverture (quartiles)**
Rang 4	**IPL (annuel)**	**IPL (annuel)**	**TME (annuel)**	**IPL (annuel)**	**IPL (annuel)**	**IPL (annuel)**
Rang 5	CAC 40 (au 31/12 de l'année N)	TME (annuel)	CAC 40 (au 31/12 de l'année N)	TME (annuel)	CAC 40 (au 31/12 de l'année N)	CAC 40 (au 31/12 de l'année N)
Rang 6	TME (annuel)	CAC 40 (au 31/12 de l'année N)	IPL (annuel)	CAC 40 (au 31/12 de l'année N)	TME (annuel)	TME (annuel)
Rang 7	Taux de chômage (annuel)	Taux de chômage (annuel)	Taux de chômage (annuel)	Taux de chômage (annuel)	Taux de chômage (annuel)	Taux de chômage (annuel)

TABLEAU 19.1 : Classement des variables explicatives obtenu avec l'algorithme de forêts aléatoires (pour chaque tranche d'âge).

2nd approche : utilisation de la fonction "train()" pour calibrer les paramètres d'exécution des algorithmes des forêts aléatoires.

Une solution pour optimiser l'utilisation de l'algorithme des forêts aléatoires et plus précisément le choix du paramétrage est l'utilisation de la librairie **« caret ».**
En effet, la fonction « train » de cette librairie est ainsi utilisée notamment avec la fonction « trainControl » pour obtenir une validation croisée de l'erreur de prédiction (argument de la fonction : « cv ») sur un échantillonnage test dont la taille est précisée par le nombre de blocs considérés (argument de la fonction : « number = »). L'idée est d'obtenir une validation croisée de l'erreur de prédiction estimée.

Limites rencontrées
L'utilisation de cette librairie pour optimiser l'exécution de l'algorithme des forêts aléatoires sollicite davantage les capacités de l'ordinateur.

Comparaison des résultats obtenus (librairie « caret » versus librairie « randomForest »)
Compte tenu des capacités de l'ordinateur utilisé, des résultats ont pu être obtenus uniquement pour la 6ème base de données pour un âge de 77 à 117 ans qui est la base la plus petite en taille (114 571 lignes).
Avec la librairie « caret », le taux d'erreur diminue de 0,005%, inversant le classement uniquement des variables les plus significatives (ancienneté et âge). Le classement des autres variables reste inchangé.

Librairie "randomForest"	Librairie "caret"
77 à 117 ans	
3,864%	3,859%
Ancienneté	Age
Age	Ancienneté
Niveau de la PM d'ouverture (quartiles)	Niveau de la PM d'ouverture (quartiles)
IPL (annuel)	IPL (annuel)
CAC 40 (au 31/12 de l'année N)	CAC 40 (au 31/12 de l'année N)
TME (annuel)	TME (annuel)
Taux de chômage (annuel)	Taux de chômage (annuel)

TABLEAU 19.2 : Classement des variables explicatives obtenus par l'algorithme « Forêts aléatoires » pour la tranche d'âge 77 à 117 ans : librairie « caret » versus librairie « randomForest » (taux d'erreur)

Etant donné que la base de données a été découpée en plusieurs bases par tranche d'âge et que la démarche retenue est d'identifier les tendances communes sur l'ensemble des tranches en créant un classement global avec des groupes de variables et enfin compte tenu de la diminution de l'erreur de prédiction qui reste à la marge, l'utilisation de la librairie « randomForest » peut être considérée comme la plus appropriée.

3.2.3. Application de l'algorithme XGBoost

Par construction, l'alogrithme « XGBoost » sollicite moins les capacités de l'ordinateur que l'algorithme des forêts aléatoires (comme expliqué en partie 2). L'algorithme « XGBoost » a ainsi pu être utilisé avec la fonction « train() » pour chacune des tranches d'âge.

Résultats de la classification (librairie « caret », algorithme XGBoost) :

	0 à 30 ans	31 à 45 ans	46 à 55 ans	56 à 64 ans	65 à 76 ans	77 à 117 ans
Taux d'erreur (cross-validation)	5,65%	5,33%	5,40%	6,51%	5,21%	3,86%
Rang 1	Niveau de la PM d'ouverture (quartiles)	Niveau de la PM d'ouverture (quartiles)	Niveau de la PM d'ouverture (quartiles)	Ancienneté	Ancienneté	Ancienneté
Rang 2	Ancienneté	Ancienneté	Ancienneté	Niveau de la PM d'ouverture (quartiles)	Niveau de la PM d'ouverture (quartiles)	Niveau de la PM d'ouverture (quartiles)
Rang 3	Age	Age	TME (annuel)	IPL (annuel)	Age	Age
Rang 4	Taux de chômage (annuel)	CAC 40 (au 31/12 de l'année N)	Taux de chômage (annuel)	Age	IPL (annuel)	IPL (annuel)
Rang 5	IPL (annuel)	IPL (annuel)	IPL (annuel)	TME (annuel)	Taux de chômage (annuel)	CAC 40 (au 31/12 de l'année N)
Rang 6	CAC 40 (au 31/12 de l'année N)	Taux de chômage (annuel)	Age	Taux de chômage (annuel)	TME (annuel)	Taux de chômage (annuel)
Rang 7	TME (annuel)	TME (annuel)	CAC 40 (au 31/12 de l'année N)	CAC 40 (au 31/12 de l'année N)	CAC 40 (au 31/12 de l'année N)	TME (annuel)

TABLEAU 20 : Classement des variables explicatives obtenu avec l'algorithme XGBoost (pour chaque tranche d'âge).

Point d'attention concernant le taux d'erreur :
Même si le taux d'erreur par tranche d'âge calculé lors de l'exécution des algorithmes « Forêts aléatoires » et XGBoost » semble assez proche (à 0,01% près), la performance des deux modèles doit être formellement

testée à partir d'un échantillon et comparée sur la base d'indicateurs pertinents (exemple : RMSE, AUC présentés en partie 2.4.). Les résultats de cette comparaison sont formalisés dans la partie suivante (3.2.4.).

3.2.4. Comparaison des méthodes de classification

Comme précisé précédemment, les résultats obtenus avec la méthode ascendante (« forward stepwise selection » AIC) ne présentent pas une discrimination des variables explicatives assez forte pour être exploités.

L'analyse de la performance des deux algorithmes Machine Learning (XGBoost et forêts aléatoires), en construisant chaque modèle de prédiction à partir de 75% des données et l'échantillon de test à partir de 25% de données restantes, permet de conclure que l'algorithme XGBoost est le plus performant des deux (comme le démontrent les indicateurs de performance présentés ci-après). En effet, pour le modèle XGBoost, le critère AUC est plus proche de 1 et l'erreur quadratique moyenne (RMSE) est plus faible pour chaque tranche d'âge.

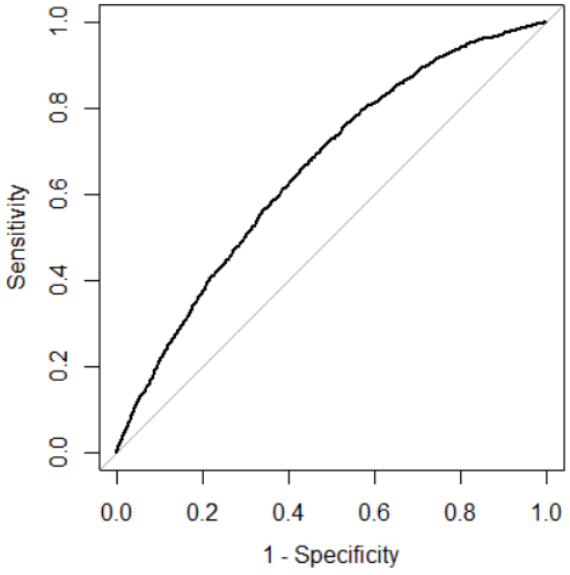
Remarque : *au regard du critère AUC, la performance du modèle « forêts aléatoires » semble assez faible (AUC proche de de 0,5).*

AUC	0 à 30 ans	31 à 45 ans	46 à 55 ans	56 à 64 ans	65 à 76 ans	77 à 117 ans
XGBoost	0,6559	0,6460	0,6558	0,6260	0,6095	0,6013
Forêts aléatoires	0,5399	0,5254	0,5137	0,5103	0,5097	0,5314

RMSE	0 à 30 ans	31 à 45 ans	46 à 55 ans	56 à 64 ans	65 à 76 ans	77 à 117 ans
XGBoost	0,2292	0,2232	0,2243	0,2453	0,2211	0,1917
Forêts aléatoires	0,2373	0,2306	0,2324	0,2551	0,2280	0,1959

TABLEAU 21 : Choix du meilleur modèle de prédiction (critères AUC et RMSE)

Algorithme « XGBoost » (0 à 30 ans)



Algorithme « Forêts aléatoires » (0 à 30 ans)

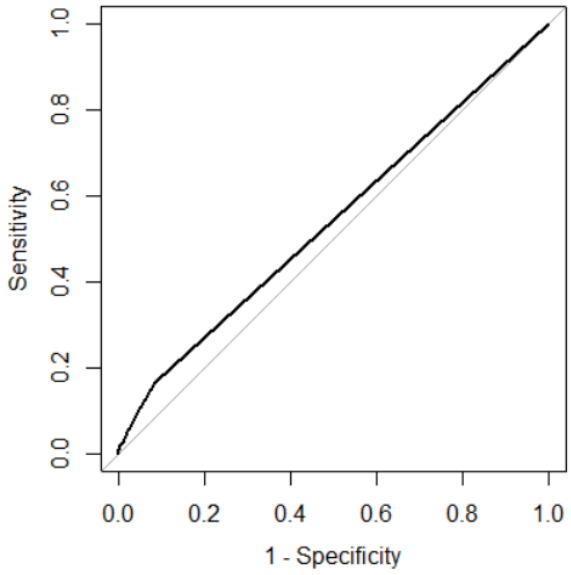


FIGURE 19 : Choix du meilleur modèle de prédiction (Courbes ROC : exemple pour la tranche d'âge de 0 à 30 ans)

Remarque : la classification des variables les moins significatives avec l'algorithme des forêts aléatoires est plus stable entre les tranches que celle obtenue avec l'algorithme XGBoost. Les classifications obtenues pour les deux tranches 65 à 76 ans et 77 à 117 ans matérialisent ce constat. Ce qui est cohérent avec les qualités de l'algorithme des forêts aléatoires présentées en partie 2.3.3.

3.2.5. Constitution et classement de groupes de variables explicatives

L'analyse des résultats obtenus avec les différentes méthodes de classification met en évidence l'existence de 3 groupes de variables explicatives classés par ordre d'importance (de la plus significative à la moins significative) :

Groupes	Variables explicatives du groupe
1er Groupe	- Ancienneté - Niveau de la PM d'ouverture (quartiles)
2ème Groupe	- Age - IPL (annuel)
3ème Groupe	- TME (annuel) - Taux de chômage (annuel) - CAC 40 au 31/12 de l'année N

TABLEAU 22 : Groupes de variables explicatives classés par ordre d'importance

3.3. Modélisation de la fréquence de versement libre

Introduction :

L'objectif de ce chapitre est de proposer différentes méthodes pour modéliser les versements libres dans le cadre IFRS 17 (construction de lois de versement libre).

Estimation du montant de versements libres par contrat
Pour estimer le montant de versements libres par contrat, la formule retenue dans le cadre de ce mémoire est la suivante :

Montant de VL par contrat = fréquence de VL (%) x Montant moyen de VL (€)

Point d'attention : Les méthodes de modélisation retenues dans le cadre de cette étude doivent être intégrables en l'état dans le cadre des « Model points » du modèle de projection d'AXA France.

Dans un premier temps, les travaux seront consacrés principalement à la modélisation de la fréquence de versement libre. Dans un second temps, les travaux porteront sur la modélisation du montant moyen en euros de versement libre (partie 3.4.).

Lois de versement libre (fréquence de versement) :

Dans cette partie, les méthodes de modélisation des VL qui ont été mises en œuvre sont les suivantes :

	Méthodes de modélisation utilisées	
N°	**Modélisation de la fréquence des VL**	**Modélisation du montant moyen de VL**
1	**Loi par ancienneté** (construite à partir de triangles)	**Loi par ancienneté** (construite à partir de triangles)
2	- **Loi par ancienneté spécifique pour la tranche d'âge considérée comme la plus concentrée en nombre de VL** (construite à partir de triangles) - Taux moyen global de VL pour les autres tranches d'âge (fréquence moyenne globale)	
3	**Loi GLM (logistique)** avec la variable explicative **Ancienneté ou PM d'ouverture**	
4	**Loi GLM (logistique)** avec les variables explicatives suivantes : **- Ancienneté ou PM d'ouverture** **- Age** **- avec ou sans variable macro-économique** (TME, IPL ou taux de chômage)	
5	**Loi GAM (logistique)** avec les variables explicatives suivantes : **- Ancienneté ou PM d'ouverture** **- Age avec une fonction spline de lissage cubique ou « à plaques minces »** **- avec ou sans variable macro-économique** (TME, IPL ou taux de chômage)	

TABLEAU 23 : Méthodes de modélisation des versements libres mises en œuvre

Points d'attention :

- Compte tenu du niveau de corrélation observé entre les variables Ancienneté, Age et Niveau de la PM d'ouverture (quartiles), qui sont classées dans les 2 premiers groupes des variables explicatives, la pertinence d'utiliser chacune des variables devra être analysée dans le cadre du « backtesting » des modèles de prédiction mis en œuvre (estimation versus observation) ;
- Compte tenu du niveau de corrélation assez significatif qui existe entre les variables macro-économiques, il a été décidé de n'utiliser qu'une seule variable à la fois dans le cadre des travaux de modélisation des versements libres.
- La variable CAC 40 au 31/12 de l'année N, très corrélée à la variable TME (annuel) et non classée dans le 1er groupe des variables explicatives ne sera finalement pas prise en compte dans le cadre des travaux de modélisation des versements libres. La variable TME est considérée comme étant plus « proche » de l'activité d'assurance vie (notamment, pour la calibration du taux technique au regard des exigences règlementaires, Art. A. 132-18 du code des assurances) ;
- Pour les modèles additifs généralisés (GAM), l'utilisation de la fonction spline de lissage a été privilégié pour la variable « âge » qui est classée dans le 2ème groupe des variables les plus explicatives (dont la relation avec la variable à expliquer semble non linéaire).

3.3.1. Loi de versement par ancienneté

Pour construire une loi de versement libre par ancienneté, l'approche retenue a été de construire des triangles de cadencement des versements libres avec en ligne la date d'effet du contrat et en colonne l'ancienneté du contrat.

	Nombre de versements libres par année police			
Génération	1	2	3	(…)
	4 152,00	4 712,00	3 562,00	
(…)				
2 007				
2 008			356,00	
2 009		593,00	484,00	
2 010	617,00	612,00	389,00	
2 011	447,00	471,00	401,00	
2 012	471,00	607,00	518,00	
2 013	557,00	723,00	554,00	
2 014	647,00	677,00	467,00	
2 015	562,00	552,00	393,00	
2 016	432,00	477,00		
2 017	419,00			

FIGURE 20 : Extrait du triangle de cadencement du nombre de versement libre par ancienneté en fonction de l'année d'effet du contrat (« année police »)

Ces triangles ont été obtenus à partir de l'exécution d'un programme informatique de retraitement des données (programme WPS[2] déjà existant, développé et validé par le département Risk Management d'AXA France).

Les triangles obtenus présentent notamment le cadencement du nombre, de la fréquence et du montant moyen en euros de versements libres par contrat. A partir de ces triangles obtenus, des lois de VL (fréquence et montant moyen) sont construites par année de mouvement (de 2010 à 2017).

[2] World Programming System (équivalent à Statistical Analysis System - SAS)

La loi de versement libre par ancienneté retenue est la loi moyenne sur l'historique 2010 à 2017 de la base de données.

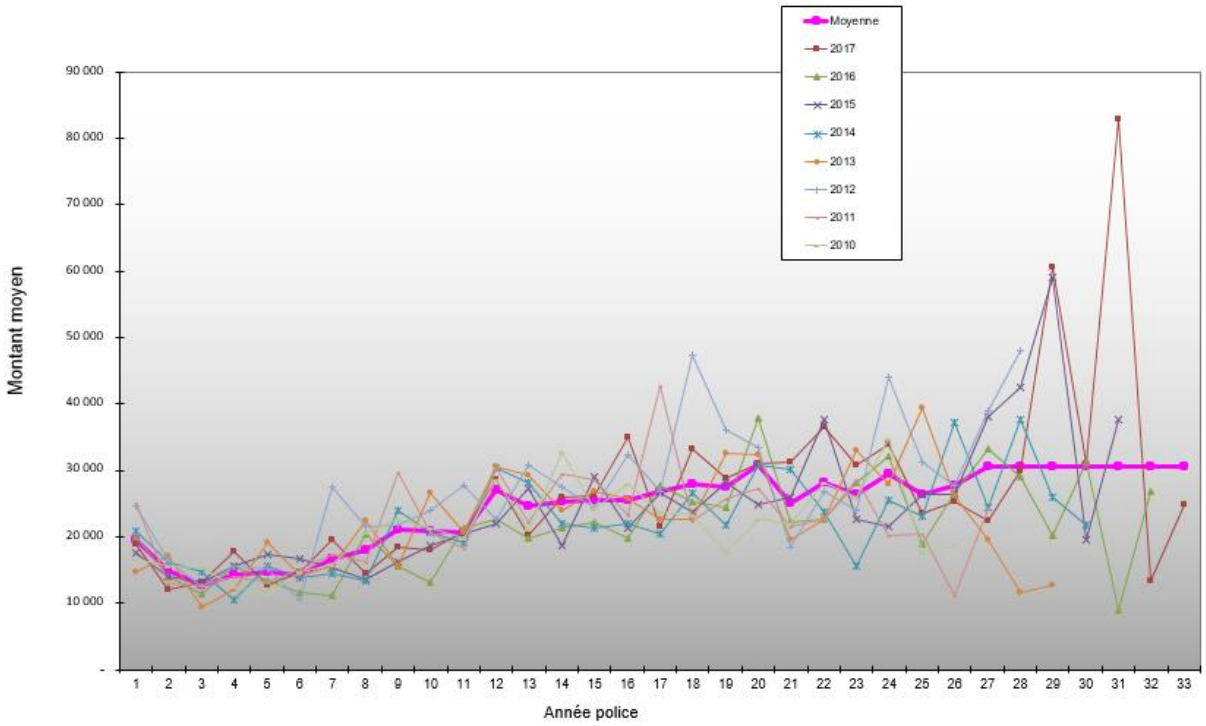


FIGURE 21 : Evolution du montant moyen en euros de versement libre en fonction de l'ancienneté des contrats

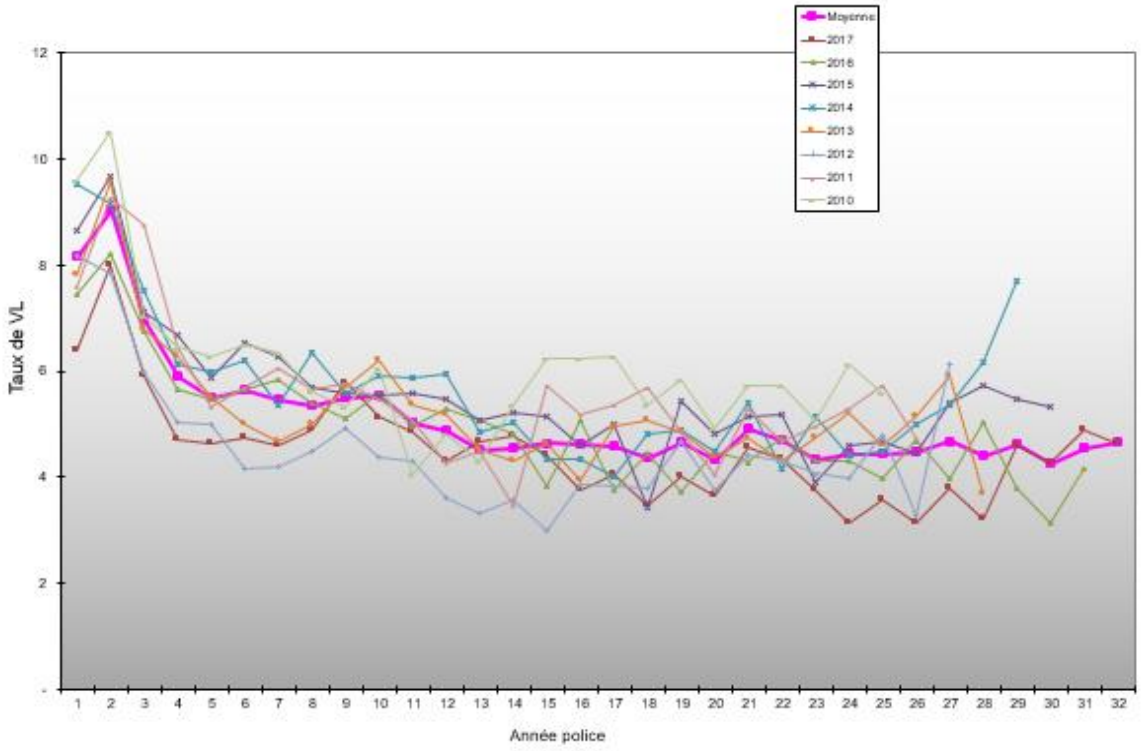


FIGURE 22 : Evolution du taux de versement en fonction de l'ancienneté des contrats (fréquence en %)

3.3.2. Loi par ancienneté « par tranche d'âge » pour la fréquence de versement

L'observation de l'évolution de la fréquence et surtout du nombre de versement libre en fonction de l'âge a permis de mettre en évidence un « pic » de versements sur la tranche d'âge entre 51 et 70 ans.

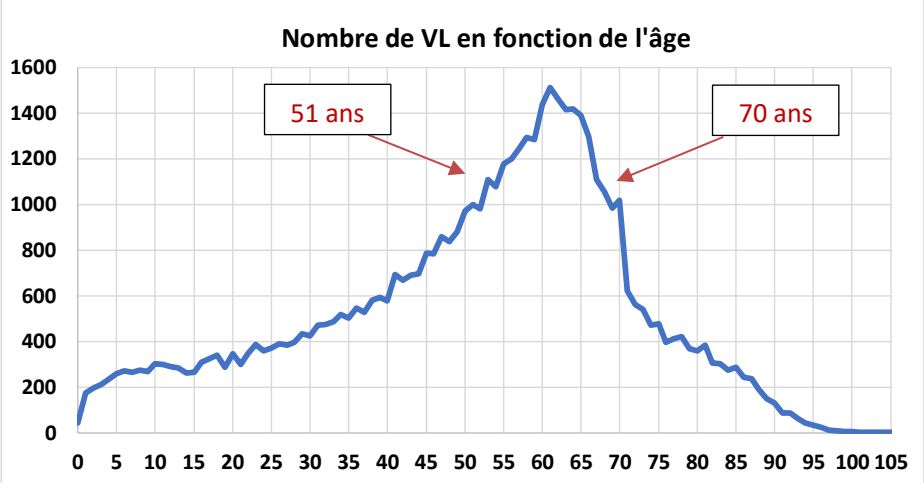


FIGURE 23 : Evolution du nombre de versement libre en fonction de l'âge (cumul des mouvements de 2010 à 2017)

L'objectif de cette partie est de présenter les modalités de construction d'une **loi de VL par ancienneté pour la fréquence de versement qui serait spécifique à la tranche d'âge 51-70 ans**, considérée comme la tranche la plus concentrée en nombre de versement libre.

Tout d'abord, pour confirmer les limites de la tranche observée, une classification des données a été réalisée par classification ascendante Hiérarchique (« CAH ») et par moyennes mobiles (« K-means »). Ces outils mathématiques ont été présentés dans la partie 2.

<u>Résultats des classification CAH et K-means :</u>

Les résultats suivants confirment un regroupement des âges entre environ 50 ans et 70 ans (groupe 5 pour la méthode CAH et groupe 6 pour la méthode K-Means). Ainsi, la tranche observée [51 ans ; 70 ans] a été conservée pour la construction de la loi de VL.

CAH-Groupes	AGES																												
1	0	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22												
2	1	2	3	4	5	6																							
3	23	24	25	26	27	28	29	30	31	33																			
4	32	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	54	55	68	69					
5	53	56	57	58	59	60	61	62	63	64	65	66	67	70															
6	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	103		
7	97	98	99	100	101	102	104	105																					

KMEANS-Groupes	AGES																												
1	39	40	41	42	43	45	46	47	48	49	50	51	52	54	55	56	68	69	96										
2	44	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	102		
3	97	98	99	100	101	104	105																						
4	0	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21													
5	1	2	3	4	5	6																							
6	53	56	57	58	59	60	61	62	63	64	65	66	67	70															
7	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38												

FIGURE 24 : Résultats des classifications CAH et K-Means du nombre de VL par tranche d'âge

Construction de la loi de versement spécifique à la tranche d'âge 51-70 ans (pour la fréquence)
Les flux pour les âges d'assuré compris dans l'intervalle de 51 à 70 ans inclus ont été extraits de la base de données. Ainsi, la loi de fréquence de VL pour la tranche [51 ans ;70 ans] a pu être construite à partir de triangles (même méthode que celle décrit dans la partie 3.3.1.).

3.3.3. Loi GLM pour la fréquence de versement

L'objectif de cette partie est de présenter les modalités de construction d'une **loi de VL paramétrique pour la fréquence de versement** à partir d'un modèle GLM avec régression logistique (probabilité de verser ou pas). Cet outil mathématique a été présenté dans la partie 2.
L'idée est ici de présenter avec un exemple les étapes de la construction de loi GLM de fréquence de VL en utilisant le logiciel R.

Exemple considéré : modélisation de la fréquence de versement libre avec une loi GLM considérant les variables explicatives Ancienneté, Age et TME (annuel).
En utilisant la fonction « glm() » sous R, on obtient donc les coefficients β du modèle GLM :

Nom du coefficient	Valeur du coefficient	P-value associée
β_0 (constante)	-2,6418484	$< 2 \times 10^{-16}$
β_1 (associé à la variable « ancienneté »)	-0,0314235	$< 2 \times 10^{-16}$
β_2 (associé à la variable « âge »)	0,0021520	$< 2 \times 10^{-16}$
β_3 (associé à la variable « TME (annuel) »)	0,0108243	0,0119

TABLEAU 24 : coefficients β du modèle GLM

Si l'on considère un seuil de significativité α = 5% (« marge d'erreur »), on observe que la p-value est inférieur à α pour tous les coefficients (ce qui est considéré comme satisfaisant pour notre modèle de prédiction).

Ce constat est confirmé par les résultats obtenus à partir d'un test de Wald appliqué à chaque coefficient à l'exception de la constante (résultats du test présentés en annexe).

L'intervalle de confiance de l'estimation des coefficients avec un seuil de significativité à 5% est le suivant :

	Intervalle de confiance	
Nom du coefficient	**2,5%**	**97,5%**
β_0 (constante)	-2,669886411	-2,613810464
β_1 (associé à la variable « ancienneté »)	-0,032848599	-0,029998463
β_2 (associé à la variable « âge »)	0,001683592	0,002620464
β_3 (associé à la variable « TME (annuel) »)	0,002390104	0,019258579

TABLEAU 25 : intervalle de confiance (coefficients β du modèle GLM)

Calcul de la fréquence de versement libre à partir des coefficients β obtenus avec le modèle GLM

Rappel : Pour tout i entier, la loi de probabilité du modèle « Logit » est définie par :

$$\mathbb{P}[y_i = 1|X_i] = \frac{1}{1 + e^{-X_i\beta}}$$

Exemple d'application numérique

En utilisant cette formule aux données disponibles d'ancienneté (X_1), d'âge (X_2) et de TME annuel (X_3), la valeur de la fréquence de VL alors être calculée.

En considérant X_1 = 10 ans, X_2 = 52 ans et X_3 = 0,8458333 (pour l'année 2017).

$$\boldsymbol{X_i\beta} = (\beta_0 = -2{,}6418484) + (X_1 = 10) \times (\beta_1 = -0{,}0314235) + (X_2 = 52) \times (\beta_2 = 0{,}0021520) + (X_3 = 0{,}8458333) \times (\beta_3 = 0{,}0108243)$$

La fréquence calculée en utilisant la loi de probabilité du modèle « Logit » est égale 5,5460636%.

3.3.4. Loi GAM pour la fréquence de versement

L'objectif de cette partie est de présenter les modalités de construction **d'une loi pour la fréquence de versement à partir d'un modèle additif GAM avec régression logistique** (probabilité de verser ou pas). Cet outil mathématique a été présenté dans la partie 2.

L'idée ici est de présenter avec un exemple les étapes de la construction de loi GAM de fréquence de versement libre en utilisant le logiciel R.

Exemple considéré : modélisation de la fréquence de versement libre avec une loi GAM considérant les variables explicatives Ancienneté, Age et TME (annuel).

- **Pour la variable « âge », utilisation de la fonction spline de lissage cubique avec 9 degrés de liberté**

Point d'attention : la fonction spline de lissage « s() » est appliquée à la variable Age avec comme argument le type de lissage (« cr » correspond au lissage cubique) et le nombre de degré de liberté (k = 10 correspond à k-1 degrés de liberté).

				Splines associées à la variable « âge »								
Nom du coefficient	β_0 (constante)	β_1 (associé à la variable « ancienneté »)	β_2 (associé à la variable « TME (annuel) »)	**s(AGE).1**	**s(AGE).2**	**s(AGE).3**	**s(AGE).4**	**s(AGE).5**	**s(AGE).6**	**s(AGE).7**	**s(AGE).8**	**s(AGE).9**
Valeur du coefficient	-2,543950926	-0,030661109	0,007850484	-0,279464777	0,022500923	-0,208759339	-0,072448492	0,137323566	-0,263809365	-0,173326038	-1,239135292	-2,405255032

TABLEAU 26 : coefficients du modèle GAM (fonction spline de lissage cubique avec 9 degrés de liberté)

Les coefficients peuvent être retenus selon le seuil de significativité α considéré (comme dans le cas du modèle GLM).

Calcul de la fréquence de versement libre à partir des coefficients obtenus avec le modèle GAM :

- De la même manière que pour un modèle GLM, le calcul de la fréquence est effectué en utilisant la formule de la loi de probabilité du modèle « Logit ».
- **La sélection de la valeur de la fonction spline s(AGE).i optimale (avec i entier naturel Є [1 ;9]) est réalisée à partir de l'analyse des résultats du « backtesting » du modèle additif GAM.**

3.3.5. Présentation et interprétation des résultats

L'objectif de cette partie est de confronter les principaux résultats du backtesting obtenus avec chacune des méthodes de modélisation des versements libres présentés précédemment.

L'écart de backtesting est calculé à partir de la formule suivante :

$$Ecart\ backtesting\ (\%) = \frac{Montant\ de\ VL\ estimé\ pour\ l'année\ N\ (€) - Montant\ de\ VL\ observé\ en\ année\ N\ (€)}{Montant\ de\ VL\ observé\ en\ année\ N\ (€)}$$

Méthode 1 à 3 :

- **Une loi par ancienneté (construite à partir de triangles)**
- Versus **une loi par ancienneté spécifique pour la tranche d'âge considérée comme la plus concentrée en nombre de VL (construite à partir de triangles)**
- Versus **une loi GLM (logistique) par ancienneté**
- Versus **une loi GLM (logistique) par montant de PM d'ouverture**

Résultats du backtesting :

La confrontation des résultats du backtesting obtenus notamment pour les années 2016 et 2017 montrent que :

- la loi GLM est plus performante que la loi construite à partir de triangles,
- la variable ancienneté est plus déterminante que le montant en euros de la PM d'ouverture. En effet, l'erreur quadratique moyenne (RMSE) est plus faible en utilisant la variable ancienneté.
  L'utilisation des deux variables ancienneté et PM d'ouverture donne quasiment les mêmes résultats que ceux obtenus uniquement avec la variable ancienneté. La variable PM d'ouverture n'a donc pas été retenue pour la modélisation des versements libres.

ECARTS BACKTESTING (estimé versus observé en %)	Loi par ancienneté (triangles) : Taux de VL et Montant moyen de VL	-Taux de VL : Loi par ancienneté par tranche d'âge (triangles ou taux moyen global) -Montant moyen de VL : Loi par ancienneté (triangles)	-Taux de VL : Loi GLM_Logit avec ancienneté -Montant moyen de VL : Loi par ancienneté (triangles)	-Taux de VL : Loi GLM_Logit avec PM d'ouverture -Montant moyen de VL : Loi par ancienneté (triangles)
2010	-8,79%	-7,19%	-7,26%	-7,01%
2011	-7,19%	-5,44%	-6,27%	-4,71%
2012	2,36%	4,93%	3,18%	6,12%
2013	-0,63%	1,66%	-0,82%	2,84%
2014	-2,97%	-0,87%	-3,96%	0,49%
2015	-5,24%	-2,99%	-6,60%	-1,15%
2016	**13,29%**	**16,30%**	**11,28%**	**19,27%**
2017	9,47%	**13,55%**	7,66%	**16,42%**
RMSE 2010-2017	**10 396 263**	**11 755 769**	**9 369 885**	**13 581 239**
Moyenne des écarts en valeur absolue 2010-2017	**6,24%**	**6,62%**	**5,88%**	**7,25%**
Maximum des écarts en valeur absolue 2010-2017	**13,29%**	**16,30%**	**11,28%**	**19,27%**

TABLEAU 27 : Résultats du backtesting - Méthode 1 à 3 (modélisation des VL)

Méthode 3 à 5 :

- **Une loi GLM (logistique) par ancienneté**
- Versus **une loi GLM (logistique) par ancienneté considérant l'âge avec ou sans variable macro-économique (TME, IPL ou taux de chômage)**
- Versus **une loi GAM (logistique) par ancienneté considérant l'âge avec ou sans variable macro-économique (TME, IPL ou taux de chômage)**

<u>Résultats du backtesting :</u>

1. **Choix de la variable macro-économique**

En effet, au regard des résultats du backtesting suivants, le choix d'utiliser de la variable « TME (annuel) » pour la modélisation de la fréquence de VL apparaît le plus pertinent. L'erreur quadratique moyenne (RMSE) est notamment la plus faible en utilisant la variable « TME (annuel) ».

ECARTS BACKTESTING (estimé versus observé en %)	-Taux de VL : Loi GLM_Logit avec ancienneté, âge & TME -Montant moyen de VL : Loi par ancienneté (triangles)	-Taux de VL : Loi GLM_Logit avec ancienneté, âge & IPL -Montant moyen de VL : Loi par ancienneté (triangles)	-Taux de VL : Loi GLM_Logit avec ancienneté, âge & Taux de chômage -Montant moyen de VL : Loi par ancienneté (triangles)
2010	-5,58%	-0,61%	-9,37%
2011	-4,51%	**-14,98%**	-8,74%
2012	4,15%	-5,11%	2,82%
2013	-0,37%	-3,61%	1,11%
2014	-4,15%	-2,51%	-2,17%
2015	-7,65%	-1,27%	-4,43%
2016	9,48%	**13,87%**	**12,20%**
2017	6,18%	0,55%	5,34%
RMSE 2010-2017	**8 401 269**	**10 548 217**	**9 545 906**
Moyenne des écarts en valeur absolue 2010-2017	**5,26%**	**5,31%**	**5,77%**
Maximum des écarts en valeur absolue 2010-2017	**9,48%**	**14,98%**	**12,20%**

TABLEAU 28 : Résultats du backtesting - Méthode 4 (choix de la variable macro-économique)

2. Loi considérée comme la plus pertinente

La confrontation des résultats du backtesting obtenus montre que :

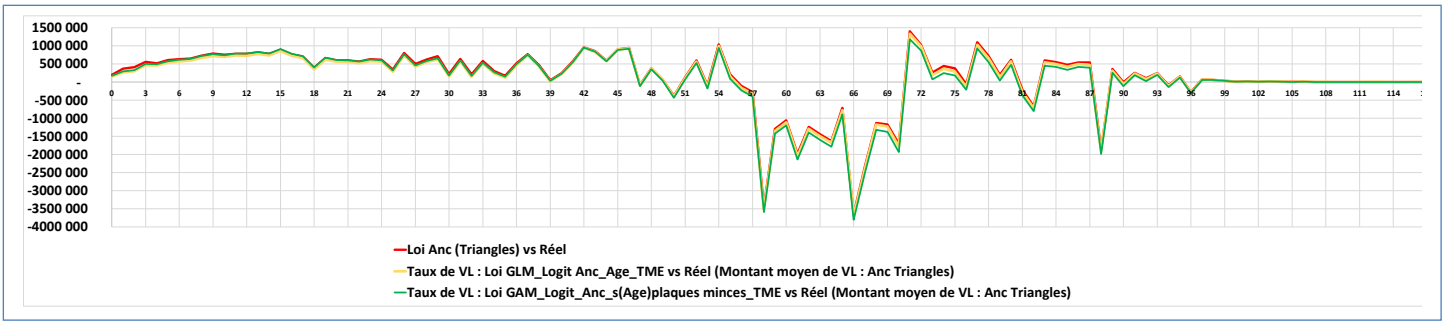
- Si on considère uniquement l'erreur quadratique moyenne (RMSE), les deux méthodes de modélisation de la fréquence de versement libre les plus performantes semblent être :
  - la loi GLM (Logit) avec les variables ancienneté, âge et TME (annuel),
  - et la loi GAM (Logit) avec la variable ancienneté, la variable âge en lui appliquant un spline de lissage « à plaques minces » et la variable TME (annuel).
- Si on considère également l'écart maximum en valeur absolue constaté entre l'estimation et le réel observé, cette dernière loi (GAM) semble être la plus performante des deux (avec 9,65% pour l'année de survenance 2018). En effet, la fonction spline de lissage « à plaques minces » semble particulièrement adaptée à la relation non linéaire qui existe entre la variable explicative âge et la variable à expliquer (fréquence de versement).

ECARTS BACKTESTING (estimé versus observé en %)	-Taux de VL : Loi GLM_Logit avec ancienneté -Montant moyen de VL : Loi par ancienneté (triangles)	-Taux de VL : Loi GLM_Logit avec ancienneté et âge -Montant moyen de VL : Loi par ancienneté (triangles)	-Taux de VL : Loi GLM_Logit avec ancienneté, âge & TME -Montant moyen de VL : Loi par ancienneté (triangles)	-Taux de VL : Loi GAM_Logit avec Ancienneté, s(âge) cubique & TME -Montant moyen de VL : Loi par ancienneté (triangles)	-Taux de VL : Loi GAM_Logit avec ancienneté, s(âge) à plaques minces & TME -Montant moyen de VL : Loi par ancienneté (triangles)
2010	-7,26%	-6,83%	-5,58%	-5,23%	-8,10%
2011	-6,27%	-5,97%	-4,51%	-4,17%	-7,07%
2012	3,18%	3,38%	4,15%	4,81%	1,63%
2013	-0,82%	-0,79%	-0,37%	0,47%	-2,58%
2014	-3,96%	-4,00%	-4,15%	-3,17%	-6,11%
2015	-6,60%	-6,73%	-7,65%	-6,45%	-9,29%
2016	**11,28%**	10,97%	9,48%	**11,01%**	**7,70%**
2017	7,66%	7,26%	6,18%	7,67%	4,39%
Model Point 2018	**13,11%**	**13,12%**	**11,93%**	**13,09%**	**9,65%**
RMSE 2010-2018	**10 829 963**	**10 646 084**	**9 760 733**	**10 307 303**	**9 934 561**
Moyenne des écarts en valeur absolue 2010-2018	**6,68%**	**6,56%**	**6,00%**	**6,23%**	**6,28%**
Maximum des écarts en valeur absolue 2010-2018	**13,11%**	**13,12%**	**11,93%**	**13,09%**	**9,65%**

TABLEAU 29 : Résultats du backtesting - Méthode 3 à 5 (modélisation des VL)

Par ailleurs, il est intéressant d'observer les résidus pour certains des modèles appliqués (écarts en euros entre l'estimation et le réel observé). A titre d'exemple, pour l'année de survenance 2017, l'observation des résidus en fonction de l'âge met notamment en évidence les points suivants :

- Pour la période d'âge entre 58 ans et 70 ans, la loi par ancienneté construite à partir de triangles semble la plus performante. Cette tranche d'âge correspond au « pic » du nombre de versements libres (beaucoup d'historique).
- Pour la période d'âge de 71 ans à 90 ans, la loi la plus performante est la loi GAM en considérant la variable ancienneté, la variable âge en appliquant un spline de lissage « à plaques minces » et la variable TME (annuel). Ce type de modélisation apparaît donc comme plus adaptée pour cette tranche d'âge où la fréquence de versement libre est plus faible. La fonction spline de lissage « à plaques minces » appliquée à l'âge permet d'améliorer le pouvoir prédictif du modèle (la relation entre les variables explicatives considérées et la variable à expliquer étant loin d'être linéaire).



ECARTS EN EUROS	Loi Anc (Triangles) vs Réel	Taux de VL : Loi GLM_Logit Anc_Age_TME vs Réel (Montant moyen de VL : Anc Triangles)	Taux de VL : Loi GAM_Logit_Anc_s(Age)plaques minces_TME vs Réel (Montant moyen de VL : Anc Triangles)
Entre 0 et 50 ans	27 817 456	25 354 773	26 808 838
Entre 51 et 57 ans	1 519 775	1 207 057	836 163
Entre 58 et 70 ans	-22 774 632	-23 465 223	-24 980 867
Entre 71 et 90 ans	6 429 113	5 265 905	3 347 637
Entre 91 et 117 ans	688 283	569 052	336 469
ECART TOTAL	**13 679 995**	**8 931 565**	**6 348 240**

FIGURE 25 : Evolution des résidus en fonction de l'âge (année de survenance 2017)

3.4. Modélisation du montant moyen de versement libre

L'objectif de cette partie est de trouver une alternative pertinente à la loi par ancienneté construite à partir de triangles pour améliorer la qualité de la prédiction des montants moyens en euros de versement libre.
L'idée est d'utiliser le modèle linéaire généralisé avec une loi de probabilité continue en considérant uniquement la distribution des montants de versements libres non nuls.
Dans le cadre de cette étude, le choix s'est porté sur la loi Log-Gamma et la loi Log-Normale (lois de probabilité continues).

Remarque : ces deux lois sont fréquemment utilisées dans le cadre de la modélisation des coûts individuels de sinistres en assurance non-vie.

Résultats du backtesting :

La confrontation des résultats du backtesting obtenus notamment pour les années 2016 et 2018 montre que la méthode de modélisation du montant moyen de versement libre la plus performante est celle utilisant un modèle GLM Log-Normale en considérant uniquement la variable âge. En effet, l'erreur quadratique

moyenne (RMSE) et l'écart maximum en valeur absolue constaté entre l'estimation et le réel observé sont les plus faibles avec cette méthode.

Remarques :

- la variable ancienneté semble ici moins déterminante,
- il serait opportun de réaliser des travaux supplémentaires pour identifier d'autres variables explicatives à prendre en compte pour modéliser le montant moyen de versement libre par contrat.

ECARTS BACKTESTING (estimé versus observé en %)	-Taux de VL : Loi GAM_Logit avec ancienneté, s(âge) à plaques minces & TME -Montant moyen de VL : Loi par ancienneté (triangles)	-Taux de VL : Loi GAM_Logit avec ancienneté, s(âge) à plaques minces & TME -Montant moyen de VL : Loi GLM_Log-Gamma avec ancienneté	-Taux de VL : Loi GAM_Logit avec ancienneté, s(âge) à plaques minces & TME -Montant moyen de VL : Loi GLM_Log-Normale avec ancienneté	-Taux de VL : Loi GAM_Logit avec ancienneté, s(âge) à plaques minces & TME -Montant moyen de VL : Loi GLM_Log-Normale avec âge
2010	-8,10%	-8,58%	-8,22%	-3,81%
2011	-7,07%	-8,12%	-7,85%	-4,75%
2012	1,63%	1,11%	1,32%	3,41%
2013	-2,58%	-3,03%	-2,89%	-2,17%
2014	-6,11%	-6,06%	-5,99%	-5,86%
2015	-9,29%	-8,28%	-8,28%	-9,11%
2016	**7,70%**	**9,29%**	**9,20%**	**6,91%**
2017	4,39%	6,55%	6,37%	3,11%
Model Point 2018	**9,65%**	**9,27%**	**9,12%**	**8,55%**
RMSE 2010-2018	**9 934 561**	**10 412 751**	**10 215 325**	**8 510 310**
Moyenne des écarts en valeur absolue 2010-2018	**6,28%**	**6,70%**	**6,58%**	**5,30%**
Maximum des écarts en valeur absolue 2010-2018	**9,65%**	**9,29%**	**9,20%**	**9,11%**

TABLEAU 30 : Résultats du backtesting – Travaux de modélisation du montant moyen de VL

Conclusion

La nouvelle norme IFRS 17 encadrant la comptabilisation des contrats d'assurance (dont l'entrée en vigueur est prévue le 1er janvier 2023) introduit différentes évolutions sur le calcul des engagements, et modifie notamment la frontière des contrats en Epargne en intégrant désormais les versements libres futurs non programmés sur les contrats en portefeuille à la date de comptabilisation.
De ce fait, les compagnies d'assurance doivent désormais mettre en place des processus de calcul permettant de modéliser les comportements de versement libre des assurés en minimisant le plus possible l'erreur de prédiction. L'étude descriptive de la base de données (historique des flux de 2010 à 2017) et l'application de méthodes de classification avancées (algorithmes Machine Learning des forêts aléatoires et XGBoost) ont permis d'établir un premier classement des variables explicatives les plus déterminantes pour modéliser la fréquence de versement libre (l'algorithme XGBoost s'est révélé le plus performant). L'application des différentes méthodes de modélisation a permis d'affiner ce classement. Il en ressort que l'ancienneté est la variable la plus déterminante, suivie de l'âge et que l'utilisation de l'indice macro-économique TME (annuel) peut également s'avérer pertinente. Il convient de souligner que la qualité des prédictions est étroitement liée au volume et à la qualité des données. L'étude pourrait être étendue en identifiant d'autres variables potentiellement explicatives à analyser et en utilisant des outils informatiques de calcul plus puissants.
Enfin, nous avons mis en œuvre différents types de modélisation plus ou moins avancées pour modéliser les versements libres : des lois de versement libre construites à partir de triangles de versement aux lois construites en utilisant le modèle linéaire généralisé et son extension le modèle additif généralisé.
Le backtesting des méthodes appliquées sur l'historique 2010-2017 et sur le Model Point 2018 du modèle de projection de la compagnie a permis de faire ressortir les modèles de prédiction les plus performants :

- Pour la fréquence de versement libre par contrat, le modèle additif généralisé (GAM Logit) en prenant en compte la variable ancienneté, la variable âge en lui appliquant la fonction spline de lissage « à plaques minces » et la variable TME (annuel). Cette fonction spline de lissage semble particulièrement adaptée à la relation non linéaire qui existe entre la variable explicative âge et la variable à expliquer (fréquence de versement).
- Pour le montant moyen de versement libre par contrat, le modèle GLM Log-Normale en prenant en compte la variable âge (la variable ancienneté semble ici moins déterminante. D'autres variables explicatives pourraient être identifiées en vue d'améliorer la performance du modèle de prédiction).

Le déploiement opérationnel de cette approche est en cours sur l'ensemble des produits épargne individuelle de la compagnie et se décline en cinq étapes clés :

1. Cartographie des produits éligibles à une modélisation de type GLM ou GAM,
2. Identification et classification des variables explicatives,
3. Backtesting des méthodes de modélisation appliquées,
4. Intégration opérationnelle de la loi de versement libre considérée comme la plus performante dans le modèle interne de la compagnie,
5. Analyse des impacts de l'utilisation de cette loi sur les sorties du modèle (par exemple : impact sur la Value In Force).

La première étape du déploiement est clé dans la mesure où l'efficacité de cette approche s'est confrontée à différentes limites liées à la nature du portefeuille (par exemple : historique insuffisant, faible activité sur des produits en run-off).
Certains travaux d'actuariat récemment réalisés sur la modélisation des rachats (Bibliographie 12.) ont mis en évidence le fait que les algorithmes de Machine Learning semblent très performants pour la classification des variables explicatives, mais également pour la modélisation (construction d'une loi de rachat). Il serait opportun de vérifier ce constat pour la modélisation des versements libres.

Bibliographie

MEMOIRES

1. AIT M'BARK R. (2018), *Approche d'agrégation des contrats d'assurance sous IFRS 17*, Mémoire présenté devant l'ISFA
2. ANDRE B. (2018), *IFRS 17 : L'allocation de la CSM en P&L pour un contrat d'épargne en euros mono-support*, Mémoire présenté devant l'ISFA
3. ASSARAF K. (2020), *Modélisation des versements libres sous IFRS 17 par des méthodes de machine learning*, Mémoire présenté l'Université Paris Dauphine
4. BAILLY R. et GUEMIN N. (2019), *IFRS 17 : Interprétation de la norme, premiers résultats et leviers de pilotage pour un portefeuille dommages*, Mémoire présenté devant l'Institut du Risk Management
5. BELABED A. (2016), *Modélisation de la sinistralité atypique en RC automobile avec prise en compte des spécificités d'un versement en rente*, Mémoire présenté devant l'ISUP
6. BELLINA R. (2014), *Méthodes d'apprentissage appliquées à la tarification non-vie*, Mémoire présenté devant l'ISFA
7. BENABDELKRIM FZ. (2017), *Modélisation des versements libres en assurance-vie : utilisation de méthodes de scoring*, Mémoire présenté devant l'ISUP
8. BOUCHTA G. (2017), *Mise en œuvre de méthodes innovantes de tarification*, Mémoire présenté devant l'Université Paris Dauphine
9. BOUEDDINE M. (2013), *Assurance Automobile : Analyse de l'impact d'une variation du tarif sur le comportement des assurés lors de l'acte de souscription et de résiliation*, Mémoire présenté devant l'EURIA
10. CHASSERAY P., ELDIN G., LEFEBVRE A. (2017), *P&C Resinsurance Modeling - Pure Premium Estimation and Creation of a Reinsurance Program*, Mémoire présenté devant l'EURIA
11. COTE S. (2016), *Modèles additifs généralisés dans la modélisation de l'impact du kilométrage et de l'exposition au risque en assurance automobile*, Mémoire présenté comme exigence partielle de la Maîtrise en Mathématiques (Université du Québec à Montréal).
12. FALL ML. (2017), *Modélisation des rachats par une approche machine learning*, Mémoire présenté devant l'ENSAE
13. *GHOSN A. (2010), Efficient Thin Plate Spline Interpolation and its Application to Adaptive Optics,* Université de Linz en Autriche (JKU)
14. GUILLOT A. (2015), *Apprentissage statistique en tarification non-vie : quel avantage opérationnel ?*, Mémoire présenté devant l'ENSAE
15. JEMLI H. (2019), *Étude du risque de rachat de produits d'épargne italiens par des données agrégées et individuelles*, Mémoire présenté devant l'Université Paris Dauphine
16. KARAMOKO FOFANA CH. (2017), *Approche tarifaire des contrats collectifs Frais de Santé à l'aide des méthodes d'apprentissage*, Mémoire présenté devant l'IFSA
17. KERNEIS J. (2018), *IFRS 17 : Enjeux et application en assurance emprunteur*, Mémoire présenté devant l'ISFA
18. MECHERGUI MA. (2018), *Evaluation du capital économique sous Solvabilité 2 : Mise en place de l'approche Curve Fitting*, Mémoire présenté devant l'Université Paris Dauphine
19. LAUR M. (2018), *Anticipation des changements de notes des obligations du portefeuille d'un assureur par méthode de machine learning*, Mémoire présenté devant l'Université Paris Dauphine
20. LUO Y. (2015), *Amélioration de la modélisation de sinistres graves à l'aide d'une approche d'apprentissage*, Mémoire présenté devant l'ISFA

21. NANA NJOYA ES. (2016), *Prédiction des comportements de rachat en épargne inidviduelle : une approche machine learning*, Mémoire présenté devant l'ENSAE
22. OSSENI Z. (2014), *Optimisation de la prise en compte de la sinistralité dans la tarification Automoteur Agricole*, Mémoire présenté devant l'Université de Strasbourg
23. OTTOU P. (2017), *Méthodes d'apprentissage automatique appliquées au provisionnement ligne à ligne en assurance non-vie*, Mémoire présenté devant l'Université Paris Dauphine
24. POUNA SIEWE V. (2010), *Modèles additifs généralisés : Intérêts de ces modèles en assurance automobile*, Mémoire présenté devant l'ISUP
25. VANDAL N. (2005), *La régression non paramétrique multidimensionnelle - Théorie et application à une étude portant sur la densité mammaire*, Mémoire présenté devant la Faculté des études supérieures de l'Université Laval (Québec) dans le cadre du programme de Maîtrise en statistique pour l'obtention du grade de Maître des sciences (M.Sc.)

LIVRE

26. BEL L., DAUDIN J-J., ETIENNE M., LEBARBIER E., MARY-HUARD T., ROBIN S., VUILLET C. (2016), *Le Modèle Linéaire et ses Extensions*, Editions Ellipses (version « open source » du 14 septembre 2016 https://www6.inrae.fr/mia-paris/content/download/4281/40718/version/1/file/ModeleLineaireEt_Extensions.pdf)

SUPPORTS DE COURS

27. CHOUQUET C. (année scolaire 2009-2010), *Modèles linéaires*, Supports de cours (Laboratoire de Statistique et Probabilités - Université Paul Sabatier – Toulouse)
28. DAMAS V. (2018), *La comptabilité et le provisionnement - Éléments de comptabilité des assurances*, Supports de cours (Centre d'Etudes Actuarielles)
29. MARCOU G., JOST P. (2011), Support *« Cours de statistique »,* Université Strasbourg
30. MOUTARDE F. (2017), *Arbres de décision et forêts aléatoires*, Supports de cours (MINES ParisTech)
31. ROUVIERE L. (2015), *Analyse du modèle de régression logistique*, Supports de cours (Université Rennes 2)
32. VAUDOR L. (2015), *Classification par forêts aléatoires*, Support de cours en ligne (ENS Lyon) http://perso.ens-lyon.fr/lise.vaudor/classification-par-forets-aleatoires/

REVUES / ETUDES SCIENTIFIQUES

33. DUCHON J. (1976), I*nterpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces*. RAIRO Analyse Numérique.
34. DUTANG C. (2013), *A survey of GNE computation methods: theory and algorithms. 2013. hal-00813531.*

OUTILS INFORMATIQUES

35. Logiciel R version 3.5.1 (avec RStudio)
36. Package R 'xgboost' : BENESTY M., CANO I., CHEN K., CHO H., GENG Y., HE T., KHOTILOVICH V., LI M., LI Y., LIN M., MITCHELL R, TANG Y., TIANQUI CHEN A., XIE J. and ZHOU T. (2019), *xgboost: Extreme Gradient Boosting*. R package version 0.90.0.2. https://CRAN.R-project.org/package=xgboost
37. Package R 'pcaPP': FILZMOSER P., FRITZ H. and KALCHER K. (2018), *pcaPP: Robust PCA by Projection Pursuit*. R package version 1.9-73. https://CRAN.R-project.org/package=pcaPP
38. Package R 'caret' : KUHN M. with contributions from BENESTY M., CANDAM C., COOPER T., ENGELHARDT A., HUNT T., KEEFER C., KENTEL B., LESCARBEAU R., MAYER Z., SCRUCCA L., TANG Y.,

WESTON S., WILLIAMS A., WING J., ZIEM A and the R Core Team (2019), *caret: Classification and Regression Training*. R package version 6.0-84 https://CRAN.R-project.org/package=caret

39. Package R 'aod' : LANCELOT R., LESNOFF M. (2012), *aod: Analysis of Overdispersed Data*. R package version 1.3.1. URL http://cran.r-project.org/package=aod
40. Package R 'randomForest' : LIAW A. and WIENER M. (2002), *Classification and Regression by randomForest*. R News 2(3), 18--22. Version 4.6-14. https://cran.r-project.org/web/packages/randomForest/index.html
41. Package R 'pROC' : MULLER M., HAINARD A., LISACEK F., ROBIN X., SANCHEZ J-C., TIBERTI N., TURCK N. (2011), *pROC: an open-source package for R and S+ to analyze and compare ROC curves*. Version 1.16.2. https://cran.r-project.org/web/packages/pROC/index.html
42. Package R 'doParallel': WESTON S. and Microsoft Corporation (2019), *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.15. https://CRAN.R-project.org/package=doParallel
43. Package R 'mgcv' : WOOD S. with contributions and/or help from HORNIK K., KNEIB T., LONERGAN M., NILSSON H., PYA N., RIPLEY B and SCHEIPL F. (2019), *Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation and GAMMs by REML/PQL*. Version 1.8-29. https://cran.r-project.org/web/packages/mgcv/index.html

Annexe : traitements informatiques

Optimisation du temps de calcul avec R

Lignes de commande R :

```
> library(doParallel)

> c1 <- makePSOCKcluster(7)
> registerDoParallel(c1)
```

Application de la méthode ascendante (« forward stepwise selection » AIC)

1ère étape - Fixation des limites *mo* et *mf* d'exécution du programme d'ajout de variable par itération :

Lignes de commande R :

```
> m0 <- glm(VL_BOOL ~ 1,data = baseVL1704     , family = binomial)

> mf <- glm(VL_BOOL ~ .,data = baseVL1704     , family = binomial)
```

Remarque : l'argument « family = binomial » correspond à la fonction lien par défaut « logit » (équivalent à l'argument « family = binomial(link = "logit") »).

2nd étape - Lancement de la classification des variables avec la méthode ascendante

Ligne de commande R :

```
> step(m0, scope = list(lower = m0, upper = mf), data = baseVL1704     , direction = "forward")
```

Application de l'algorithme des forêts aléatoires

1ère approche : utilisation de la fonction randomForest() pour exécuter l'algorithme des forêts aléatoires.

Ligne de commande R :

```
> output.forest <- randomForest(VL_BOOL ~ . , data = baseVL1704     )
```

2nd approche : utilisation de la fonction "train()" pour calibrer les paramètres d'exécution des algorithmes des forêts aléatoires.

Ligne de commande R :

```
> model170420 <- train(VL_BOOL ~., data = baseVL1704     , method = "rf", trControl = trainControl("cv", number = 10))
```

Application de l'algorithme XGBoost (avec la fonction « train() »)

Ligne de commande R :

```
> model170420 <- train(VL_BOOL ~., data = baseVL1704     , method = "xgbTree", trControl = trainControl("cv", number = 10))
```

Calcul des indicateurs RMSE, AUC et traçage d'une courbe ROC (pour évaluer la performance d'un algorithme « forêts aléatoires » à partir d'un échantillon de test correspondant à 25% des données de la base initiale)

Lignes de commande R :

```
> splitIndex <- createDataPartition(baseVL1704[,"VL_BOOL"], p = .75, list = FALSE, times = 1)
> trainDF <- baseVL1704[ splitIndex,]
> testDF <- baseVL1704[ -splitIndex,]
> model170420 <- train(VL_BOOL ~., data = trainDF, method = "xgbTree", trControl = trainControl("cv", number = 10))
>
> predictions <- predict(object = model170420, testDF[,predictorsNames], type='prob')
> outcomeName <- 'VL_BOOL'
> predictorsNames <- names(baseVL1704)[names(baseVL1704) != outcomeName]
>
> predictions <- predict(object = model170420, testDF[,predictorsNames], type='prob')
> auc <- roc(ifelse(testDF[,outcomeName]=="VRAI",1,0), predictions[[2]])
> erreur <-RMSE(ifelse(testDF[,outcomeName]=="VRAI",1,0), predictions[[2]])
> auc <- roc(ifelse(testDF[,outcomeName]=="VRAI",1,0), predictions[[2]], plot=TRUE,  grid=TRUE,
print.auc=TRUE)
> par(pty="s")
> plot(auc, xlab = "1 - Specificity")
> plot(auc, legacy.axes = TRUE)
```

Construction des modèles GLM

Modèles GLM pour modéliser la fréquence de versement libre

Lignes de commande R :

```
> glm1704 <- glm(formula = VL_BOOL ~ ANCIENNETE + AGE + TME_ANNUEL, family = binomial, data = baseVL1704       )
> summary(glm1704)
Call:
glm(formula = VL_BOOL ~ ANCIENNETE + AGE + TME_ANNUEL, family = binomial, data = baseVL1704       )

Deviance Residuals:
   Min      1Q      Median      3Q     Max
-0.4125  -0.3578  -0.3300  -0.3024  2.6647

Coefficients:
                Estimate   Std. Error   z value     Pr(>|z|)
(Intercept)    -2.6418484  0.0143054 -184.676  <2e-16 ***
ANCIENNETE     -0.0314235  0.0007271  -43.218  <2e-16 ***
AGE             0.0021520  0.0002390    9.004  <2e-16 ***
TME_ANNUEL      0.0108243  0.0043033    2.515   0.0119 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 422888 on 1002459 degrees of freedom
Residual deviance: 420753 on 1002456 degrees of freedom
AIC: 420761
Number of Fisher Scoring iterations: 5
```

Réalisation du test de Wald (GLM)

Lignes de commande R :

```
> wald.test(b = coef(glm1704), Sigma = vcov(glm1704), Terms = 2:4)
Wald test:
----------
Chi-squared test:
X2 = 2071.0, df = 3, P(> X2) = 0.0
> wald.test(b = coef(glm1704), Sigma = vcov(glm1704), Terms = 3:4)
Wald test:
----------
Chi-squared test:
X2 = 90.6, df = 2, P(> X2) = 0.0
> wald.test(b = coef(glm1704), Sigma = vcov(glm1704), Terms = 4:4)
Wald test:
----------
Chi-squared test:
X2 = 6.3, df = 1, P(> X2) = 0.012
```

Calcul d'intervalle de confiance (GLM)

Ligne de commande R :

```
> confint.default(glm1704)
                    2.5 %      97.5 %
(Intercept)   -2.669886411 -2.613810464
ANCIENNETE    -0.032848599 -0.029998463
AGE            0.001683592  0.002620464
TME_ANNUEL     0.002390104  0.019258579
```

Modèles GLM pour modéliser le montant moyen de versement libre

Lignes de commande R :

```
> glm1704 <- glm(formula = VL ~ AGE, family = Gamma(link ="log"), data = baseVL1704       )
> glm1704 <- glm(formula = VL ~ AGE, family = gaussian(link ="log"), data = baseVL1704       )
```

Construction des modèles GAM (pour modéliser la fréquence de versement libre)

Ligne de commande R :

```
> gamVL1704 <- gam(VL_BOOL ~ ANCIENNETE + s(AGE, k=10, bs="cr") + TME_ANNUEL, data = baseVL1704     , family = binomial)
```

Ligne de commande R pour afficher les coefficients (9 degrés de liberté pour la fonction spline) :

```
> coef(gamVL1704)
  (Intercept)    ANCIENNETE    TME_ANNUEL      s(AGE).1      s(AGE).2      s(AGE).3      s(AGE).4      s(AGE).5      s(AGE).6
-2.543950926 -0.030661109  0.007850484 -0.279464777  0.022500923 -0.208759339 -0.072448492  0.137323566 -0.263809365
     s(AGE).7      s(AGE).8      s(AGE).9
-0.173326038 -1.239135292 -2.405255032
```