Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du
Diplôme d'Actuaire EURIA
et de l'admission à l'Institut des Actuaires

le 7 Septembre 2022

Par : Marine Dulaurans
Titre : Modélisation des versements libres en Retraite individuelle dans un contexte IFRS 17

Confidentialité : Oui 2 ans

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

Membre présent du jury de l'Institut des Actuaires :
Marine HABART
Jérôme VIGNANCOUR

Signature :

Membres présents du jury de l'EURIA :
Françoise PENE

Entreprise :
GENERALI France
Signature :

Directeur de mémoire en entreprise :
Yoann Gouyen
Signature :

Invité :

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 du candidat :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

Ce mémoire porte sur la modélisation des versements libres en Retraite individuelle dans un contexte IFRS 17. Un versement libre est un versement qui peut être effectué à tout moment au cours de la vie du contrat contrairement aux primes uniques ou périodiques. Il est par nature aléatoire et donc complexe à modéliser. Or l'entrée en vigueur de la norme IFRS 17 change la définition de la frontière des contrats et rend obligatoire l'intégration de ces versements dans les projections des flux best estimate.

Une première approche vise à construire des lois de versements libres par rapport à l'assiette de provisions mathématiques. L'idée a été de construire une modélisation simple des versements libres facilement implémentable dans les outils de projection de best estimate de Generali. Plus précisément, les taux sont obtenus en calculant le rapport entre les montants de versements libres et l'assiette de provisions mathématiques. Ces taux sont construits selon une maille d'agrégation qui a été établie grâce à une analyse descriptive des données. Cette approche permet de mesurer l'enjeu financier de l'intégration des versements libres dans l'estimation du best estimate.

Ensuite, les lois construites seront intégrées dans les modèles sous Prophet et permettront de projeter les versements libres. Leur impact sera analysé dans le bilan et le compte de résultat en norme IFRS 17. Des indicateurs tels que le Best Estimate ou la PVFP seront ainsi observés.

Pour compléter cette première approche, une modélisation plus fine des versements libres est mise en place afin de comprendre les facteurs prépondérants influençant le déclenchement d'un versement libre. Plusieurs méthodes de modélisation seront testées avec différents niveaux d'interprétabilité. Ce mémoire recourra également à l'utilisation de méthodes de data augmentation car les versements libres en retraite ont une fréquence assez faible ce qui rend leur modélisation difficile.

Mots clefs: Retraite individuelle, IFRS 17, Frontière des contrats, Versements libres, Prévision, Machine Learning, Passifs du Bilan économique

Abstract

This essay deals with the modeling of future flexible premiums in individual retirement in an IFRS 17 context. A future flexible premium is a payment that can be made at any time during the life of the contract unlike single or periodic premiums. It is random by nature and therefore complex to predict. However, the coming into effect of IFRS 17 changes the definition of the contract boundary and makes it mandatory to include these payments in the best estimate flow projections.

A first approach aims at building laws of future flexible premiums in relation to the base of mathematical reserves. The idea was to create a simple model of future flexible premiums that could be easily implemented in Generali's best estimate projection tools. More precisely, the rates are obtained by calculating the ratio between the amounts of future flexible premiums and the mathematical reserves base. These rates are constructed according to an aggregation grid that has been established through a descriptive analysis of the data. This approach makes it possible to measure the financial impact of including future flexible premiums in the best estimate.

Then, the laws built will be entered into the models in Prophet and will make it possible to forecast the future flexible premiums. Their impact will be analyzed in the balance sheet and the income statement under IFRS 17. Indicators such as the Best Estimate or the PVFP will be observed.

To complete this first approach, a more complex modeling of the future flexible premiums is set up in order to understand the predominant factors influencing the triggering of a future flexible premium. Several modeling methods will be tested with different levels of interpretability. This essay will also use data augmentation methods because future flexible premiums in retirement have a relatively low frequency which makes their modeling difficult.

Keywords: Individual Retirement, IFRS 17, Contract Boundary, future flexible premiums, Forecasting, Machine Learning, Economic Balance Sheet Liabilities

Note de synthèse

La publication de la norme comptable internationale IFRS 17 donne une nouvelle définition de la frontière des contrats. La norme sera applicable dès janvier 2023 et aura plusieurs conséquences sur les passifs du bilan comptable. La redéfinition de la frontière des contrats d'assurance implique l'intégration de nouveaux éléments comptables. Ainsi, les versements libres doivent désormais être comptabilisés dans les flux Best Estimate, ce qui n'était jusqu'à présent pas le cas sous Solvabilité II.

Les versements libres sont des versements qui peuvent être effectués à tout moment dans la phase de capitalisation des contrats. Ils sont donc par nature aléatoires et dépendent de la volonté des assurés de les réaliser. Leur intégration dans la frontière des contrats représente donc un réel enjeu pour les compagnies d'assurance.

Ce mémoire porte ainsi sur la compréhension et la modélisation des versements libres. Les travaux se sont orientés sur le périmètre de la retraite individuelle. En effet, des études internes à Generali avaient déjà été construites sur le périmètre de l'épargne. L'idée était donc de pouvoir en faire autant sur le périmètre de la retraite. Les produits concernés amènent ainsi des réflexions intéressantes sur divers sujets comme la réalisation de la projection de ces montants dans les outils internes ou encore la gestion de données déséquilibrées dans le cadre de modélisation par des méthodes de machine learning.
Seuls les produits individuels ont été retenus car ils concentrent le plus de versements libres. Ils offrent également des perspectives de compréhension des comportements des assurés à réaliser ou non un versement libre.

Approches explorées afin d'appréhender les versements libres

Les versements libres ont été abordés au travers de deux approches principales.
La première est une approche simplifiée visant à construire des taux de versements libres grâce à des calculs utilisant l'historique disponible dans les bases de données récupérées. Cette approche se veut simple dans sa réalisation afin de pouvoir analyser les impacts dans les outils internes de Generali. De plus, cette méthode présente un avantage opérationnel car les mises à jour des taux et leur intégration dans Prophet sont relativement simples d'une année à l'autre.
La seconde est une approche visant à modéliser les versements libres grâce à des méthodes

de *machine learning*. A travers une modélisation plus fine, l'objectif sera de comprendre les facteurs influençant la décision d'investissements via des versements libres sur les contrats retraite. Nous serons en mesure de critiquer les choix d'implémentation faits pour la méthode plus simple mais déjà opérationnelle, et de faire des propositions d'amélioration.

Les données

Plusieurs sources de données ont été exploitées afin de répondre aux différentes approches envisagées.

La première approche est volontairement simplifiée et utilise une segmentation sur quelques critères. Les principales informations retenues sont les montants de versements libres et de provisions mathématiques ainsi que des renseignements sur les produits concernés avec les supports d'investissements, l'ancienneté des contrats, les taux garantis et les taux de frais de gestion. Ces éléments ont été récupérés via des bases internes et ce sur un historique compris entre 2015 et 2021.

La seconde approche s'est basée sur les travaux précédents. Toutefois, de nouvelles variables ont été ajoutées afin de réussir à comprendre les investissements par versements libres. Ainsi, des données concernant les assurés ainsi que des données externes ont permis d'étoffer notre analyse. Ces trois types de variables sont synthétisés ci-dessous :

- Les données contrats : les produits, le support d'investissement, la fiscalité, le motif d'entrée du contrat, le taux d'intérêt technique, le taux de frais de gestion, le fractionnement et le montant des primes périodiques, le montant de versements libres et de provisions mathématiques, le montant de versements libres de l'année précédente, l'année du versement et l'ancienneté du contrat ;
- Les données assurés : les âges atteints au moment du versement, le sexe, le département de résidence, le panier de contrats que le client possède chez Generali, le revenu moyen de la commune de résidence, la tranche d'unité urbaine ;
- Les données externes : le taux de livret A, le taux moyen du PEL, le taux d'inflation, le niveau de PIB, le cours du CAC 40.

Approche simplifiée de construction des taux de versements libres

Les taux de versements libres ont été construits en suivant les travaux réalisés à ce sujet pour le périmètre de l'épargne. Cette méthodologie a été suivie afin de permettre d'intégrer ces travaux dans l'outil Prophet nécessaire à la projection des flux futurs.

Ainsi, les données ont été agrégées selon la granularité produit et ancienneté dans cette logique. Des statistiques descriptives ont été réalisées afin de définir de potentiels regroupements entre les produits afin de prédire au mieux les tendances de versements

libres. Des courbes de taux ont ainsi été construites pour chaque produit. Ces études ont permis de construire neuf lois de versements libres différentes.

Les taux ont été calculés grâce aux formules suivantes :

$$\forall\ N \neq 0\ alors\ VL_N\ =\ \frac{Montant\_Versement\_Libre_N}{Montant\_PM\_Ouverture_N} \tag{1}$$

$$Pour\ N = 0\ alors\ VL_0\ =\ \frac{Montant\_Versement\_Libre_0}{Montant\_PM\_Ouverture_1 - Montant\_Versement\_Libre_0} \tag{2}$$

où N est l'ancienneté du contrat
où le montant de PM d'ouverture N correspond au montant de PM de clôture N-1

L'outil Prophet impose d'avoir des taux de versements libres sur l'intégralité de l'horizon de projection. En ce sens, les taux ont été extrapolés à l'aide d'une méthode de moyenne mobile jusqu'à l'ancienneté demandée. Des taux à 0% ont également été imposés pour les anciennetés des versements n'admettant plus d'assurés. Par exemple, certains produits sont en *run-off* depuis 2015, ils ne doivent donc plus enregistrer de versements libres sur les sept premières anciennetés. Les graphes de la figure 1 ont ainsi été produits :

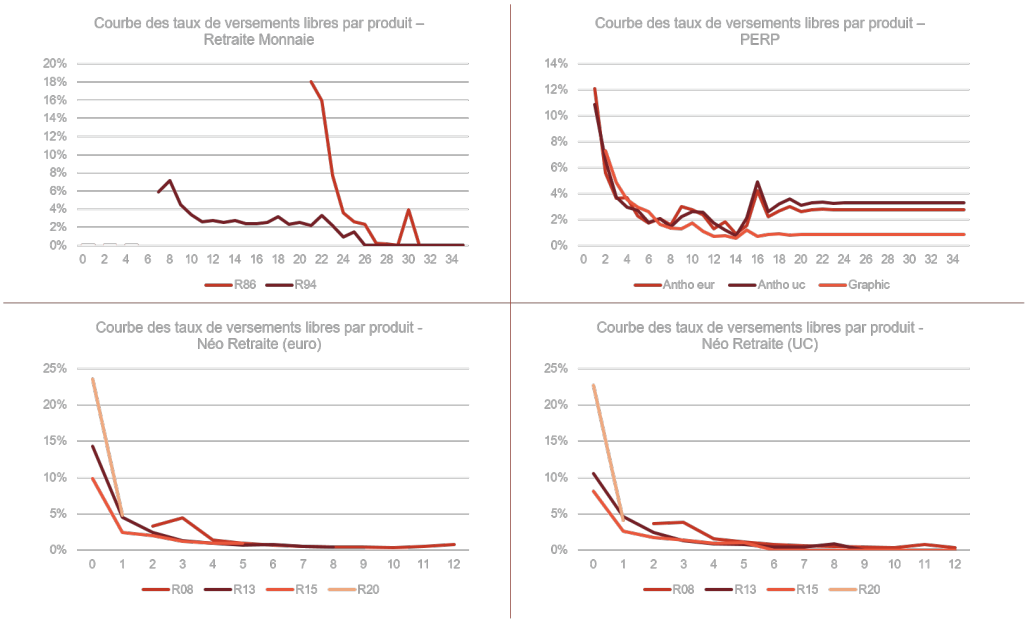


FIGURE 1 – Courbe des taux de versements libres par produit

Ces différentes courbes ont permis de réaliser plusieurs analyses :

— Des tendances de versements différentes ont été perçues entre les produits retraite.
— Les assurés réalisent davantage de versements sur les supports euro. Ils adoptent ainsi des comportements plus prudents. Ce constat était inversé sur les produits de l'épargne.
— Les taux sont décroissants par rapport à l'ancienneté.

Les taux de versements libres ont ensuite été renseignés dans Prophet. Cependant, l'implémentation de la variable nécessaire à la projection des versements libres n'existait pas encore dans l'outil. J'ai donc débuté des travaux afin que cela soit possible. Différentes méthodes ont été mises en place afin d'intégrer ces versements. La compréhension de l'environnement Prophet ainsi que la validation des codes réalisés se sont avérées complexes. Les travaux ont été validés et des résultats ont ainsi pu être extraits de l'outil.

Les montants des primes brutes ont été recueillis afin d'analyser les prédictions des montants de versements libres sur l'année 2021. Les résultats de la figure 2 ci-dessous ont ainsi été obtenus :

Produits retenus	Prévisions 2021	Observations 2021	Observations 2020	Observations 2019	Observations 2018	Observations 2017	Observations 2016	Observations 2015
Anthologie	1 745K€	11 237K€	7 824K€	2 979K€	3 622K€	4 921K€	4 676K€	5 340K€
Graphic	147K€	180K€	199K€	157K€	159K€	209K€	253K€	411K€
R08	5 881K€	2 959K€	2 976K€	3 796K€	4 433K€	1 252K€	789K€	1 738K€
R13	3 359K€	1 312K€	1 477K€	1 791K€	2 174K€	641K€	60K€	1 160K€
R15	4 255K€	2 330K€	3 020K€	3 475K€	4 114K€	3 531K€	1 265K€	373K€
R20	2 674K€	9 046K€	3 547K€	204K€				
R80	0K€				1K€			
R86	144K€	124K€	138K€	296K€	268K€			
R94	12 894K€	4 746K€	5 568K€	5 933K€	7 241K€			
	31 100K€	31 933K€	24 749K€	18 630K€	22 012K€	10 555K€	7 044K€	9 022K€

FIGURE 2 – Historique des versements libres ainsi que les prédictions obtenues sur 2021

Les montants ont été modifiés pour des raisons de confidentialité sans pour autant nuire aux analyses principales. On constate que les prédictions globales sont très satisfaisantes. Néanmoins, des compensations s'opèrent entre les différents produits.
Anthologie est un PERP qui a enregistré beaucoup de versements libres au moment de l'annonce de la la fin de souscription en 2019. Ces versements ont été effectués et encouragés par la loi PACTE afin d'anticiper les transferts vers les nouveaux PER. Cette dynamique n'a pas été intégrée dans les lois de versements libres construites.
Certains produits La Retraite sont quant à eux surestimés. Ce phénomène provient de la méthode de calcul des taux de versements libres. En effet, ces produits ne sont plus commercialisés et connaissent une décroissance des versements libres dans le temps. Cette baisse n'a pas été suffisamment captée par les lois construites. En effet, la loi empirique construite ne prend pas en compte de composante 'tendance' comme on peut l'avoir sur des modèles plus complexes.

Par la suite, l'impact de ces montants a été étudié au travers du Best Estimate et de la Present Value of Future Profits. Les résultats de la figure 3 ont ainsi été obtenus.

TOTAL + VL_RET - TOTAL			
Portefeuille IFRS 17	BE	PVFP	MV
Produits La Retraite	-40 587 125	40 407 430	-179 695
PERP Grafic	310 690	-310 690	0
PERP Anthologie	-1 050 229	1 048 507	-1 722
TOTAL La Retraite	-40 595 220	40 415 526	-179 695
TOTAL PERP	-739 341	919 035	179 695
TOTAL	-41 334 561	41 334 561	0

FIGURE 3 – Présentation des impacts de l'intégration des versements libres

Les montants obtenus ont été validés et jugés cohérents par rapport aux attentes. Néanmoins, cette étude a permis de constater que l'utilisation de la variable ancienneté était parfois inadaptée et pouvait conduire à surestimer ou sous-estimer certains montants de versements libres. C'est donc en ce sens que les études de modélisation par des méthodes de *machine learning* ont été développées. L'objectif était d'analyser les variables les plus susceptibles d'influencer la réalisation d'un versement libre.

Approche de modélisation par des méthodes de *machine learning*

Le problème s'est décliné selon deux méthodes : une classification et une régression. Les deux approches ont été testées et comparées afin de vérifier si les deux modèles convergeaient en termes de comportements. Ainsi, deux méthodes de modélisation ont été envisagées :

— Une classification binaire de la survenance des versements libres
— Un modèle de fréquence du nombre de versements libres

Les performances du premier modèle ont été calculées à partir du taux de rappel et de précision et du F1 score car les données sont déséquilibrées. Le second modèle a été prédit à l'aide du RMSE et du MAE.

Pour ces deux méthodes, les algorithmes de GLM pénalisé, d'arbre de décision, de forêt aléatoire et d'histogram gradient boosting ont été retenus. Les performances de ces modèles ont été comparées et ont permis de privilégier pour les deux méthodes l'algorithme d'histogram gradient boosting. Les hyperparamètres de ces modèles ont été optimisés afin d'améliorer les taux d'erreur. Finalement, les résultats suivants ont été constatés :

Pour le modèle de classification :

Taux de précision	0,84
Taux de rappel	0,35
F1 score	0,493
Accuracy	0,97
AUC	0,87

Pour le modèle de fréquence :

RMSE	0,28
MAE	0,075

Compréhension des modèles

Par la suite, une analyse du comportement des modèles a été réalisée grâce à l'algorithme SHAP. Les variables importantes ont ainsi été identifiées sur les deux modèles retenus.

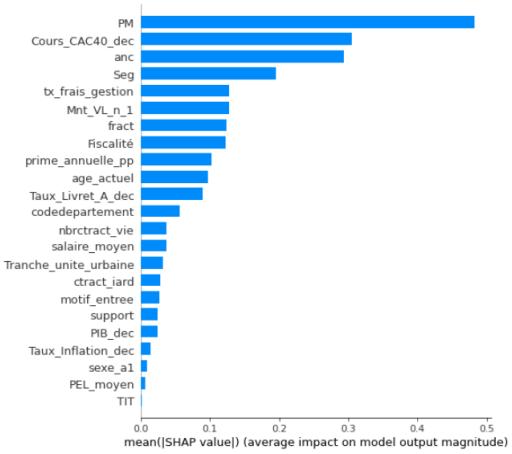


FIGURE 4 – Importances moyennes des variables du premier modèle

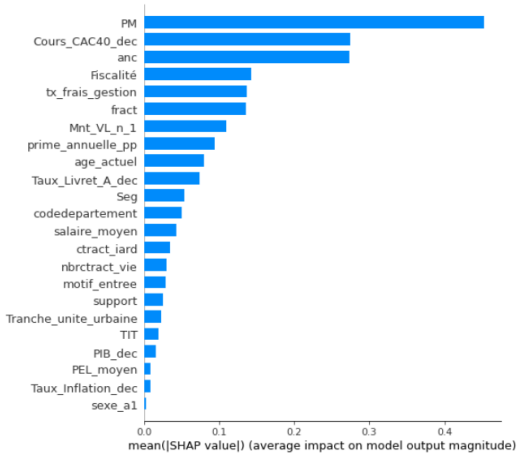


FIGURE 5 – Importances moyennes des variables du second modèle

Ces résultats ont identifié les mêmes trois premières variables nécessaires à la prédiction des versements libres. Ainsi, la provision mathématique, le cours du CAC 40 et l'ancienneté des contrats jouent un rôle important dans la prédiction des versements libres. Les résultats des différents modèles ont tous convergé vers la même sélection de variables. Cela a permis de consolider les résultats de ce mémoire.

Impact de la data augmentation

Pour terminer cette étude, des méthodes de rééchantillonnage ont été abordées sur le modèle de classification. En effet, l'idée était de vérifier la sensibilité des analyses à une étape de data augmentation.
Ainsi, des méthodes d'*oversampling* et d'*undersampling* ont été associées en ce sens. Les performances ont été améliorées, comme en témoigne le tableau 6.

Le taux de rappel a été amélioré au détriment du taux de précision. Cependant, le F1 score a été amélioré de 3,4%. Les analyses extraites de ce modèle rééchantilloné ont été sensiblement les mêmes que pour le modèle initial ce qui nous a conforté à propos de la robustesse des analyses. Les conclusions se concentreront donc sur les modèles précédents.

Taux de précision	0,57
Taux de rappel	0,49
F1 score	0,527
AUC	0,86

FIGURE 6 – Performances du modèle après rééchantillonnage

Conclusions de l'étude

Un travail d'explicabilité des modèles a été réalisé à l'aide de l'algorithme SHAP. Cette méthode dite agnostique (indépendante de l'aglorithme expliqué) fournit des outils pour comprendre quelles sont les variables les plus importantes pour les prédiction du modèle. SHAP informe également sur la compréhension du sens dans lequel les variables influencent les prédictions. Les conclusions suivantes ont ainsi été réalisées :

- Un montant élevé de provisions mathématiques encourage la réalisation de versements libres. Cette variable était déjà utilisée dans le calcul des taux de versements libres dans Prophet. Ce résultat vient donc confirmer sa nécessité.
- Un contexte financier moins favorable favorise les investissements sur les produits retraite. Un contexte de marché actions défavorable peut encourager les investissements sur des taux garantis. Certains assurés agissent donc en connaissance d'un contexte économique particulier. Les versements libres conjoncturels ont de ce fait réussi à être identifiés au travers de ces modèles.
- L'ancienneté est apparue très nettement dans les variables importantes. Les anciennetés plus faibles contribuent à la réalisation de versements libres.
- D'autres variables telles que la segmentation ou la fiscalité des produits, le fractionnement et le montant des primes périodiques, le montant du versement libre de l'année précédente et le taux de frais de gestion sont également revenues fréquemment dans les modèles. Ces variables jouent donc un rôle dans les investissements par versements libres.

Finalement, ce mémoire a permis dans un premier temps de confirmer le bien fondé des hypothèses prises pour modéliser les versements libres dans un outil Prophet qui est déjà opérationnel. Il a permis d'identifier plus finement les facteurs influençant le déclenchement d'un versement libre. Ainsi, nous avons par exemple vu que l'environnement économique, le niveau de primes périodiques programmées, les habitudes passées de versements libres avaient un impact non négligeable.
Dans la prolongation de ce mémoire, la question se posera de voir comment améliorer les projections de versements libres dans l'outil Prophet compte tenu de ces conclusions. D'ores et déjà, une piste pourrait être de prendre en compte l'impact des marchés financiers à travers les scénarios économiques déjà gérés dans l'outil Prophet.

Synthesis note

The publishing of the international accounting standard IFRS 17 gives a new definition of the contract boundary. The standard will be applicable as of January 2023 and will have several consequences on the liabilities of the accounting balance sheet. The redefinition of the boundary of insurance contracts implies the integration of new accounting elements. Thus, future flexible premiums must now be accounted for in the Best Estimate flows, which was not the case under Solvency II.

The future flexible premiums are payments that can be made at any time during the capitalization period of the contracts. They are therefore random by nature and depend on the policyholders' willingness to make them. Their integration in the contract boundery is therefore a real challenge for insurance companies.

This essay focuses on the understanding and modeling of future flexible premiums. The research focused on the individual retirement scope. Indeed, internal studies at Generali had already been carried out on the savings scope. The idea was to be able to do the same for retirement. The products concerned lead to interesting considerations on various subjects such as the projection of these amounts in internal tools or the management of unbalanced data in the context of modeling by machine learning methods.
Only individual products were selected because they concentrate the most future flexible premiums. They also offer perspectives for understanding the behavior of policyholders in making a future flexible premium.

Approaches explored to understand future flexible premiums

The future flexible premiums have been addressed through two main approaches.
The first is a simplified approach that aims to build up the rates of future flexible premiums through calculations using the historical data available in the recovered databases. This approach is intended to be simple in its implementation so that the impacts can be analyzed in Generali's internal tools. Moreover, this method has an operational advantage because rate updates and integration into Prophet are relatively simple from year to year.
The second approach aims to model future flexible premiums using machine learning methods. Through a more detailed modeling, the objective will be to understand the

factors influencing the investment decision via future flexible premiums on retirement contracts. We will be able to criticize the implementation choices made for the simpler but already operational method, and to make proposals for improvement.

The data

Several data sources were used in order to respond to the different approaches considered.

The first approach is deliberately simplified and uses a segmentation on a few criteria. The main information used was the amounts of future flexible premiums and mathematical reserves, as well as information on the products concerned, including the investment supports, the age of the contracts, the guaranteed rates and the management fee rates. These elements were retrieved from internal databases for a period of time between 2015 and 2021.

The second approach was based on the previous work. However, new variables were added in order to gain insight into future flexible premiums. Thus, policyholder and external data were added to our analysis. These three types of variables are summarized below :

- The product data : the products, the investment support, the taxation, the reason of entrance of the contract, the technical interest rate, the rate of management fees, the frequency and the amount of the periodic premiums, the amount of future flexible premiums and mathematical reserves, the amount of future flexible premiums of the previous year, the year of the payment and the age of the contract ;
- The policyholder data : ages reached at the time of payment, gender, department of residence, all the contracts that the client has with Generali, average income of the town of residence, the urban unit range ;
- The external data : the Livret A rate, the average PEL rate, the inflation rate, the GDP level, the CAC 40 share price.

Simplified approach to building future flexible premium rates

The rates of future flexible premiums were constructed following the work done on this subject for the savings scope. This method was used in order to integrate this work into the Prophet tool, which is required to project future cash flows.

Thus, the data were aggregated according to product and age granularity in this way. Descriptive statistics were performed in order to define potential groupings between products in order to better predict trends in future flexible premiums. Rate curves were thus created for each product. These studies made it possible to construct nine different future flexible premium laws.

The rates were calculated using the following formulas :

$$\forall\ N \neq 0\ so\ FP_N\ =\ \frac{Amount\_Future\_Flexible\_Premium_N}{Amount\_Opening\_Reserve_N} \tag{3}$$

$$For\ N = 0\ so\ FP_0\ =\ \frac{Amount\_Future\_Flexible\_Premium_0}{Amount\_Opening\_Reserve_1 - Amount\_Future\_Flexible\_Premium_0} \tag{4}$$

where N is the age of the contract
where the opening mathematical reserve amount N corresponds to the closing mathematical reserve amount N-1

The Prophet tool requires that future flexible premium rates be available over the entire projection horizon. To do so, the rates were extrapolated using a moving average method up to the requested age. Rates at 0% were also imposed for the ages of premiums that no longer admit insureds. For example, some products have been in run-off since 2015, so they should no longer record future flexible premiums on the first seven vesting periods. The graphs in figure 7 were thus produced :

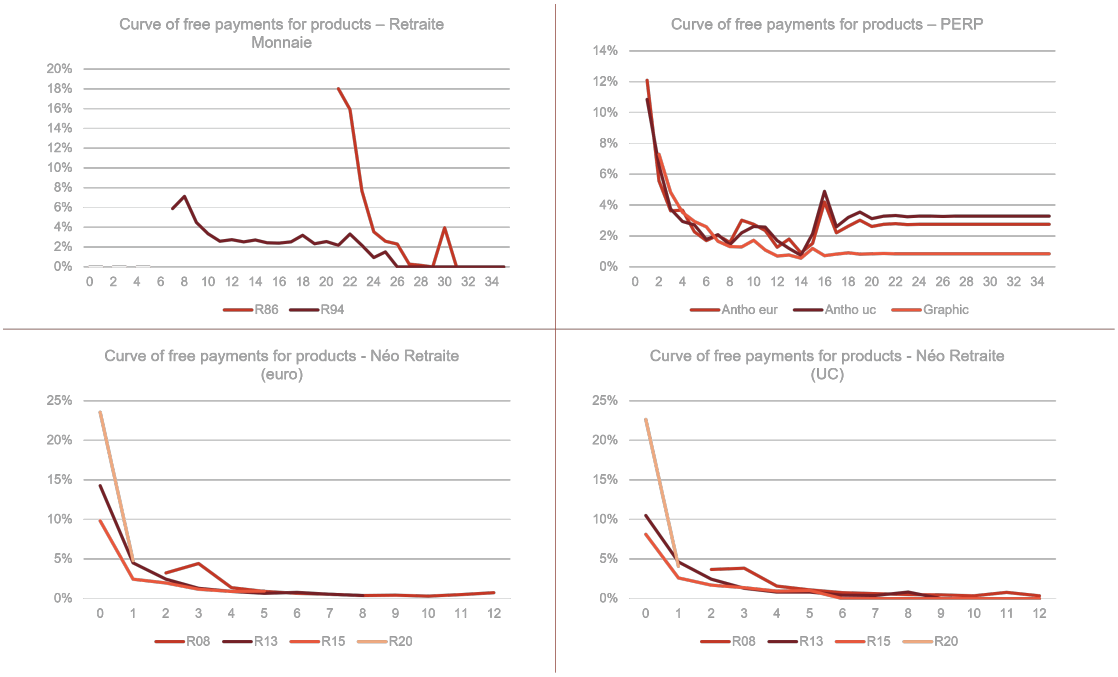


FIGURE 7 – Curve of the rates of future flexible premiums by product

These different curves allowed us to carry out several analyses :

— Different payout patterns were seen between retirement products.
— Policyholders are making more payments into euro funds. They are thus adopting a more cautious attitude. This observation was reversed for savings products.
— The rates decrease with age.

The future flexible premium rates were then entered into Prophet. However, the implementation of the variable necessary for the projection of the future flexible premiums did not yet exist in the tool. I therefore started work to make it possible. Different methods were put in place to integrate these payments. The understanding of the Prophet environment as well as the validation of the realized codes proved to be complex. The work has been validated and results have been extracted from the tool.

Gross premium amounts were collected to analyze the predictions of future flexible premiums over the year 2021. The results in figure 8 below were thus obtained :

Selected Products	Forecast 2021	Observations 2021	Observations 2020	Observations 2019	Observations 2018	Observations 2017	Observations 2016	Observations 2015
Anthologie	1 745K€	11 237K€	7 824K€	2 979K€	3 622K€	4 921K€	4 676K€	5 340K€
Graphic	147K€	180K€	199K€	157K€	159K€	209K€	253K€	411K€
R08	5 881K€	2 959K€	2 976K€	3 796K€	4 433K€	1 252K€	789K€	1 738K€
R13	3 359K€	1 312K€	1 477K€	1 791K€	2 174K€	641K€	60K€	1 160K€
R15	4 255K€	2 330K€	3 020K€	3 475K€	4 114K€	3 531K€	1 265K€	373K€
R20	2 674K€	9 046K€	3 547K€	204K€				
R80	0K€				1K€			
R86	144K€	124K€	138K€	296K€	268K€			
R94	12 894K€	4 746K€	5 568K€	5 933K€	7 241K€			
	31 100K€	31 933K€	24 749K€	18 630K€	22 012K€	10 555K€	7 044K€	9 022K€

FIGURE 8 – Historic of future flexible premiums and predictions obtained on 2021

The amounts have been changed for confidentiality reasons without affecting the main analyses. We can see that the overall predictions are very satisfactory. Nevertheless, there are compensations between the different products.
Anthologie is a PERP that recorded a lot of future flexible premiums at the time of the announcement of the end of subscription in 2019. These payments were made and encouraged by the PACTE law in order to anticipate transfers to the new PER. This dynamic was not incorporated in the future flexible premium laws built.
Some La Retraite products are overestimated. This phenomenon is due to the method used to calculate the rates of future flexible premiums. These products are no longer being marketed and are experiencing a decline in future flexible premiums over the years. This decline has not been sufficiently captured by the laws that have been drafted. Indeed, the empirical law constructed does not take into account a 'trend' component as it can be found in more complex models.

The impact of these amounts was then studied through the Best Estimate and the Present Value of Future Profits. The results shown in figure 9 were thus obtained.

TOTAL + VL_RET - TOTAL			
IFRS 17 Portfolio	BE	PVFP	MV
La Retraite Products	- 40 587 125	40 407 430	- 179 695
PERP Graphic	310 690	- 310 690	-
PERP Anhtologie	- 1 050 229	1 048 507	- 1 722
TOTAL La Retraite	- 40 595 220	40 415 526	- 179 695
TOTAL PERP	- 739 341	919 035	179 695
TOTAL	- 41 334 561	41 334 561	-

FIGURE 9 – Presentation of the impacts of the integration of future flexible premiums

The amounts obtained were validated and deemed consistent with the expectations. Nevertheless, this study showed that the use of the seniority variable was sometimes inappropriate and could lead to over/underestimating some amounts of future flexible premiums. It is therefore in this sense that modeling studies using machine learning methods were developed. The objective was to analyze the variables most likely to influence the realization of a future flexible premium.

Modeling approach using machine learning methods

The problem was addressed using two methods : a classification and a regression. The two approaches were tested and compared in order to verify whether the two models converged in terms of behavior. Thus, two modeling methods were considered :

— A binary classification of the occurrence of future flexible premiums
— A frequency model for the number of future flexible premiums

The performance of the first model was calculated from the recall and precision rate and the F1 score as the data were unbalanced. The second model was predicted using RMSE and MAE.

For both methods, the penalized GLM, decision tree, random forest and histogram gradient boosting algorithms were selected. The performances of these models have been compared and have led to the preference for the histogram gradient boosting algorithm for both methods. The hyperparameters of these models were optimized in order to improve the error rates. Finally, the following results were found :

For the classification model :

Precision rate	0,84
Recall rate	0,35
F1 score	0,493
Accuracy	0,97
AUC	0,87

For the frequency model :

RMSE	0,28
MAE	0,075

Understanding the models

Then, an analysis of the behavior of the models was performed using the SHAP algorithm. The important variables were thus identified in the two models selected.

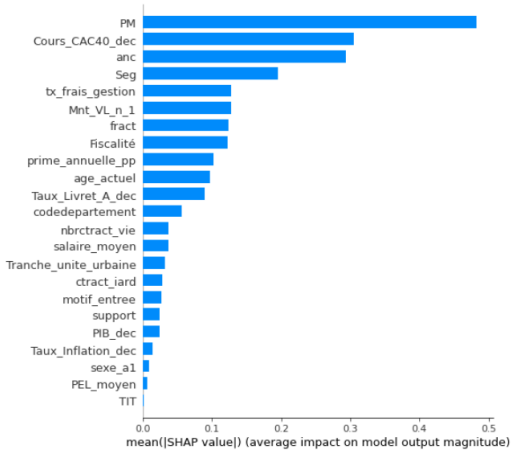


FIGURE 10 – Average importance of variables in the first model

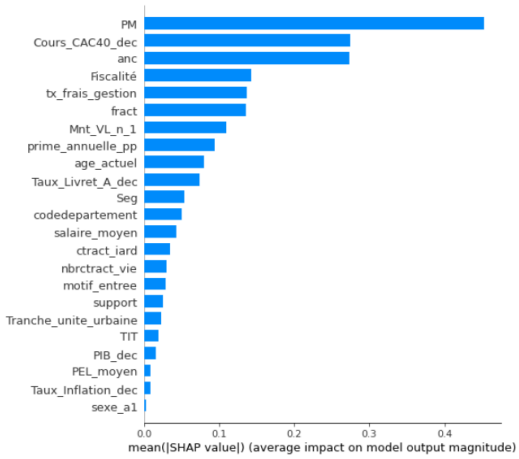


FIGURE 11 – Average importance of variables in the second model

These results identified the same first three variables necessary to predict future flexible premiums. Thus, the mathematical reserve, the CAC 40 stock price and the age of the contracts play an important role in the prediction of future flexible premiums. The results of the different models all converged towards the same selection of variables. This has allowed us to consolidate the results of this essay.

Impact of data augmentation

To finish this study, resampling methods have been discussed on the classification model. Indeed, the idea was to check the sensitivity of our analyses to a data augmentation step
Thus, oversampling and undersampling methods have been associated in this sense. The performances were improved, as shown in the table 12.

The recall rate was improved at the detriment of the precision rate. However, the F1 score was improved by 3.4%. The analyses extracted from this resampled model were essentially the same as for the original model which confirmed the robustness of the analyses. The conclusions will therefore focus on the previous models.

Precision rate	0,57
Recall rate	0,49
F1 score	0,527
AUC	0,86

FIGURE 12 – Performance of the model after resampling

Conclusions of the study

A work of explicability of the models was carried out using the SHAP algorithm. This so-called agnostic method (independent of the explained aglorithm) provides tools to understand which variables are the most important for the predictions of the model. SHAP also informs the understanding of the direction in which the variables influence the predictions. The following conclusions were reached :

- A high level of mathematical reserves encourages the realization of future flexible premiums. This variable was already used in the calculation of the rates of future flexible premiums in Prophet. This result confirms its necessity.
- A less favorable financial environment favors investments in retirement products. An unfavorable equity market environment may encourage investments in guaranteed rates. Some policyholders are therefore operating with knowledge of a particular economic environment. The cyclical future flexible premiums have therefore been successfully identified through these models.
- Seniority appeared very clearly in the important variables. Lower tenure levels contribute to future flexible premiums.
- Other variables such as product segmentation or taxation, frequency and periodic premium amounts, the amount of the previous year's future flexible premium, and the management fee rate were also frequently included in the models. These variables therefore play a role in the free-payment investments.

Finally, this essay has allowed us to confirm the validity of the hypotheses taken to model future flexible premiums in a Prophet tool that is already operational. It has allowed us to identify in greater detail the factors that influence the triggering of a future flexible premium. For example, we have seen that the economic environment, the level of scheduled periodic premiums and past habits of future flexible premium have a significant impact.

As an extension of this essay, the question will arise as to how to improve the projections of future flexible premiums in the Prophet tool in light of these conclusions. One possibility would be to take into account the impact of the financial markets through the economic scenarios already managed in the Prophet tool.

Remerciements

Je tiens tout d'abord à remercier mon tuteur d'alternance Yoann GOUYEN, responsable d'études actuarielles. Son encadrement et ses conseils m'ont beaucoup appris. Il a su me mettre en confiance et se rendre disponible afin que la réalisation de mes missions se passe bien. Je le remercie également pour les relectures qu'il a pu faire pour mon mémoire et le rapport d'alternance ainsi que pour ses retours toujours constructifs.

Je souhaite ensuite adresser mes remerciements à Etienne GUILLOU, responsable de la plateforme actuariat de Nantes et manager de l'équipe Wiz'You, qui a initié mon sujet d'alternance. Je le remercie pour sa disponibilité et ses conseils sur les différents sujets que j'ai pu aborder.

J'aimerais remercier Hakim M'MADI, actuaire au sein de l'équipe IFRS 17, et Cédrik PERINA, consultant senior en actuariat, qui m'ont apporté leur aide pour ma mission au sein de leur équipe. Ils ont répondu à mes questions sur le périmètre de la retraite ainsi que sur Prophet et ont contribué au bon déroulé de ma mission. Je les remercie pour leur disponibilité ainsi que leur bonne humeur au cours des réunions.

Je saisis également cette occasion pour remercier Vincent SADE, chargé d'études actuarielles à la TA Retraite, Maxime MARIX, actuaire à la TA Retraite, Jiahui ZHU, data scientist à Wiz'You, Yassine LAGHZALI, actuaire à la TA IARD, pour leurs conseils au cours de la réalisation de mon mémoire.

Je souhaiterais adresser mes remerciements à Julius QUIQUET, responsable d'études actuarielles à Wiz'You, qui m'a encadrée lors de mon stage de Master 1 et qui a tout mis en œuvre pour que mon intégration se passe bien. Je souhaiterais enfin plus globalement remercier l'équipe Wiz'You pour leur accueil très chaleureux qui m'a permis de prendre rapidement mes marques.

Merci à Yoann GOUYEN, Etienne GUILLOU, Maxime MARIX, Vincent SADE, Thibault CAUSSE et Hakim M'MADI pour les relectures qu'ils ont pu apporter à ce mémoire.

Je remercie également ma famille qui m'a soutenue tout au long de mes études.

Table des matières

Introduction

La norme IFRS 17 est applicable depuis le 1er janvier 2023 et constitue une refonte des passifs du bilan comptable. Elle apporte également une nouvelle définition des contrats d'assurance. L'une des divergences notables entre la norme comptable IFRS 17 et la norme prudentielle Solvabilité II est la définition de la frontière des contrats. En effet, cette redéfinition implique désormais la prise en compte des versements libres dans les flux *best estimate*.

Les versements libres sont des versements qui peuvent être effectués à tout moment dans la phase de constitution des contrats d'épargne retraite. Deux types de versements libres sont principalement observés auprès des assurés :

- Les versements libres conjoncturels : sont des versements qui sont réalisés dans un objectif d'optimisation financière et fiscale. Les placements sont effectués en connaissance d'un contexte économique et financier particulier.
- Les versements libres structurels : sont des versements qui sont réalisés pour alimenter le contrat en question. Ces placements ne sont pas nécessairement optimisés mais sont pensés dans la continuité de la constitution de l'épargne retraite.

Ces versements sont par nature imprévisibles et aléatoires. Leur modélisation n'est donc pas simple.

Ce mémoire portera sur leur comptabilisation dans le périmètre de la retraite. Des études ont déjà été mises en oeuvre pour le périmètre de l'épargne. Ainsi, les travaux de ce mémoire vont se concentrer plus précisément sur la retraite individuelle. En effet, les volumes de versements libres sont essentiellement présents sur ce périmètre.

L'intégration de ces montants dans le *best estimate* représente donc un réel enjeu pour les compagnies d'assurance qui vont devoir modéliser ces flux. L'idée de ce mémoire sera ainsi de comprendre les dynamiques en analysant les facteurs favorisant les versements libres.

Ce mémoire s'est donc orienté autour de deux approches de modélisation pour répondre à la problématique.
La première est une approche simple visant à modéliser des versements libres selon une maille imposée par les outils internes à Generali. Ces taux seront construits par rapport

à l'historique disponible. Cette méthode permettra d'établir plusieurs interprétations au sujet des comportements des assurés par rapport aux versements libres.
Par la suite, l'idée est de challenger les hypothèses imposées par les outils internes. Ainsi, une autre approche sera envisagée grâce à des méthodes de *machine learning*. Dans un premier temps, une classification sera réalisée afin de prédire la survenance des versements libres. Dans un second temps, un modèle actuariel classique de fréquence sera développé afin de prédire le nombre de versements libres. Dans les deux cas, des analyses d'explicabilité des modèles à l'aide de SHAP seront mises en place afin de comprendre leur comportement.

Finalement, la dernière partie de ce mémoire permettra également d'établir des recommandations pour la modélisation des versements libres dans les outils internes de Generali.

Chapitre 1

Contexte réglementaire de l'étude

Cette partie vise à introduire le contexte réglementaire en vigueur susceptible d'impacter les investissements par versement libre. Dans un premier temps, les normes IFRS 4 puis 17 seront introduites car elles sont à l'origine de la problématique de modélisation des versements libres. La seconde partie traitera des enjeux et des changements apportés par l'intégration de ces montants.

1.1 Les normes IFRS

Depuis le 1er janvier 2005, les sociétés établies dans un Etat membre de l'Union Européenne dans un marché réglementé ont l'obligation de publier leurs comptes consolidés en norme IFRS (International Financial Reporting Standard) selon le règlement européen 1606/2002 [1]. Les IFRS sont des normes comptables visant à changer la communication financière pour plus de transparence. Elles cherchent à homogénéiser les enregistrements comptables des états financiers entre secteurs d'activité et entre pays. Des éléments comptables sont rendus publics pour permettre aux investisseurs de déterminer la situation économique d'une entreprise. Au-delà de l'Union Européenne, les normes IFRS sont appliquées dans 150 pays dans le monde dont le Canada, l'Afrique du Sud, le Japon etc.

Les normes IFRS sont la continuité des normes IAS (International Accounting Standard). Ces deux types de normes ont les mêmes objectifs à savoir de transformer les enregistrements comptables pour plus de transparence et d'homogénéité entre les entreprises. Les normes IAS étaient publiées par l'IASC (International Accounting Standards Committee) de 1973 à 2001 tandis que les normes IFRS le sont par l'IASB (International Accounting Standard Board) depuis 2001. Aujourd'hui, les normes IFRS prévalent sur les normes IAS car elles constituent une mise à jour de ces dernières.

1. Source : `https://acpr.banque-france.fr/europe-et-international/cadre-comptable/standards-internationaux/normes-comptables-internationales-ifrs`

Plus précisément, les IFRS sont des normes comptables internationales faisant intervenir différentes instances. Elles sont rédigées par l'IASB et appuyées par l'European Financial Reporting Advisory Group (EFRAG) qui évalue l'intérêt de l'application de la norme dans l'Union Européenne. L'IASB est l'organisme responsable de la rédaction des normes comptables IAS et IFRS. L'EFRAG intervient ensuite pour encourager la rédaction de normes internationales et s'assurer de leur conformité. De plus, le CFO Forum regroupe tous les directeurs financiers des principales assurances européennes qui interviennent dans le processus de validation de la norme. Finalement, la Commission Européenne vote la norme en tenant compte de l'avis de l'EFRAG. Il existe à ce jour dix-sept normes IFRS.

1.1.1 La norme IFRS 4

IFRS 4 est l'une de ces dix-sept normes. C'est une norme transitoire qui prévoit une comptabilisation des passifs **des contrats d'assurance** en normes locales. Elle vise à limiter les changements de pratiques comptables pour tout ce qui se réfère aux contrats d'assurance. Elle s'applique à tous les contrats d'assurance émis ainsi qu'à tous les contrats de réassurance détenus par la compagnie d'assurance. Cette norme a été publiée en deux phases : une première en 2007 puis la seconde en 2012.

Les principaux changements apportés par la norme IFRS 4 sont les suivants :

— Une définition des contrats d'assurance : les contrats sont rattachés à IFRS 4 ou à IAS 39 suivant la significativité du risque d'assurance couvert. Les contrats classés en IFRS 4 sont des contrats d'assurance et ceux attribués à IAS 39 sont des contrats d'investissement. Pour qu'un contrat soit qualifié de contrat d'assurance dans la norme IFRS 4, "une partie (l'assureur), en accord avec une autre partie (l'assuré) doit accepter un risque significatif en dédommageant l'assuré ou un tiers bénéficiaire en cas de survenance d'un risque futur qui l'affecterait défavorablement."

  Lorsque cette définition n'est pas satisfaite, les contrats sont classés en contrat d'investissement que ce soit avec ou sans participation discrétionnaire. Seuls les contrats d'investissement sans participation discrétionnaire sont comptabilisés en juste valeur. Les autres catégories, elles, sont enregistrées en coût historique.

— Le Shadow Accounting : selon les normes actuelles, les passifs sont calculés en coût historique (en norme locale) alors que les actifs le sont en juste valeur. Il réside donc un écart dans les engagements entre les actifs et les passifs du bilan comptable des compagnies d'assurance. Cet écart est visible pour les contrats avec participation aux bénéfices. La figure 1.1 permet d'illustrer ce mécanisme :

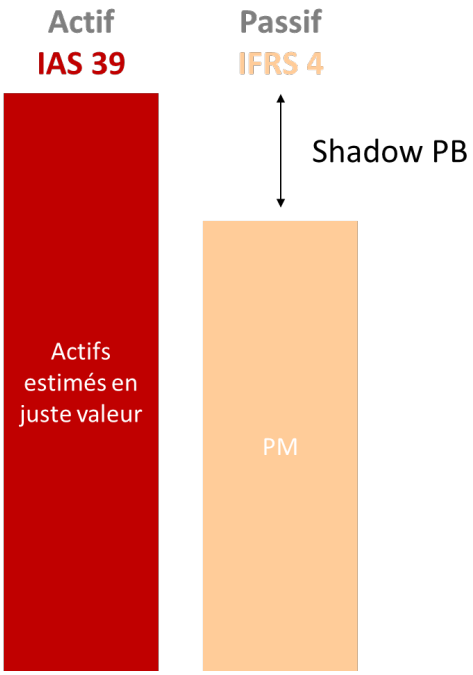


FIGURE 1.1 – Vision simplifiée du bilan en IAS 39 et IFRS 4

L'IASB a donc permis aux compagnies d'assurance de pratiquer le Shadow Accounting, autrement appelé "la comptabilité reflet". Cette méthode consiste à créer des sous-comptes au passif autorisant d'ajuster les provisions techniques, les coûts d'acquisition ainsi que les valeurs du portefeuille. Ainsi, des plus ou moins values latentes sont enregistrées en placements. Ces options ont un caractère temporaire et ne sont mises en place qu'en attente de la norme IFRS 17. Cette pratique reste néanmoins contrôlée par le mécanisme du Liability Adequacy Test.

— Le test LAT : les compagnies détenant des contrats d'assurance (classés en IFRS 4) sont tenues d'effectuer un test de suffisance des passifs (Liability Adequacy Test : LAT). Son objectif est de comparer les passifs à une estimation des flux futurs de trésorerie.

— Les provisions techniques sont majoritairement comptabilisées en normes locales.

Comme cela a été dit précédemment, la norme IFRS 4 a un rôle transitoire. De ce fait, plusieurs mesures sont mises en place dans un but temporaire en attente de la norme suivante. Ainsi, cette norme présente plusieurs limites qui ont conduit à la publication d'IFRS 17 :

— Les passifs ne sont pas évalués de façon économique. Le recours aux normes locales

rend complexe la comparaison entre pays, entre contrats d'assurance et entre secteurs d'activités.
— La norme ne prescrit pas de méthode de calcul pour les provisions. De ce fait, la vision rentabilité des produits n'est pas comparable.
— Un manque de transparence est également dénoncé notamment sur les informations relatives aux contrats d'assurance et sur leur rentabilité.

Finalement, il devient difficile de se fier aux résultats entre compagnies d'assurance car il existe trop de leviers de pilotage de ces derniers.

1.1.2 La norme IFRS 17

C'est dans ce contexte que la norme IFRS 17 [2] a été publiée le 18 mai 2017 et sera applicable le 1er janvier 2023 [12], [19].

Cette norme vise à apporter une information comparable entre secteurs et pays différents. Elle a pour objectif de donner une évaluation unifiée et **économique** pour ainsi fournir des indicateurs de rentabilité homogènes. IFRS 17 allie à la fois une approche basée sur la valeur actuelle, comme cela pouvait être le cas dans Solvabilité II, tout en reconnaissant les revenus sur la période de projection.

Pour cela, les normes locales ne sont plus admises pour les assureurs publiant leurs comptes en IFRS à savoir pour les entités cotées au sein de l'Union Européenne qui émettent des contrats d'assurance et qui ont pour activité principale l'assurance ou la réassurance. Chaque société doit appliquer cette norme aux contrats d'assurance qu'elle émet, aux contrats de réassurance qu'elle détient et aux contrats d'investissement comportant un élément de participation discrétionnaire qu'elle émet dans le cas où elle possède également des contrats d'assurance.

Le passage à IFRS 17 fait émerger trois notions essentielles :
— Le Best Estimate (BE) : est la valeur actuelle de l'ensemble des cash-flows futurs dans la limite de la frontière des contrats. Les provisions sont évaluées à leur valeur économique selon une méthode proche de Solvabilité II.

— L'ajustement pour risque (RA) : permet de couvrir l'incertitude autour de la provision établie. Elle correspond ainsi au montant que pourrait réclamer une entité tierce pour accepter l'incertitude des flux dans le cas où cette entité tierce souhaiterait reprendre le portefeuille d'assurance avec ses provisions.

2. Source : `https://www.ifrs.org/projects/completed-projects/2017/insurance-contracts/`

La marge de service contractuelle

La troisième notion apparue de la norme IFRS 17 est la marge de service contractuelle ou CSM [3] . Un point d'attention particulier est porté sur cette notion du fait de sa complexité.

La CSM représente les profits liés aux services futurs au-delà de la marge pour risque. Son calcul est réalisé au début de la vie du contrat afin d'éviter de comptabiliser du résultat au commencement de ce dernier. Ainsi, elle correspond au profit attendu par l'assureur. Dans le cas d'une perte, elle est immédiatement constatée. Son amortissement est basé sur la notion de "coverage units". Elle considère ainsi la durée attendue et la taille des contrats du groupe. Tout profit éventuel doit être étalé dans le temps tandis que les pertes sont enregistrées immédiatement.

A la transition, la CSM peut être valorisée de trois manières différentes :

— Full Retrospective Approach (FRA) : Comme son nom l'indique, cette méthode fonctionne intégralement de manière rétrospective. Elle impose ainsi que la norme s'applique à tous contrats dès leur commencement. Néanmoins, cela n'est pas toujours possible pour tous les assureurs. Une étude doit être réalisée pour estimer la capacité de la société à récupérer l'historique nécessaire. Cette méthode peut s'avérer coûteuse et n'est pas toujours applicable au sens du paragraphe C5 de la norme. Si tel est le cas, les deux prochaines méthodes sont alors privilégiées.

— Modified Retrospective Approach (MRA) : Cette seconde méthode vise à se rapprocher de la précédente mais étant donné que certaines informations sont manquantes, la rétrospective est seulement partielle. Dans cette méthode, l'idée est de tendre au maximum vers les dates d'origine des contrats. Néanmoins, les dates de début des contrats proches de la date de transition vers la nouvelle norme doivent être considérées. La tolérance s'applique surtout pour les contrats anciens afin de se rapprocher le plus possible de la première méthode.

— Fair Value Approach (FVA) : Cette dernière méthode ne se base pas sur l'historique comme les précédentes. Les contrats sont alors traités comme débutant à la date de transition.

Le fait de considérer l'une des méthodes peut avoir un impact non négligeable sur le montant de la CSM. Finalement, le dernier point de cet aparté va s'attarder davantage sur le fonctionnement de la CSM et la manière dont elle est calculée. Son mécanisme est synthétisé dans le schéma 1.2.

3. Source : `https://www.optimind.com/medias/documents/6511/tf-ifrs-17.pdf`

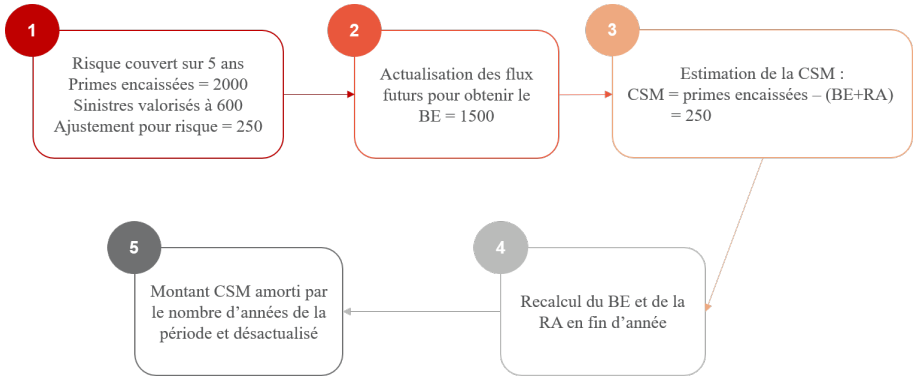


FIGURE 1.2 – Explication synthétique du calcul de la CSM

Plus globalement, le passage du bilan en norme IFRS 4 à IFRS 17 peut être représenté comme ci-dessous :

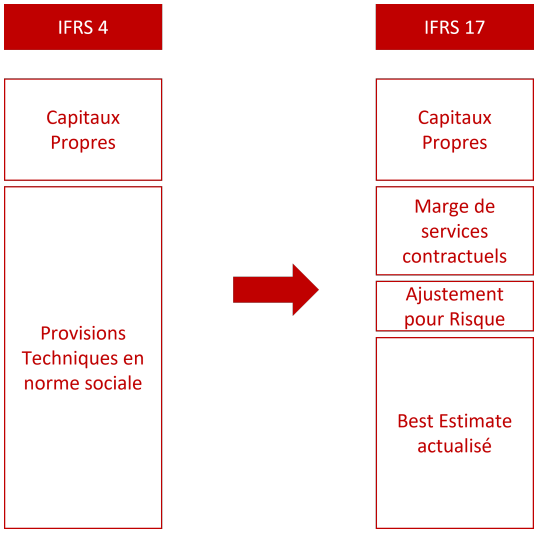


FIGURE 1.3 – Passage du bilan comptable IFRS 4 à IFRS 17

Ces trois notions offrent une vision plus économique. Toutes les compagnies devront communiquer sur leur BE. La *Risk Adjustment* permettra de communiquer sur l'appétence au risque en divulguant les quantiles utilisés pour quantifier l'incertitude. Finalement, la CSM informe sur la profitabilité de la compagnie.

Au-delà du bilan comptable, la vision de la performance dans le compte de résultat est également revue dans la norme IFRS 17. Sa présentation diffère de celle observée en norme IFRS 4. Le souhait d'harmoniser les résultats des différents secteurs d'activité

se traduit par un compte de résultat plus détaillé. Ce changement est expliqué dans le schéma 1.4 :

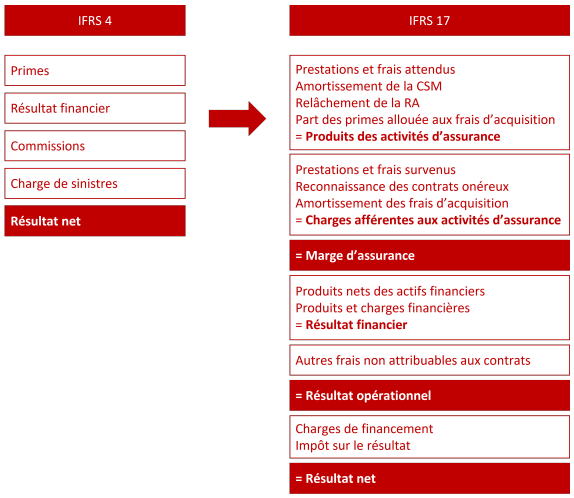


FIGURE 1.4 – Passage du compte de résultat IFRS 4 à IFRS 17 [19]

La comptabilisation et l'évaluation des sociétés doivent se faire au niveau du groupe de contrats. Des groupes de profitabilité sont ainsi définis par la norme. Le niveau d'agrégation choisi est déterminant en terme de résultat. Les contrats sont classés lors de la première comptabilisation entre contrats onéreux et contrats qui n'ont pas de possibilité significative de devenir onéreux. Un contrat est qualifié d'onéreux lorsqu'il est comptablement en perte. Cette nouvelle segmentation conduit les assureurs à montrer concrètement les générations de contrats en perte. Il n'est plus possible de masquer les pertes engendrées par une génération de contrat du fait de l'allocation immédiate de la CSM. La granularité imposée par IFRS 17 est représentée dans la figure suivante, les contrats sont séparés à la maille portefeuille, cohorte et contrat onéreux ou non.

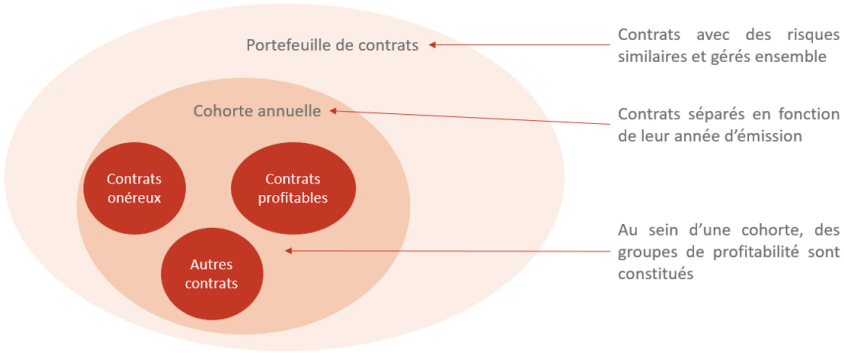


FIGURE 1.5 – Niveaux d'agrégation sous IFRS 17

1.1.3 Les enjeux liés à la modélisation des versements libres

La norme IFRS 17 fait émerger une nouvelle définition de la frontière des contrats. Elle est définie comme comprenant tous les contrats pour lesquels l'assureur est engagé et ne peut changer les conditions tarifaires.
Dans le cas des versements libres, l'assureur est engagé par les primes futures de l'assuré. De ce fait, les versements libres sont désormais intégrés dans la frontière des contrats. Ils vont impacter le Best Estimate.

Le paragraphe B65 de la norme comptable IFRS 17 liste les flux de trésorerie compris dans le périmètre du contrat, dont les primes et appuie ainsi les propos précédents :

> *"Les flux de trésorerie compris dans le périmètre du contrat d'assurance sont ceux qui sont directement liés à l'exécution du contrat, y compris ceux dont le montant ou l'échéancier sont à la discrétion de l'entité. Les flux de trésorerie compris dans ce périmètre comprennent a) Les primes (y compris les ajustements de primes et les primes à versements échelonnés) que verse le titulaire de contrat d'assurance et tout flux de trésorerie supplémentaire qui résulte de ces primes."*

Ce second extrait de la norme définit plus précisément la frontière des contrats [4] :
"IFRS 17§34 : Les flux de trésorerie sont compris dans le périmètre d'un contrat d'assurance s'ils découlent de droits et obligations substantiels qui existent au cours de la période de présentation de l'information financière dans laquelle l'entité peut contraindre le titulaire de contrat d'assurance à payer les primes ou dans laquelle elle a une obligation substantielle de lui fournir des services. Une obligation substantielle de fournir des services cesse dans l'une ou l'autre des situations suivantes :

a) L'entité a la capacité pratique de réévaluer les risques posés spécifiquement par le titulaire de contrat d'assurance et peut, en conséquence, fixer un prix ou un niveau de prestations qui reflète intégralement ce risque ;

b) Les deux critères ci-dessous sont remplis :
- ☐ *L'entité a la capacité pratique de réévaluer les risques posés par le portefeuille de contrats d'assurance dont fait partie le contrat en cause et peut, en conséquence, fixer un prix ou un niveau de prestation qui reflète intégralement le risque posé par le portefeuille*
- ☐ *L'établissement du prix de la couverture s'étendant jusqu'à la date de réévaluation des risques ne tient pas compte des risques liés aux périodes antérieures."*

Ce dernier paragraphe induit la prise en compte des versements libres, ils vont désormais être ajoutés aux primes périodiques.
En d'autres termes, la norme stipule que sont à prendre dans la frontière des contrats les flux générés suite à l'exécution d'options contractuelles qui impliquent la modification

4. Source : `https://www.institutdesactuaires.com/global/gene/link.php?doc_id=14379&fg=1`

de l'obligation de l'assureur. C'est dans ce cadre que s'inscrivent les versements libres qui sont des options de paiement de primes. Une fois cette option exercée, elle induit l'augmentation de l'engagement de l'assureur.

De ce fait, dans le cadre de contrats Retraite, la norme comptable rend désormais obligatoire la considération des versements libres dans la frontière des contrats. Jusqu'à présent, ces flux ne sont pas comptabilisés dans les résultats des assureurs. Cet aspect de la norme représente donc une réelle nouveauté pour les compagnies. Ce mémoire va donc viser à appréhender le comportement des assurés concernant leurs investissements par versements libres.

1.2 La loi PACTE

Les rendements générés par les placements en retraite ont été réduits du fait de taux d'intérêt bas et d'un contexte réglementaire en pleine transformation. La loi PACTE a donc eu pour vocation à redynamiser l'épargne retraite. Pour ce faire, plusieurs éléments majeurs ont été repensés comme la transférabilité entre contrats ou encore la possibilité pour les assurés d'investir sur des fonds en UC plus risqués pour augmenter leurs rendements. Cette partie va donc introduire la loi PACTE plus en détails.

La loi PACTE relative à la croissance et à la transformation des entreprises a été promulguée le 22 mai 2019. Elle souhaite populariser l'Epargne Retraite en regroupant les différents produits existants. En effet, la gestion de ces contrats était complexe et généralement peu ou pas optimisée. Les assurés cumulaient plusieurs produits sans pouvoir effectuer de transferts entre les contrats. Les fiscalités ainsi que les modalités de sortie différaient entre les produits.

La loi PACTE vise à permettre aux anciens contrats d'être transférés vers un autre produit dans une compagnie d'assurance tout en conservant la fiscalité du produit initial. Plus particulièrement, la loi PACTE a élargi provisoirement les possibilités de transferts des produits d'assurance vie vers les Plans d'Epargne Retraite (PER) [5]. Les PER servent à préparer la retraite tout en faisant des économies d'impôts. De ce fait, les versements volontaires sont déductibles des impôts dans la limite de plafonds établis. Ces produits sont perçus comme des compléments des produits d'assurance vie. Là où l'assurance vie permet une sortie en capital à tout moment, le nouveau PER offre un placement à plus long terme.
Jusqu'au 1er janvier 2023, des incitations fiscales sont mises en place de telle sorte que les assurés transfèrent leur assurance vie vers le nouveau PER. Dans ce cas, les rachats d'un contrat de plus de 8 ans offriront un abattement fiscal doublé.

Les épargnants possèdent ainsi un unique produit compartimenté en trois poches : les PER avec versements individuels, les PER avec versements collectifs et les PER avec

5. Source : `https://www.economie.gouv.fr/PER-epargne-retraite#`

versements obligatoires. Le nouveau PER est compartimenté entre l'épargne retraite au titre de l'entreprise et de l'individu :

— Les produits relevant du PERCO ou de l'Article 83 ont respectivement été revus par un PER collectif facultatif et un PER obligatoire.
— Les Madelin et les PERP sont également remplacés par un PER à titre individuel.

Les transferts sont facilement réalisables entre ces deux compartiments. Il est possible de sortir une partie des montants placés dans certains cas comme un accident par exemple. L'épargne investie est retirée au moment de la retraite en capital ou en rente.

Chapitre 2

Présentation du périmètre de l'étude

L'étude porte sur le périmètre de la retraite individuelle dont les produits sont commercialisés via des agents ou des courtiers. Ce sont les produits qui concentrent le plus de volumes de versements libres. En effet, les produits collectifs enregistrent moins de versements libres. Les assurés auront généralement tendance à privilégier des investissements sur leur propre contrat individuel retraite plutôt que sur un contrat d'entreprise. De plus, certaines fiscalités pour les contrats individuels encouragent les versements libres. Par exemple, pour les Madelin, on retrouve beaucoup de professions libérales. Ainsi, les assurés se constituent eux-mêmes leur retraite et vont donc verser bien plus dans leur intérêt. Cette dynamique n'est pas observée sur les produits collectifs.
Pour construire une loi de versements libres, il est important d'appréhender le périmètre dans lequel s'inscrit l'étude. Des hypothèses peuvent être fixées, elles doivent clairement être explicitées et justifiées par cette connaissance du portefeuille.
Les deux premières parties viseront donc à introduire les notions essentielles concernant l'assurance vie et le périmètre de la retraite. La dernière partie s'attachera à définir les produits composant le portefeuille étudié de Generali.

2.1 Généralités sur l'assurance vie

L'assurance vie [1] est aujourd'hui le placement préféré des français. Elle se décompose en deux types d'assurance : l'assurance en cas de vie ou en cas de décès.

Elle consiste à permettre à l'assuré de valoriser le capital qu'il investit. L'assuré peut réaliser des versements sur son contrat lorsqu'il le souhaite. En échange de ces versements, l'assureur s'engage à verser le capital constitué à l'assuré directement ou à un bénéficiaire (selon les clauses du contrat).

L'assurance peut s'adresser à des profils hétérogènes. En effet, il est possible de réaliser des placements sur des fonds en euro qui permettent une garantie en capital mais

1. Source : `https://www.economie.gouv.fr/cedef/assurance-vie`

aussi d'investir sur les marchés financiers plus risqués pour lesquels les espérances de rendement sont plus élevées. Il n'y a pas de limite de versements et le retrait de capital est faisable à tout moment sans pénalité. Aujourd'hui, les assureurs souhaitent diminuer leur stock de placements en euro qui sont de plus en plus difficiles à garantir dans un contexte de taux bas. Les placements sur des fonds euro, s'ils sont plébiscités par les assurés, restent néanmoins une spécificité française qui oblige les assureurs à garantir un taux d'intérêt quel que soit le contexte actuel.

Pour les raisons citées précédemment, les objectifs des assurés peuvent différer :

1. Epargner pour financer des projets personnels, préparer sa retraite ou une potentielle perte d'autonomie
2. Anticiper la succession de son patrimoine
3. Investir ses économies dans un but d'optimisation financière : ces contrats sont prisés des investisseurs car l'assurance vie offre un cadre fiscal souple et avantageux.

Il existe donc plusieurs types de contrats en assurance vie : les contrats d'épargne, de retraite et de prévoyance. Sur ces différents contrats, plusieurs types de versements sont réalisables :

— Les versements initiaux sont réalisés à la souscription du contrat. Le capital constitué est ensuite complété par les deux autres types de versements.
— Les versements périodiques sont cadencés selon un montant et une périodicité choisie ou fixée dans les conditions du produit.
— Les versements libres peuvent être effectués à tout moment. Ils sont par nature aléatoires et peuvent dépendre de nombreux facteurs : économiques, financiers, comportementaux. Cette spécificité explique la complexité pour les assureurs de les prédire.

Au sein des produits d'épargne, retraite et prévoyance, il existe deux types de contrats d'assurance qui sont en réalité complémentaires : les contrats individuels et les contrats collectifs. Les contrats individuels sont souscrits par un assuré et ne présentent généralement pas d'intermédiaires vis-à-vis de l'assureur. Dans le cas des contrats collectifs, c'est une personne morale qui est le souscripteur, c'est-à-dire une entreprise, une association etc. Cette entité représente donc un groupe d'assurés. Elle n'agit donc pas nécessairement dans l'intérêt de chaque assuré. Dans les deux cas, ces contrats visent principalement à couvrir la retraite complémentaire, l'épargne salariale ainsi que la potentielle perte d'autonomie ou de décès.

Ce mémoire portera plus particulièrement sur les produits de retraite. Un point d'attention sera donc apporté à ce sujet dans la prochaine partie.

2.2 Généralités sur les produits Retraite

Les contrats de retraite sont souscrits pour permettre de constituer un capital afin d'anticiper la retraite. Ce capital constitué sera ensuite reversé à l'assuré sous la forme d'une rente viagère ou d'un capital selon les conditions des contrats.

Comme cela a été développé dans la partie relative à la loi PACTE, les produits retraite ont été repensés pour plus de simplicité. Cette partie va permettre de détailler le regroupement des différents produits et leurs spécificités. Ainsi, les contrats d'épargne retraite sont compartimentés selon trois parties :

— Les PER à versements individuels : regroupent tous les régimes supplémentaires à adhésion individuelle. De ce fait, les versements sont faits par les assurés à titre individuel. Les assurés réalisent des versements volontaires ou des transferts venant d'autres contrats sans contrainte vis-à-vis de l'entreprise par exemple. Ces contrats ont pour objectif de capitaliser au cours de la vie active afin de percevoir une rente viagère ou sur option un capital.
Ce PER a succédé aux produits PERP et Madelin. Les PERP agissent sur le même principe que le PER individuel actuel. Ce produit fonctionne par capitalisation jusqu'au moment de la retraite. Les Madelin sont des produits plus spécifiques dédiés aux travailleurs non salariés. Ils leur permettent d'épargner pour leur retraite complémentaire. Les montants versés sur les contrats sont déductibles au niveau des impôts du résultat de l'entreprise. Ces produits sont encore accessibles aux assurés ayant souscrits avant la loi PACTE. Néanmoins, la loi PACTE permet désormais de réaliser des transferts tout en conservant les avantages fiscaux initiaux de ces produits.

— Les PER à versements collectifs : sont souscrits par des entreprises et sont à adhésion facultative ou obligatoire selon les conventions. Ce nouveau PER d'entreprise succède aux contrats dits Article 39, 82 et 83 ainsi qu'au Plan d'Epargne pour la Retraite Collectif (PERCO). Ils visent également à apporter un revenu complémentaire au moment du départ en retraite du salarié. Ce revenu complémentaire sera proportionnel au montant épargné par l'employeur tout au long de la période active du salarié.
Des subtilités existent entre ces types de contrats. Les potentiels transferts effectués entre ces produits garantissent le maintien du cadre fiscal de l'ancien produit. Ainsi, les contrats de type article 39 offrent la possibilité d'une sortie en rente viagère, elle sera néanmoins imposable. Les articles 83 fonctionnent sur ce même principe mais permettent une exonération partielle des impôts. Ils octroient une meilleure transférabilité du contrat d'une entreprise à une autre. Les articles 82 sont constitués par les employeurs en considérant un pourcentage du revenu du salarié. Deux sorties de l'épargne sont alors possibles que ce soit sous la forme d'un capital non imposable ou d'une rente viagère imposable comme c'est le cas pour les autres contrats. Actuellement, les contrats dits d'article 83 sont les plus

populaires. Selon les données de la DREES, le service statistique public du Ministère des solidarités et de la santé, le total des encours pour les produits de type article 83 s'élevait à 73,4 milliards d'euros en 2018, soit avant le passage de la loi PACTE. Les encours atteignaient 41,5 milliards d'euros pour les articles 39 et 4,2 milliards pour les articles 82.

— Les PER à versements obligatoires : sont les régimes légalement obligatoires. Les employeurs ainsi que les salariés sont obligés légalement de verser sur leur contrat le cas échéant. Ces montants alloués sont ensuite déductibles dans leur intégralité (versements des salariés et des employeurs) à hauteur de 8%. Au moment de la retraite, les fonds sont distribués sous la forme d'une rente viagère.

Les investissements sur ces contrats sont effectués :
— En euro : ce sont des placements plus sécurisés où un rendement minimal est garanti. Le contexte de taux bas induit des performances qui sont de moins en moins attractives.
— En UC : ici le capital assuré n'est pas garanti. Cela a pour conséquence de créer plus d'incertitudes sur la rentabilité du contrat. Ainsi, les gains et les pertes peuvent potentiellement être élevés.

Les contrats retraite sont des contrats dits de long terme. Ainsi, ils sont moins sensibles aux variations ponctuelles du contexte économique et financier. La gestion des contrats diffère donc davantage en fonction du nombre d'années restant avant le départ en retraite. Un assuré plus proche de l'âge de la retraite sera donc probablement plus prudent dans ses investissements pour anticiper la sortie de son capital.
De plus, en fonction des différentes fiscalités, les incitations sur les investissements ne sont pas les mêmes. Comme cela a été évoqué, certains placements offrent la possibilité de défiscaliser les versements réalisés. Certains assurés réalisent ainsi des versements en fin d'année afin d'optimiser leurs investissements. C'est notamment le cas sur les contrats Madelin. Cette saisonnalité des versements sera également évoquée dans la suite de ce mémoire.

2.3 Les produits Retraite au sein de Generali

C'est donc dans ce contexte qu'il convient d'introduire le périmètre de cette étude. Generali regroupe des produits individuels et collectifs. Les volumes présents sur la partie individuelle ont été jugés plus intéressants. La retraite individuelle est concentrée dans un système d'information appelé GB2000. Il correspond aux contrats individuels commercialisés via des agents ou des courtiers. Il est cependant possible de trouver encore quelques produits collectifs mais dans un moindre degré. Ils ne seront d'ailleurs pas développés car leur présence n'est pas suffisamment importante pour permettre une étude dédiée.

Cette étude portant sur l'analyse des versements libres, seuls les produits en comptabilisant ont été retenus. Ainsi, les travaux vont se concentrer sur les produits La Retraite et les PERP individuels.

Les produits La Retraite fonctionnent par génération de produits. Chaque nouvelle génération remplace la précédente, c'est-à-dire que le début des souscriptions d'un nouveau produit induit la fin de commercialisation du précédent. Ils sont divisés selon deux regroupements : La Retraite Monnaie pour les anciennes générations et La Néo Retraite pour les nouvelles générations de produits.

Plus précisément, un focus sur ces produits est proposé dans les paragraphes suivants :

— La Retraite Monnaie est constituée des produits la Retraite 80 (R80), la Retraite 86 (R86) et la Retraite 94 (R94). Chacun de ces produits correspond à une génération de produits (1980, 1986, 1994). Ils sont aujourd'hui en *run-off*, c'est-à-dire qu'ils ne sont plus commercialisés. Les contrats souscrits sur ces produits présentent aujourd'hui une ancienneté élevée et des capitaux constitués importants. Les phases de constitution des capitaux sont, pour la plupart des contrats, terminées. Les montants de versements libres seront donc de moins en moins présents sur ces produits. Il est déjà possible d'anticiper une décroissance des investissements. Pour autant, ces anciennes générations de produits ont garanti des taux d'intérêt technique très avantageux, ce qui n'est plus le cas aujourd'hui sur les nouvelles générations. Les tables de mortalité exploitées étaient également différentes. En considérant l'historique des versements, il faudra être vigilant quant au comportement des assurés. En effet, des taux d'intérêt technique intéressants peuvent encourager la décision d'effectuer un versement libre. Ces taux minimum garanti sont toujours présents aujourd'hui pour certains contrats de la R86. Il est donc possible d'anticiper un attrait particulier pour les investissements par versements libres dans ce cas.

— La Néo Retraite est composée de la Retraite 08 (R08), la Retraite 13 (R13), la Retraite 15 (R15) et la Retraite 20 (R20). Leur appellation a la même signification que pour la Retraite Monnaie. Ce nouveau regroupement a été mis en place à la suite de l'introduction des unités de compte. Les taux d'intérêt servi ainsi que les tables de mortalité ont été revus. Actuellement, il est seulement possible de souscrire des contrats sur la R20 car les autres produits sont en *run-off*. Les commentaires sont principalement les mêmes que pour les produits de la Retraite Monnaie. Les produits en *run-off* vont probablement connaître une décroissance des investissements dans le temps. A nouveau, cette information sera importante dans la compréhension de la problématique.
Néanmoins, il y a un réel sujet autour de la R20 car des incitations fiscales liées à la loi PACTE ont largement contribué à augmenter les montants de versements libres effectués sur ce produit ainsi que les transferts internes et externes. Les assurés ont bénéficié des mêmes avantages que sur d'anciens produits dans le cas des

transferts et non d'une affaire nouvelle. La loi PACTE a également encouragé les transferts externes dans le cas où un assuré possédait plusieurs contrats retraite chez des assureurs différents.
On s'attend donc à avoir des tendances de versements très différentes dans le temps par rapport aux anciennes générations de produits. C'est un produit sur lequel peu d'historique est disponible car sa commercialisation a débuté lors de la fin de l'année 2019. La dynamique actuelle de versement sera possiblement plus difficile à capter car elle s'éloigne des autres produits.

— Les Plan d'Epargne Retraite Populaire (PERP) [2] : les analyses se sont principalement orientées sur les deux produits suivants : Anthologie et Graphic. Ces produits ne sont plus commercialisés depuis respectivement fin 2020 et 2010.
Plusieurs particularités ont été décelées sur le produit Anthologie. Tout d'abord, l'annonce de la fin de commercialisation de ces produits en fin d'année 2019 et de la loi PACTE a généré beaucoup de transferts et de versements libres sur ces contrats. De nombreux assurés ont également souscrits la dernière année de commercialisation. Des comportements atypiques ont donc été observés sur 2020 et 2021 pour anticiper les transferts sur le nouveau PER individuel. Plus concrètement, pourquoi les assurés ont-ils massivement réalisé ces versements ? Cela s'explique par un changement de fiscalité. En effet, ces assurés relevaient initialement de l'article 83 avec une sortie de leur capitalisation sous forme de rente viagère. Ils étaient donc rattachés au compartiment 3 de la loi PACTE. Néanmoins, les transferts des PERP vers les nouveaux PER relèvent du compartiment 1 de la loi PACTE et induisent une sortie en capital. Cette opération leur a donc permis d'obtenir une sortie en capital et non en rente.
Désormais, les tendances de versements sont davantage à la baisse. Ce changement soudain de dynamique sera potentiellement difficile à anticiper par un modèle.
L'analyse du produit Graphic sera à nouveau sensiblement la même. Ce produit n'est plus commercialisé depuis douze années. De ce fait, les montants de versements libres sont voués à connaître une décroissance dans le temps. Cela entraînera également une baisse des montants de versements libres pour ce produit.

2. Source : `https://www.service-public.fr/particuliers/vosdroits/F10259`

Chapitre 3

Présentation des bases de données

Cette partie vise à introduire les bases de données ainsi que les variables utilisées dans ce mémoire. Plusieurs pistes de variables ont été imaginées afin de répondre à la problématique de prédiction des versements libres. Elles seront donc détaillées. Dans un second temps, la qualité des indicateurs sera discutée. Plusieurs tests seront effectués comme la présence de données manquantes ou de données aberrantes, les corrélations ainsi que le nombre de modalités pour les variables catégorielles. Finalement, les variables retenues seront décrites au travers de statistiques descriptives.

3.1 Spécificité des données

Les données utilisées sont importées depuis des bases SAS et Hadoop. Les données relatives aux contrats ont été extraites de SAS et les données des assurés des bases Hadoop. Des données externes macroéconomiques et financières ont également été récupérées pour étoffer les données de la partie modélisation. L'historique sélectionné s'étend de 2015 à 2021.
Pour rappel, le système GB2000 a été privilégié car il concentre le plus de volumes de versements libres. Les contrats conservés sont commercialisés par des agents ou des courtiers en assurance. GB2000 regroupe en majorité des produits individuels mais il est possible que certains produits collectifs soient toujours présents dans les données. La base ainsi constituée a été complétée par l'ajout des PERP qui comptabilisent également des volumes intéressants de versements libres.
GB2000 contient les différentes générations des produits La Retraite. Les PERP rassemblent les produits Anthologie et Graphic.

De plus, les spécificités des produits évoqués dans la partie précédente ont dû être prises en compte avant de débuter l'étude. Ainsi, le produit La Retraite 20 est le seul produit admettant toujours de nouvelles souscriptions. Ce produit comptabilise déjà davantage de versements libres que l'historique des anciennes générations des produits La Retraite. Tous les autres produits étudiés dans ce périmètre sont en *run off*. De ce fait,

aucun contrat supplémentaire ne pourra être souscrit. L'ancienneté des produits en *run-off* est donc vouée à augmenter dans le temps. Il est déjà possible de supposer que moins de versements libres seront observés dans le temps. Ces phénomènes ont déjà été analysés dans des études réalisées par Generali. Ainsi, les modèles devront capter la décroissance de la survenance de versements libres dans le temps.

Lorsque les spécificités des données auront été captées et les données récupérées, plusieurs bases ont été construites. En effet, deux bases de données ont été nécessaires en fonction de la partie traitée dans ce mémoire :

— L'étude préliminaire concernant la construction des taux de versements libres a nécessité une agrégation des données par produit (Anthologie, Graphic, R80 ...), support (euro et UC) et ancienneté. Seules les données contrats ont été exploitées.

— L'étude de modélisation des versements libres par des méthodes de *machine learning* a conduit à une agrégation des données par numéro de contrat, produit, année de versement et support. Cette approche vise à comprendre les versements libres et a donc intégré des données contrats, assurés ainsi que des données macroéconomiques et financières.

Un traitement spécifique est également réalisé dans les deux cas sur le calcul des montants de versements libres pour l'ancienneté 0. Ce point sera approfondi dans la section 4.3.

Cette étape de préparation des données est primordiale pour la suite des travaux. En effet, les données doivent être comprises et exploitables pour nos modèles et l'étude en question. Si ce n'est pas le cas, il sera difficile de donner un sens à nos résultats.

3.2 Description des bases de données

3.2.1 Données Mouvements et Fonds

Les données sont issues de bases internes à Generali recensant tous les mouvements liés aux contrats de la retraite. Parmi ces mouvements, on retrouve évidemment les versements libres mais également les transferts internes et externes, les participations aux bénéfices etc. De manière plus précise, les mouvements correspondent aux montants investis ou perçus sur les contrats quelle qu'en soit leur nature. La base est séparée en deux initialement entre la base contenant tous les mouvements de moins de 2 ans et ceux de plus de 2 ans.

Pour l'étude, seuls les versements libres, hors transferts, pour les contrats en cours ont été conservés. En effet, l'étude portant sur l'analyse des comportements des investissements par versements libres, il est nécessaire que les contrats soient en cours pour capter ceux qui sont sujet à faire ces versements.

Pour autant, certaines anomalies ont été constatées pour quelques contrats. Des contrats en cours présentaient des montants de primes périodiques nuls. Au total, 2049 contrats n'avaient comptabilisé aucun versement sur l'année observée. Une étude plus approfondie a permis de constater que certaines polices sont considérées comme étant "Non En Cours" dans le cas d'arrêt de paiement. La phase de capitalisation serait alors terminée dans un contexte normal. Néanmoins, la pandémie de la Covid 19 a induit plus de souplesse sur le passage de ces contrats en "Non En Cours".

Parmi les différents mouvements, les participations aux bénéfices, les capitalisations des plus-values, la division des valeurs des parts, les versements périodiques, les arbitrages des compagnies, les rachats, les versements exonérés et toutes les conversions n'ont pas été retenus car ce mémoire a uniquement porté sur les versements libres.
Plusieurs variables ont été exploitées dans cette base telles que :

Nom de la variable	Description	Exemples de valeurs observées
Code_produit	Code permettant d'identifier le produit, son support ainsi que sa fiscalité	199, 200, 201 ...
produit_commercial	Libellé du code produit	Retraite 80 ASSURANCE VIE 100% PB
Fiscalité	Fiscalité du produit	Madelin, Assurance vie ...
libelle_origine_acte	Nom des mouvements enregistrés. Cette variable permet d'identifier les versements libres	Participation aux Bénéfices, versements complémentaires ...
code_apporteur	Identifiant de l'apporteur du contrat	0291, 999AA ...
TIT	Taux d'intérêt technique net de frais du produit	0,7% ...
Date_situation_acte	Date du mouvement	01/01/2020
Mnt_net_frais_e	Montant du mouvement	940,3
N_du_fonds	Code permettant la jointure avec la table Fonds pour récupérer la ventilation euro et UC des mouvements	15675, 125905 ...

Une variable calculée a également été ajoutée à partir de cette base concernant le montant du versement libre de l'année précédente dans le cas où un versement a été effectué. Cette variable est donc une variable numérique. L'objectif est d'analyser si les assurés faisant des versements libres ont tendance à le faire tous les ans ou non.

L'importation des données dans cette section a été explicitée sur les bases de la retraite. Les traitements ont été dupliqués sur les bases de l'épargne pour récupérer les PERP. La même démarche a ainsi été suivie : les mêmes variables ont été exploitées et la jointure a été répliquée.

3.2.2 Données provisions

A nouveau, les données sont issues de bases internes à Generali. Les provisions mathématiques sont réparties sur des bases différentes en fonction des produits et des supports. Les bases étant annuelles, il a été nécessaire de répliquer l'importation des données pour chaque année de l'historique. Chaque table a été extraite à la clôture c'est-à-dire au 31 décembre de l'année.

Une fois les données récupérées, des doublons ont été constatés. Des analyses ont permis d'expliquer ces doublons par le fait que les provisions soient calculées pour tout l'historique des taux minimum garantis (TMG) de chaque contrat. Or en 2017, Generali a mis en place un versionning sur les contrats réduisant ainsi leur TMG. Pour résoudre ce point, un filtre sur les TMG négatifs ou nuls a été retenu. En effet, très peu de produits garantissent un TMG positif depuis ce versionning. Seul un produit de la R86 en garantit toujours un. De ce fait, un point d'attention particulier a été porté sur la sélection des montants de provisions. Il était important de considérer les bons montants afin de ne pas générer de façon matérielle et non conforme au réel, au moment de la construction des taux de versements libres, des montants de versements avec des taux positifs. En effet, aujourd'hui, très peu de nouvelles primes sont rentrées sur le marché avec des taux garantis positifs.

L'historique de cette étude débutant en 2015, seules les provisions mathématiques avec des taux négatifs ou nuls ont été retenus. Les taux positifs encourageaient davantage les assurés à réaliser des versements. Cette tendance était seulement observée dans le passé avant les campagnes d'avenants réalisées. L'idée était donc de conserver les montants en cohérence avec le contexte actuel.
Ces bases ont fourni les éléments suivants :

Nom de la variable	Description	Exemples de valeurs observées
TIT	Taux d'intérêt technique	0% ...
tx_frais_gestion_eur	Taux de frais de gestion pour les supports euro. Cette variable existe également pour les supports UC.	0,7% ...
support	Identifiant indiquant la ventilation euro et UC	EUR ou UC
PM	Montant de provisions mathématiques	50000€

Les deux prochaines parties vont s'attacher à récupérer le plus d'informations possible au sujet des assurés. L'objectif étant de réussir à comprendre le comportement de ces derniers, toutes les variables potentiellement liées ont été exploitées. La qualité des différents indicateurs sera analysée par la suite.

3.2.3 Données contrats Retraite et Epargne GB2000

Ces bases sont également issues des systèmes Hadoop de bases de données internes à Generali. Deux bases ont été récupérées : une base avec les contrats retraite pour les produits Retraite Monnaie et Néo Retraite ainsi qu'une base avec les contrats épargne pour les PERP : Anthologie et Graphic. L'objectif était de récupérer des informations complémentaires relatives aux assurés pour étayer la compréhension des versements libres.

Ces bases ont été utilisées pour récupérer les données suivantes :

Nom de la variable	Description	Exemples de valeurs observées
categorie_socio_prof_a1	Libellé de la catégorie socio-professionnelle de l'assuré	Profession médicale, Ouvrier ...
date_naissance_a1	Date de naissance de l'assuré	1970-01-01
sexe_a1	Sexe de l'assuré	F ou H
prime_annuelle_pp	Montant annuel de primes périodiques de l'assuré	5000€
salaire_annuel	Salaire annuel de l'assuré	0 ou vide
fract	Fréquence de paiement des versements périodiques	M, T, S ou A
motif_entree	Motif d'arrivée du contrat	Affaire nouvelle, remise en vigueur, transfert depuis les PERP ...
date_sortie	Date de fin de la phase capitalisation du contrat	2021-10-01

Un point d'attention est porté sur le choix de la variable de date de sortie du contrat. Cette étude porte sur l'analyse des comportements des assurés lors de la phase de constitution de leur capital. Or selon les types de contrats retraite observés, ils peuvent être gérés en constitution ou en restitution. En effet, les capitaux sont constitués par les assurés puis restitués par les assureurs sous la forme d'une rente ou d'un capital selon les clauses du contrat. La donnée date de sortie est effectivement gérée en constitution dans le système GB2000. De ce fait, tant que la date de sortie n'est pas dépassée, un assuré peut effectuer des versements. Cette variable a donc permis de calculer une exposition à la potentielle survenance d'un versement libre. Elle a été utilisée dans la suite de l'étude pour la modélisation de ce phénomène.

3.2.4 Données RCE : Référentiel Client Entreprise

A nouveau, les bases ont été extraites de données internes en épargne et en retraite. Précédemment, les données utilisées concernaient les contrats et les provisions mathématiques de la retraite. Les bases RCE ont été utilisées pour compléter les données concernant les assurés. Les éléments suivants ont ainsi été retenus :

Nom de la variable	Description	Exemples de valeurs observées
codedepartement et codepostal	Numéro du département ainsi que le code postal	75 et 75020
nbrctract_vie et nbrctract_iard	Nombre de contrats vie ou iard en cours à la date observée souscrits par l'assuré	0, 1, 2 ...
statutmatrimonial	Type juridique du régime matrimonial ou du contrat de Pacs entre deux personnes physiques	La communauté réduite aux acquêts, la communauté de meubles et acquêts, la séparation de biens ...
situationfamiliale	Position qui définit la situation familiale	Célibataire, marié, veuf, divorcé ...
nombreenfants	Nombre d'enfants fiscalement à charge	0, 1, 2 ...

3.2.5 Données externes

Afin d'étoffer la base de données, des informations externes ont été ajoutées. L'objectif est de comprendre si le contexte économique et financier peut également impacter les comportements de versements. Comme évoqué précédemment, les versements libres peuvent être structurels ou conjoncturels. L'ajout de ces données pourrait donc permettre de capter les comportements des assurés en cohérence avec un contexte économique et financier connu.

Des indicateurs de marché financier ainsi que des indicateurs relevant de l'épargne réglementée de type livrets ont été récupérés. En effet, il est naturel de se demander dans quelle mesure ces deux sources d'indicateurs peuvent influencer les assurés dans leurs versements libres. Ainsi, certains assurés pourraient privilégier les investissements sur certains produits ou placements en fonction des taux proposés à des fins d'optimisation financière.

Pour cela, des données INSEE, Banque de France, data gouv et finance yahoo ont été intégrées comme en témoigne le tableau ci-après.

Type de données	Nom de la variable	Source
Données macroéconomiques du mois de décembre	Taux du Livret A	Banque de France
	Taux d'inflation	Banque de France
	Evolution annuelle en pourcentage du PIB	INSEE
Données macroéconomiques annuelles	Taux de rendement moyen annuel des livrets défiscalisés	INSEE
	Taux de rendement moyen annuel du Plan Epargne Logement	INSEE
	Taux de rendement moyen annuel du Compte Epargne Temps	INSEE
	Taux de rendement moyen annuel du Livret d'épargne populaire	INSEE
	Inflation annuelle	INSEE
Données financières	Cours du CAC 40 en décembre	Yahoo Finance
	Cours EURO STOXX 50 en décembre	Yahoo Finance
Données géographiques	Salaires moyens des actifs par commune	Data Gouv
	Tranche d'unité urbaine en 2020	INSEE

Des hypothèses ont été fixées concernant la récupération des données externes. Les informations jugées pertinentes dans le cadre de notre étude sont généralement journalières ou mensuelles. De ce fait, il a fallu statuer sur la méthode de calcul de ces indicateurs. Une première approche a constitué à observer la saisonnalité des versements libres pour ainsi choisir une période de l'année adéquate. Cette analyse a démontré une saisonnalité marquée sur les mois de décembre et de janvier. 60% des versements libres tous produits confondus sont concentrés sur les mois de décembre et de janvier comme le montre le graphique 3.1.

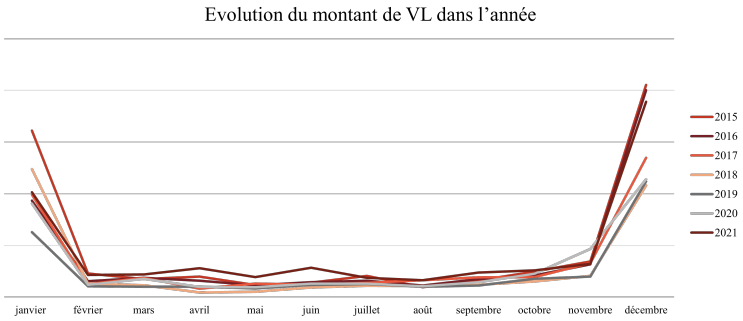


FIGURE 3.1 – Evolution des montants de versements libres dans l'année

Cette dynamique est principalement liée aux produits Madelin. En effet, la forte re-

présentativité de ces deux mois témoigne de la fiscalité avantageuse de ces produits. Les assurés ont toute l'année pour réaliser des versements sur leur contrat ; néanmoins ils réalisent très fréquemment en fin d'année que la somme de versements faite peut encore être augmentée avant de dépasser le plafond de déductibilité fiscale. De plus, les produits Madelin sont destinés aux travailleurs indépendants. De ce fait, ces investissements plus importants sont également portés par la volonté des assurés de constituer leur retraite.
La présence de montants de versements libres sur le mois de janvier peut s'expliquer par le décalage entre la date de versement et la date d'enregistrement par les gestionnaires. C'est pourquoi la récupération des indicateurs de décembre a été privilégiée. L'objectif étant d'essayer de capter les comportements des assurés cherchant à optimiser leurs placements, l'idée a été de comparer les différents niveaux des variables externes pour regarder dans quels cas la décision a été portée sur les versements libres.

La seconde approche a constitué à récupérer ces mêmes indicateurs en fonctionnant par moyenne. L'INSEE propose des taux annuels. Ils ont ainsi été exploités à titre de comparaison. Une étude des corrélations avec la variable cible n'a pas permis de privilégier l'une ou l'autre des méthodes. En fonction des variables analysées, les corrélations étaient parfois très proches. La première réflexion a finalement été retenue.

De plus, des données relatives à la situation financière et au lieu de résidence de l'assuré ont été intégrées à l'étude. Des premières analyses sur la qualité des données ont démontré que les données internes à ce sujet étaient mal renseignées, ce point sera détaillé dans la prochaine partie. Ainsi, des données externes issues de *data.gouv* ont été trouvées au sujet des revenus moyens par commune. Il est possible de supposer que des communes avec des revenus moyens plus élevés contiennent des assurés plus à même de réaliser des versements libres.
Dans la continuité de cette idée, des informations issues de l'INSEE concernant les tranches d'unité urbaine ont également été ajoutées. Cette variable segmente les communes comme étant une continuité de zone bâtie, c'est-à-dire qu'il n'y a pas de coupure de plus de 200 mètres entre deux constructions. Cela permet ainsi d'identifier les communes isolées ou denses. La densité urbaine peut amener à une sollicitation commerciale plus forte dans tous les domaines, y compris financier, ce qui pourrait amener davantage d'attrait au pilotage de son épargne. Cette variable comporte les neuf modalités suivantes :

- 0 : Commune hors unité urbaine
- 1 : Commune appartenant à une unité urbaine de 2 000 à 4 999 habitants
- 2 : Commune appartenant à une unité urbaine de 5 000 à 9 999 habitants
- 3 : Commune appartenant à une unité urbaine de 10 000 à 19 999 habitants
- 4 : Commune appartenant à une unité urbaine de 20 000 à 49 999 habitants
- 5 : Commune appartenant à une unité urbaine de 50 000 à 99 999 habitants
- 6 : Commune appartenant à une unité urbaine de 100 000 à 199 999 habitants
- 7 : Commune appartenant à une unité urbaine de 200 000 à 1 999 999 habitants
- 8 : Commune appartenant à l'unité urbaine de Paris

3.3 Analyse descriptive de la base de données

3.3.1 Analyse de la qualité des données

La première étape de préparation des données visait à réunir un maximum de variables afin d'imaginer une base idéale pour l'étude. Cette seconde étape consiste à vérifier la qualité de ces variables. Pour cela, différents indicateurs ont été analysés pour chaque variable :

— La présence de données manquantes
— La présence de données aberrantes
— Le nombre de modalités dans le cas de variables catégorielles
— La corrélation entre les variables

Ces différents indicateurs ont donc amené soit à retraiter les variables soit à les supprimer de la base de données. La suite de cette partie va être décomposée selon les quatre vérifications énoncées ci-dessus. Les variables ne répondant pas aux critères établis seront davantage détaillées.

✓ **Les données manquantes**

40% des variables de la base de données comptent au moins une valeur manquante. Leur présence est plus ou moins problématique selon la variable analysée et la quantité de données manquantes.
Des données non renseignées ont été identifiées sur les variables suivantes :

Nom de la variable	% de données manquantes
Montants de versements libres	96.3%
Montants de versements libres l'année précédente	97.8%
Taux d'intérêt technique	36%
Taux de frais de gestion euro	36%
Taux de frais de gestion UC	64%
Date de sortie	79.9%

Pour autant, cela ne vient pas de la qualité des données. En effet, dans les cas des montants de versements libres, cela signifie simplement qu'aucun versement libre n'a été réalisé sur l'année observée. Il en est de même pour la date de sortie, cela signifie qu'elle n'est pas encore connue. Pour les taux d'intérêt technique, les données manquantes correspondent aux supports UC qui ne garantissent aucun taux d'intérêt. Cette explication est également valable pour les deux variables de taux de frais de gestion. Les données non renseignées de la variable euro s'expliquent par la présence de support UC et inversement. Seuls les taux de frais de gestion des placements en UC du produit Anthologie n'étaient pas recensés. Une étude des conditions du produit a permis d'identifier le taux de frais de gestion alloué à ce produit. Les données manquantes ont ainsi été remplacées. Finalement, pour ces premières variables, seul un regroupement des variables de taux de

frais de gestion a été nécessaire. Elles ont donc toutes été conservées.

De plus, d'autres variables contenaient des données manquantes mais dans un ordre de grandeur négligeable.

Nom de la variable	% de données manquantes
Nombre de contrats vie	0.4%
Nombre de contrats iard	0.4%
Salaire moyen	3.5%
Retraite moyenne	4.8%
Revenus globaux moyens	5.2%
Numéro de département	0.4%
Code postal	0.4%

Le pourcentage de données manquantes est proche ou inférieur à 5% de la base. Dans ces conditions, les données non renseignées ont été imputées de la modalité la plus fréquente dans le cas de variables catégorielles ou d'une moyenne dans le cas des variables numériques.

Enfin, certaines variables contenaient trop d'informations manquantes et n'ont pas été conservées dans la suite de l'étude. Parmi ces variables sont listées :

Nom de la variable	% de données manquantes
Nombre d'enfants	77.5%
Salaire annuel	99.0%
Statut Matrimonial	42.3%
Situation Familiale	42.3%
Catégorie Socioprofessionnelle	58.8%

Comme évoqué précédemment, les variables relatives aux revenus annuels et à la catégorie socioprofessionnelle ont été remplacées par des données externes récoltées via l'INSEE.

✓ **Les données aberrantes**

L'objectif de cette analyse est de s'assurer de la fiabilité et de la cohérence des variables étudiées. La présence de données atypiques n'est pour autant pas toujours le signe d'une mauvaise qualité de données. En effet, il est également possible de constater la présence de profils plus particuliers. Ce paragraphe va donc détailler les variables possédant des données atypiques.

Le nombre de contrats vie : Cette variable prend ses valeurs entre 0 et 19. 99% des informations étaient concentrées sur les six premières modalités. Au-delà de cinq contrats vie, les fréquences des modalités étaient très faibles. Cette variable numérique a donc été retraitée. Une modalité "5+" a ainsi été créée.

Le nombre de contrats iard : 90% des informations correspondaient à la modalité 0. Certains assurés avaient parfois entre 20 et 61 contrats iard. Ces assurés sont néanmoins très rares. C'est donc en ce sens qu'une variable top a été créée. L'idée sera alors de tester si la présence d'un ou plusieurs contrats iard peut influencer les investissements par versements libres.

Le nombre d'enfants : Cette variable a été supprimée dans le paragraphe précédent en raison des données manquantes. Elle présente également des données atypiques notamment pour deux contrats qui recensent 17 et 35 enfants. Pour les raisons citées précédemment, cette variable n'a pas été conservée.

Les montants de provisions mathématiques : 332 contrats, soit 0.2 % des contrats, présentent des montants de provisions mathématiques inférieurs à 1€. A l'inverse, 78 contrats ont un montant de provision mathématique qui dépasse 500 000€. Ces cas sont très rares et envisageables pour plusieurs raisons. Les montants importants peuvent être constatés dans le cas d'assurés fortunés. Dans le cas des faibles montants de provisions mathématiques, des recherches complémentaires ont été réalisées. Pour 89% d'entre-eux, les contrats sont des affaires nouvelles, peu de provisions sont donc constituées. Aucun retraitement particulier n'a été réalisé car ces constats sont négligeables.

Les montants annuels de primes périodiques : 5 388 contrats enregistrent un montant annuel de prime périodique nul. 5 380 d'entre-eux sont observés sur le produit Anthologie. Ce cas avait déjà été évoqué précédemment. Si les assurés ne versent pas de primes, leur contrat passera au statut de non en cours. Ils ne pourront donc plus réaliser de versements jusqu'à la perception de la rente ou du capital lors du départ à la retraite. Moins de contrats sont concernés sur l'année 2020 du fait de la pandémie de la Covid 19. Il y avait en effet eu plus de souplesse. A l'inverse, un contrat présente un montant de primes périodiques de plus de 100 000€ sur quatre années consécutives sur le produit La Retraite 13 sur les supports euro et UC. Ces différents cas n'ont pas été retraités car ils peuvent se justifier comme pour les montants de provisions mathématiques précédemment.

Les montants de versements libres : 9 contrats ont enregistré des montants de versements libres dépassant les 100 000€. A nouveau, les assurés sont probablement fortunés car ils réalisent chaque année des versements très élevés. A l'inverse, 43 contrats ont observé des versements libres de moins de 10€. Dans ce cas, ces montants très faibles de versements libres peuvent être un choix de l'assuré ou bien la volonté de rester dans la phase de capitalisation.

L'âge actuel : 5,8% des contrats sont encore présents après l'âge légal de départ à la retraite, soit 62 ans. Les contrats passent en prorogation c'est-à-dire que le délai légal est dépassé mais pour autant la situation actuelle est conservée. Ce n'est pas anormal car certaines personnes partent en retraite après l'âge légal. Parmi ces contrats, seulement

2,7% dépassent l'âge de 70 ans. Les données n'ont pas été retraitées car les volumes concernés sont à nouveau très faibles.

L'année : Les années 2015 à 2017 représentent seulement 12% des données. Ce constat était prévu car pour ces années des taux positifs étaient encore garantis. Ces contrats n'ont pas été conservés pour ne pas fausser l'historique. Néanmoins, certains contrats proposaient déjà des taux négatifs ou nuls. C'est pour cette raison que ces années ont été conservées.

Le code du département : Dix valeurs aberrantes ont été détectées à savoir 0, 00, NW, J8, L-, M4, J9, CH et LO. Ces modalités ont été remplacées par des données manquantes. Il en est de même pour le code postal. Douze modalités aberrantes, soit dans les deux cas 0,02%, ont été remplacées par des vides.

✓ **Le nombre de modalités dans le cas des variables catégorielles**

L'objectif est d'éviter d'avoir un nombre trop important de modalités sur les variables catégorielles. Lors de la partie modélisation, cela peut s'avérer problématique dans l'encodage des variables. La démultiplication des variables catégorielles en variables binaires peut alourdir les temps de calcul. De plus, un nombre trop important de modalités peut complexifier la compréhension de la variable.

Le code apporteur : L'idée était de regarder si certains apporteurs d'affaires amenaient des assurés réalisant plus de versements libres. Pour autant, cette variable comptabilise 8395 modalités. 40% des apporteurs ont moins de dix contrats. L'apporteur ayant amené le plus de contrats en compte 8806 soit 1,6% de la base totale de données. De plus, les codes apporteur étant très peu lisibles et documentés, cette variable n'a pas été conservée.

Le code postal : La base compte 6647 codes postaux différents. Comme évoqué, certaines données aberrantes étaient également présentes. Pour la suite de cette étude, le numéro de département a été privilégié. Cette variable a essentiellement été utilisée dans le but de récupérer des informations au sujet du niveau de vie et du type de zone de résidence (rurale ou urbaine).

✓ **La corrélation entre les variables**

Cette partie vise à identifier des potentielles corrélations entre les variables afin d'éviter des répétitions dans les informations. En effet, cela pourrait nuire à la significativité de certains indicateurs. Des traitements adaptés ont donc été utilisés pour les variables numériques et catégorielles.
Pour les variables numériques, la corrélation de Pearson a été utilisée. Ce coefficient permet de quantifier si deux variables ont une relation linéaire. Ce coefficient peut prendre ses valeurs entre -1 et 1. Il est ainsi possible de regarder si les variables évoluent ou non

dans le même sens.
Ce coefficient est calculé comme suit :

$$\text{Coefficient de Pearson} = \frac{\sum_{k=1}^{n}(V_{1k}-\overline{V_1})(V_{2k}-\overline{V_2})}{\sqrt{\sum_{k=1}^{n}(V_{1k}-\overline{V_1})^2}\sqrt{\sum_{k=1}^{n}(V_{2k}-\overline{V_2})^2}}$$

où $\overline{V_1}$ et $\overline{V_2}$ représentent respectivement les moyennes des variables 1 et 2. Cette méthode a permis d'obtenir la matrice de corrélation suivante :

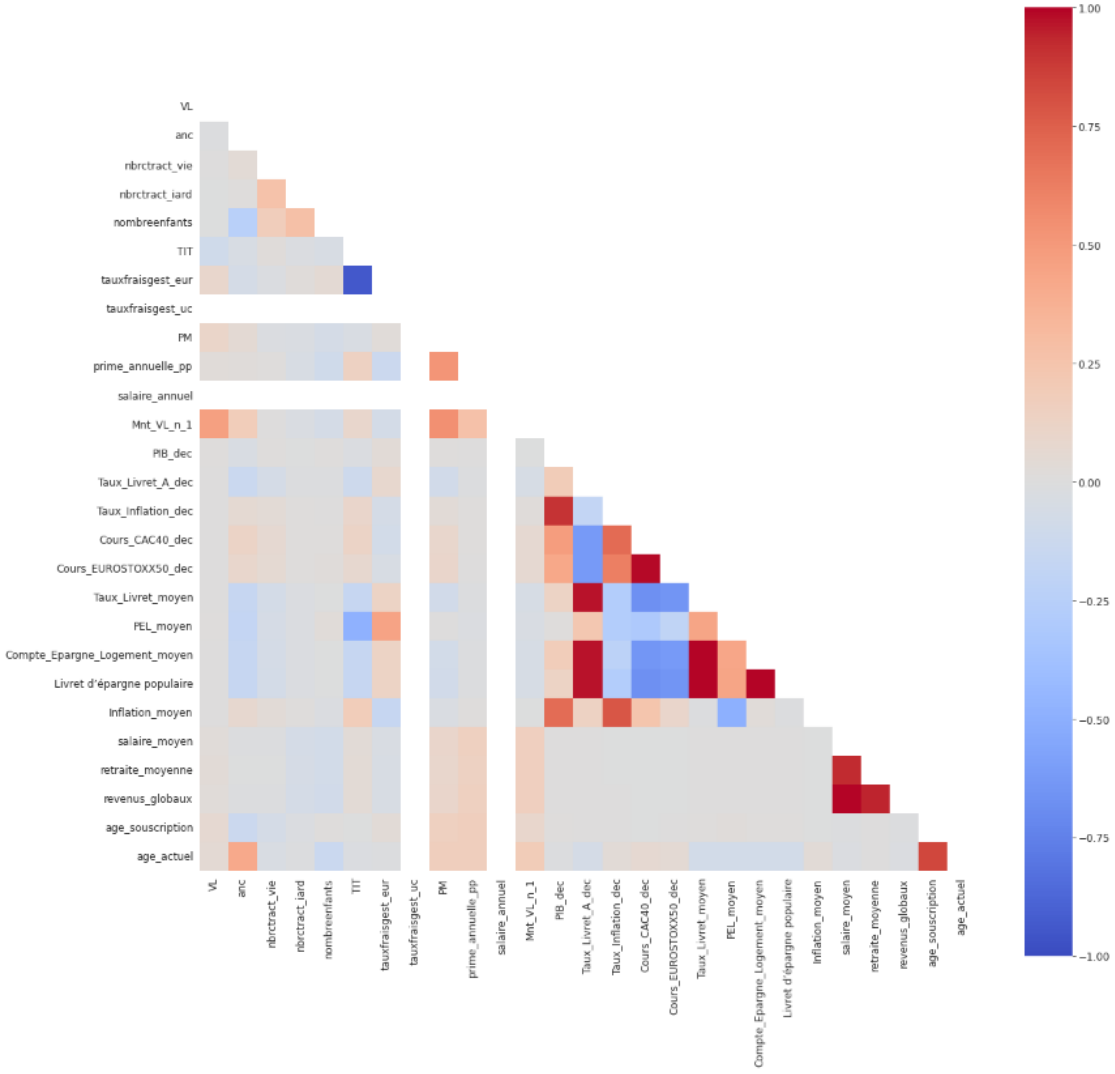


FIGURE 3.2 – Matrice de corrélation des variables numériques

Cette étude des corrélations a permis de constater des dépendances très fortes entre les variables financières. Le niveau du PIB, le cours du CAC 40 et de l'Eurostoxx 50 ainsi que les taux des différents livrets ont des coefficients de corrélation qui dépassent 50%

pour la plupart. Une sélection des variables a donc été effectuée.

Les cours Eurostoxx 50 et CAC 40 donnent une information similaire dans les modèles. De ce fait, seule la variable concernant le cours du CAC 40 a été conservée.

Dès lors qu'un choix était réalisable entre des indicateurs récupérés en décembre et une moyenne, les indicateurs de décembre ont été privilégiés. Ainsi, les variables relatives au niveau du PIB, au taux du livret A et au taux d'inflation du mois de décembre de chaque année ont été conservés dans l'étude.

Les variables relatives à l'âge actuel et à l'âge à la souscription sont également logiquement corrélées à 100%. L'âge actuel a été conservé pour les modèles. En effet, on peut supposer que les assurés se rapprochant de la retraite feront plus de versements libres.

De même, les variables relatives aux revenus sont très fortement corrélées. Les niveaux de salaires et de pensions sont corrélés à 92.2%. La corrélation atteint 99.7% entre les variables salaires moyens et revenus globaux. Cela paraît cohérent car les revenus globaux sont simplement une moyenne pondérée des deux sources de revenus précédentes. Comme l'objectif de ce mémoire est de comprendre le phénomène des versements libres, il paraît plus judicieux de conserver la variable salaire moyen par commune. En effet, cette variable renseigne sur le pouvoir d'achat des actifs qui peuvent potentiellement réaliser des versements libres s'ils possèdent un contrat retraite. A l'inverse, les retraités qui possédaient un contrat retraite sont en phase de restitution désormais, c'est-à-dire de perception de la rente. Ils ne versent donc plus sur leur contrat retraite. Ils ne rentrent donc pas dans le cadre de ce mémoire.

Pour les variables catégorielles, le V de Cramer a été retenu pour étudier les corrélations. Ce coefficient permet de quantifier l'intensité du lien entre deux variables qualitatives à l'aide du χ^2. Il prend ses valeurs entre 0 et 1. Une valeur de 0 indique que le lien entre deux variables catégorielles est inexistant. A l'inverse, une valeur de 1 signifierait une dépendance totale entre les deux variables. Il est calculé comme suit :

$$\text{V de Cramer} = \frac{\chi^2}{n \times min(l,p)}$$

où χ^2 correspond au résultat du test de χ^2, n au nombre d'observations et l et p au nombre de lignes et de colonnes de la table croisée.

Cette méthode a permis d'obtenir la matrice de corrélation suivante :

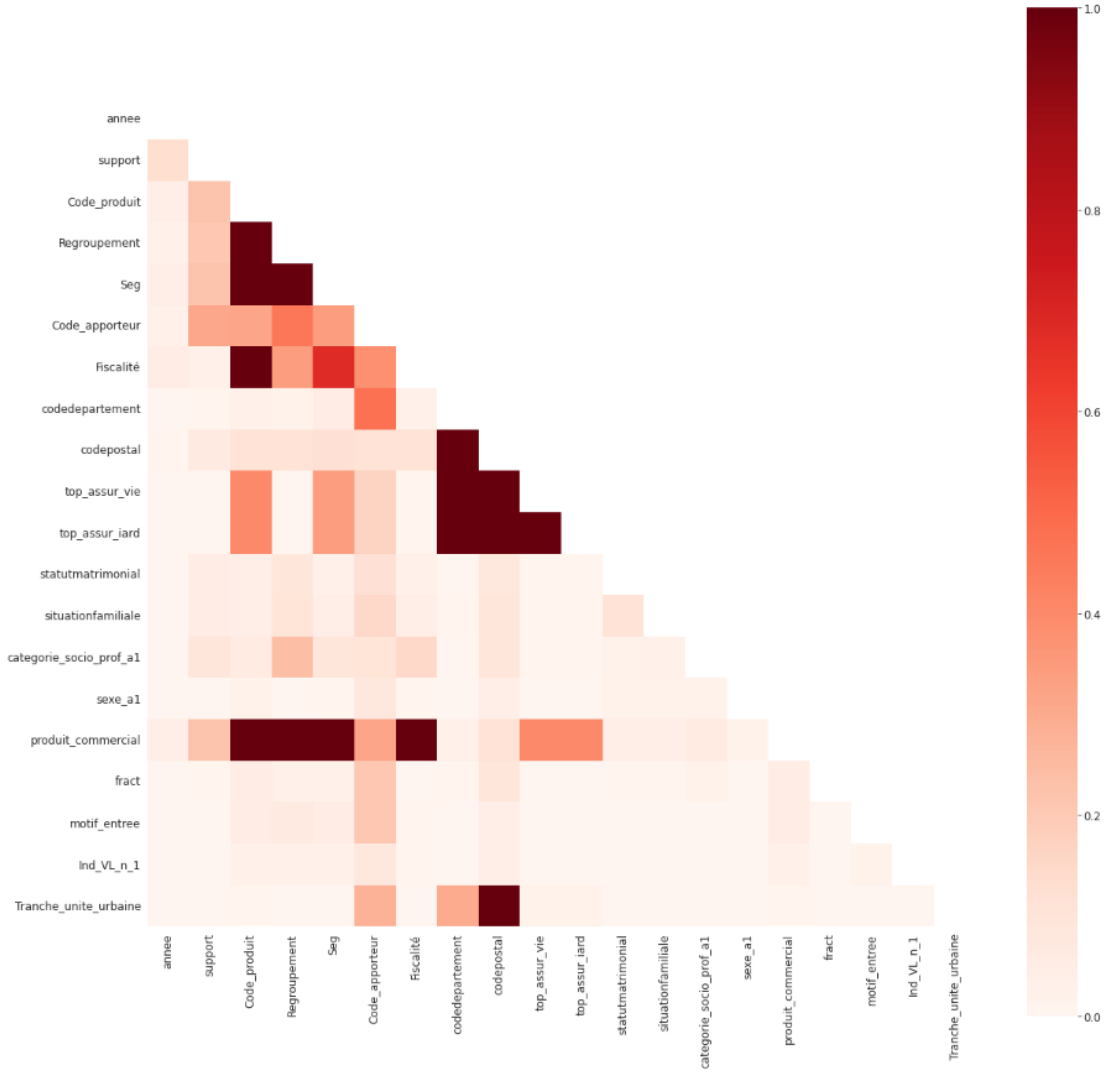


FIGURE 3.3 – Matrice de corrélation des variables qualitatives

Ces premiers résultats ont permis d'observer plusieurs corrélations à 100% entre le code produit, le regroupement, la segmentation et le nom du produit commercial. Ces variables donnent en effet la même information avec des degrés de précision différents. Le code produit et le nom du produit commercial sont les plus précis. En effet, ils renseignent à la fois sur le produit concerné, la méthode de sortie du capital constitué et le support. Ils présentent ainsi un nombre de modalités plus élevé de 25 contre 4 pour la variable Regroupement et 9 pour la variable Segmentation. A l'inverse, la variable Regroupement est beaucoup plus générale. Elle informe seulement sur la famille de produits : Néo Retraite,

Retraite Monnaie, Anthologie ou Graphic. Pour cette étude, la variable Segmentation a paru plus intéressante car elle informe sur la génération de chaque produit. Une analyse peut être mise en place afin de voir si certaines générations de produits sont plus porteuses et source de versements libres. L'information sur le support est de ce fait séparée dans une variable dédiée afin de dissocier les impacts en termes de versements.

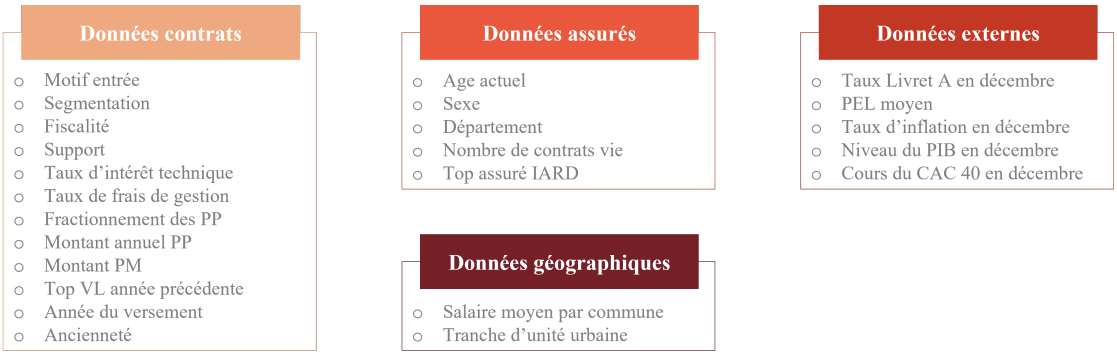
La variable code apporteur est également très corrélée aux variables produits et codes postaux et départementaux. Cette information semble cohérente. Pour autant, le manque d'interprétabilité de cette variable ainsi que les fortes corrélations observées ont conduit à la suppression de cette variable dans nos modèles.

Les variables top assuré vie et iard étaient mal renseignées, les informations ne correspondaient pas au nombre de contrats en cours des variables nombre de contrats vie et iard. De ce fait, l'analyse des corrélations n'a que confirmé l'idée de supprimer ces variables pour exploiter les variables numériques disponibles à ce sujet.

Les variables code postal et numéro de département sont également logiquement très corrélées. Comme évoqué précédemment, la variable numéro de département a été privilégiée.

3.3.2 Description des variables

La partie précédente a permis de sélectionner les variables les plus intéressantes pour la modélisation des versements libres. Le schéma ci-dessous synthétise les 24 variables retenues :



L'objet de cette partie est d'analyser les différentes variables disponibles afin d'appréhender notre base de données.

✓ Le motif d'entrée du contrat

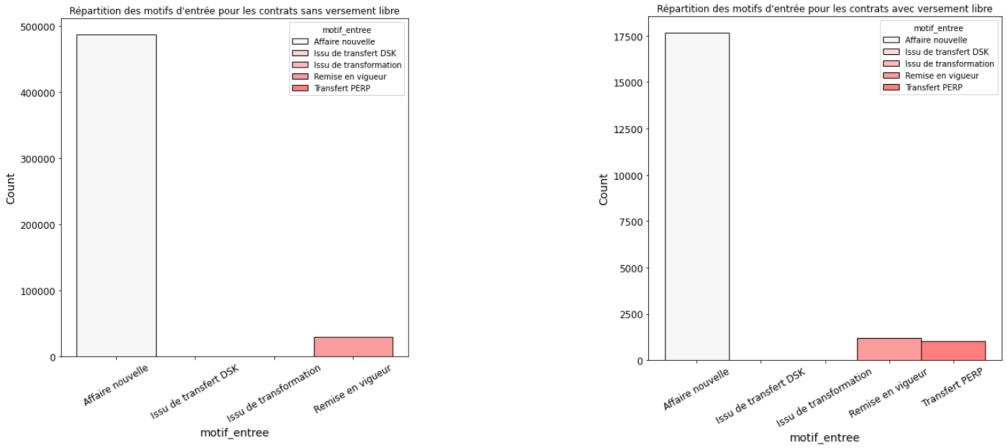


FIGURE 3.4 – Répartition des motifs d'entrée des contrats avec et sans versement libre

La majorité des contrats ont été ouverts suite à une affaire nouvelle. Certains contrats réalisant des versements libres proviennent parfois de remise en vigueur ou de transfert depuis les PERP. Les remises en vigueur correspondent à des contrats qui ont été suspendus et passés en statut de non en cours puis qui sont redevenus en cours. Cela peut survenir lorsqu'un assuré n'a pas payé une échéance de prime périodique. Lors de la pandémie de la Covid-19, plusieurs remises en vigueur ont été tolérées.
La modalité transfert PERP semble cependant isoler certains versements libres. En effet, l'annonce de la fin de commercialisation d'Anthologie a engendré beaucoup de versements libres afin d'anticiper les transferts vers les nouveaux PER.

✓ La segmentation des produits

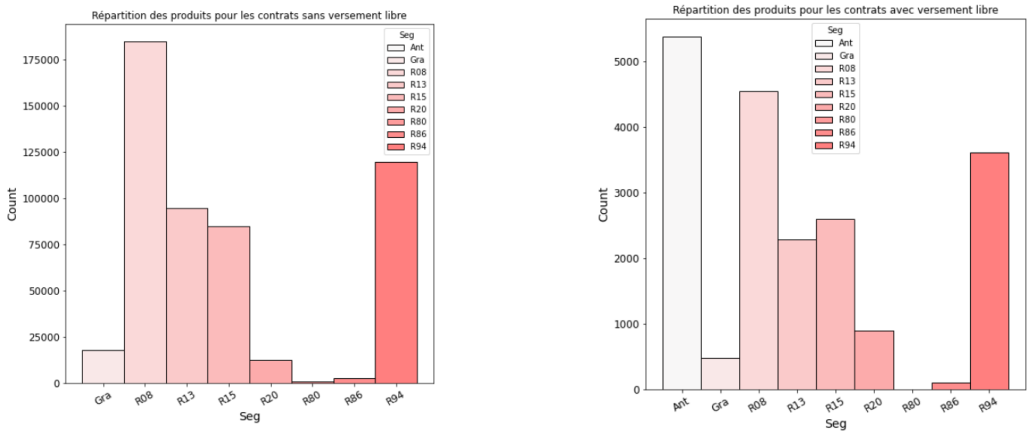


FIGURE 3.5 – Répartition des produits pour les contrats avec et sans versement libre

Globalement, les tendances des deux graphes 3.5 sont les mêmes. La survenance de ver-

sements libres semble proportionnelle au nombre de contrats dans chaque produit. Néanmoins, Anthologie se distingue des autres car tous les contrats ont effectué un versement libre. Cette information sera probablement très discriminante dans nos modèles par la suite.

Parmi ces produits, on retrouve les fiscalités suivantes :

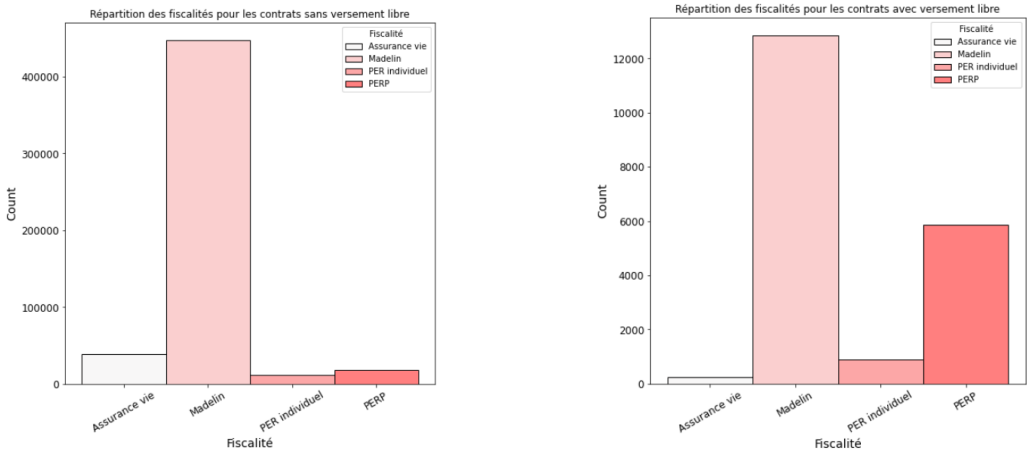


FIGURE 3.6 – Répartition des fiscalités pour les contrats avec et sans versement libre

Toute comme la variable produit, la fiscalité impacte la survenance d'un versement libre. Malgré la prédominance des Madelin, on constate que la proportion de PERP dans les données contenant les versements libres est nettement plus significative que dans les données sans versement libre.

✓ **Le support d'investissement**

	Base complète	Base sans versement libre	Base avec versements libres
EURO	64%	64%	61%
UC	36%	36%	39%

FIGURE 3.7 – Tableau récapitulant la ventilation euro et UC des investissements

Globalement, on observe plus d'investissements sur les placements en euro qu'en UC sur le tableau 3.7. Cela peut s'expliquer par le fait que les assurés privilégient la prudence. On constate sur le graphique 3.8 un changement des types d'investissement en fonction de l'ancienneté du contrat. Les données utilisées sont seulement celles de la Néo Retraite pour lesquelles l'euro et l'UC sont représentés. Ainsi, plus l'assuré se rapproche de la retraite, plus ses placements sont prudents.

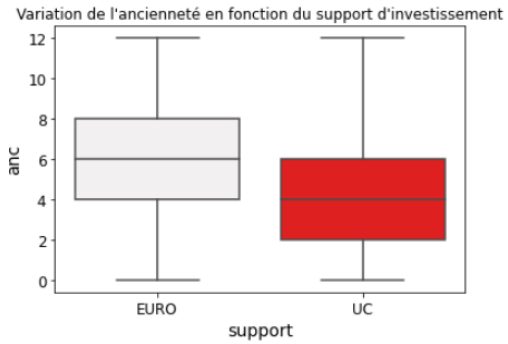


FIGURE 3.8 – Variation de l'ancienneté en fonction du support euro ou UC

✓ **Les TMG et taux de frais de gestion**

94.6% des contrats sans versement libre ont un taux d'intérêt technique de 0%. Ce pourcentage passe à 72.3% lorsque l'on considère les contrats avec des versements libres. On pouvait s'attendre à une dynamique inverse car en effet plus le taux garanti est élevé plus on réalise de versements libres. Ce phénomène peut s'expliquer par le fait que les PERP garantissent des taux d'intérêt négatifs et comptabilisent beaucoup de versements libres.

Les taux de frais de gestion dépendent essentiellement de la génération du produit. Plus la génération est récente plus le taux de frais de gestion est faible. De ce fait, ce sont davantage les nouvelles générations de produits qui pourraient possiblement comptabiliser plus de versements libres.

✓ **Le fractionnement et le montant de primes périodiques**

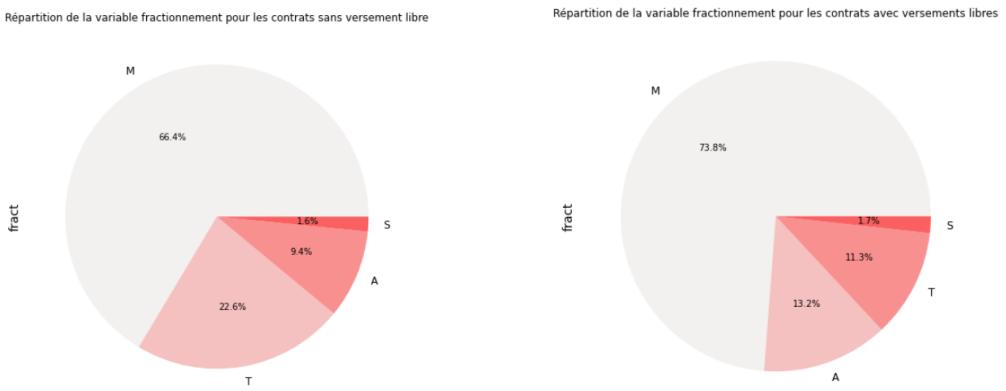


FIGURE 3.9 – Répartition de la fréquence des primes périodiques pour les contrats avec et sans versement libre

Les versements mensuels sont les plus fréquents pour 66.4% des contrats sans versement libre sur la figure 3.9. En considérant seulement les contrats avec des versements libres, ce pourcentage passe à 73.8%. Il en est de même pour les versements annuels, le pourcentage passe de 9.4% à 13.2% entre la base sans et avec les versements libres.

Cela signifie que deux types de comportements sont principalement observés :
- Les assurés qui sont actifs tous les mois sur leur contrat et qui effectuent en complément des versements libres
- Les assurés qui réalisent seulement un versement périodique et complètent avec des versements libres

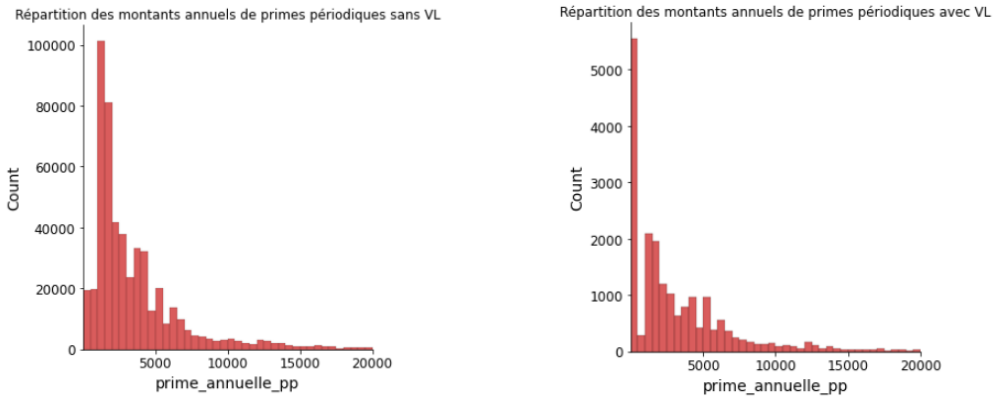


FIGURE 3.10 – Répartition des montants annuels de primes périodiques avec et sans versement libre

On constate sur la figure 3.10 que lorsqu'un versement libre est effectué, les montants annuels de primes périodiques sont moins élevés. Entre ces deux graphes, le quantile à 90% donne 12486€ dans le cas sans versement libre et 8647€ dans le cas avec versements libres. Moins le montant de prime périodique est élevé, plus l'assuré aura tendance à réaliser un versement libre.

✓ Les montants de provisions mathématiques

L'analyse des graphiques 6.5 ainsi que des différents quantiles a démontré que les niveaux de provisions mathématiques sont plus élevés dans le cas où un versement libre a été effectué. Le quantile à 95% pour les assurés réalisant des versements libres est de 69067€. Il est de 35412€ pour les assurés qui n'en réalisent pas. On constate donc des niveaux de provisions mathématiques plus élevés pour les assurés qui effectuent un versement libre.

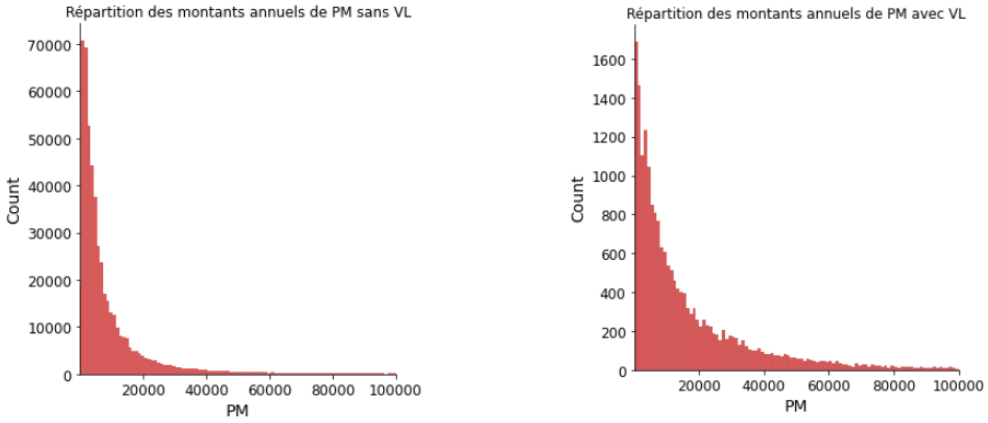


FIGURE 3.11 – Répartition des montants de provisions mathématiques pour les contrats avec et sans versement libre

✓ **L'ancienneté du contrat**

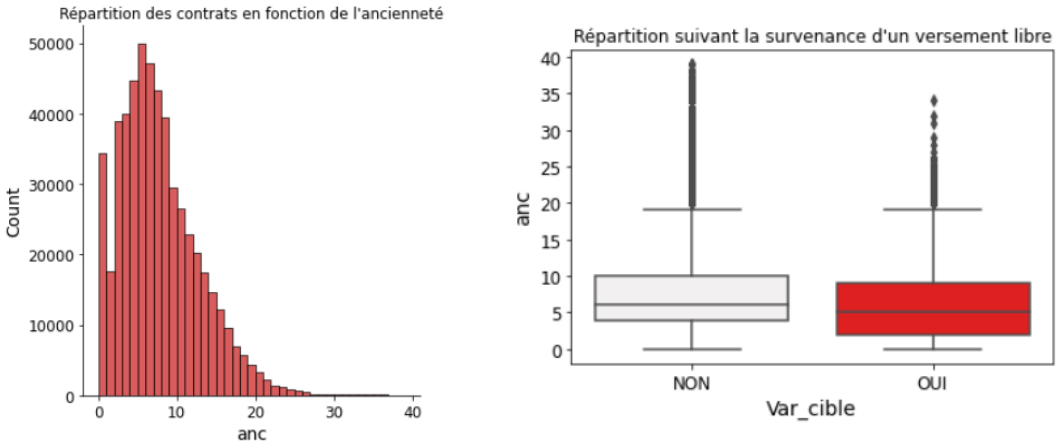


FIGURE 3.12 – Analyse de l'ancienneté des contrats

Comme en témoigne la première figure 3.12, la moitié des contrats en portefeuille ont une ancienneté comprise entre quatre et dix ans. L'ancienneté 0 est plus représentée du fait du lancement de la Retraite 20. En effet, la loi PACTE a encouragé les assurés à regrouper leurs contrats sur les nouveaux PER générant ainsi un afflux de nouvelles souscriptions. Ensuite, plus l'ancienneté augmente, plus le nombre de contrats observés diminue. Cela s'explique par le fait qu'il y ait moins de produits concernés lorsque l'ancienneté est plus grande. Par exemple, seuls les produits La Retraite 80 et 86 peuvent disposer d'une ancienneté de 30 années en portefeuille. De plus, sur ces produits, une majorité des assurés sont en phase de restitution.

Le second graphique 3.12 nous fait comprendre que l'ancienneté est légèrement inférieure dans le cas où un versement libre a été réalisé.

✓ **L'âge de l'assuré sur la période observée**

Globalement, les âges les plus représentés sur la première figure 3.13 sont concentrés entre 40 et 60 ans. L'âge de l'assuré sur la période observée est en moyenne de 48,9 ans. Lorsque l'on considère l'échantillon d'assurés réalisant des versements libres sur la seconde figure 3.13, l'âge moyen est de 51,5 ans. Les assurés réalisant un versement libre sont donc légèrement plus âgés.

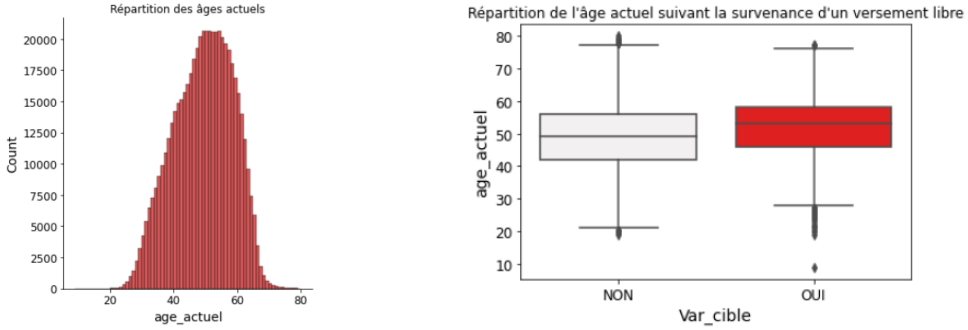


FIGURE 3.13 – Analyse de l'âge des assurés

✓ **Le sexe de l'assuré**

	Base complète	Base sans versement libre	Base avec versements libres
Homme	65%	64%	65%
Femme	35%	36%	35%

Le portefeuille contient plus d'hommes. Pour autant, il ne semble pas y avoir de différence dans les dynamiques de versements libres entre les hommes et les femmes.

✓ **Le département de l'assuré**

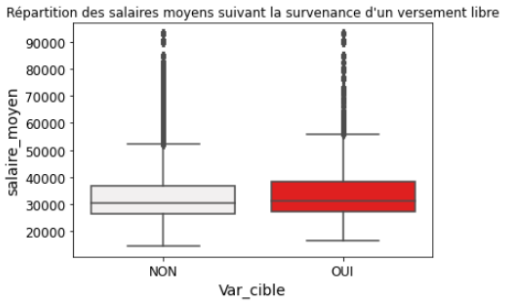
Base complète	Base sans versement libre	Base avec versements libres
75 Paris : 8.7%	75 Paris : 8.7%	75 Paris : 9.2%
59 Nord : 4.2%	59 Nord : 4.3%	69 Rhône : 5.9%
69 Rhône : 4.1%	69 Rhône : 4.0%	92 Haut de Seine : 4.0%

Les tendances de la base complète et de celle sans versement libre sont les mêmes car les volumes de versements libres sont faibles. Néanmoins, la zone géographique n'est pas tout à fait la même pour les assurés réalisant des versements libres. La région parisienne représente une part plus importante des assurés réalisant un versement libre.

✓ Le nombre de contrats vie et iard

La moyenne du nombre de contrats vie est la même entre les populations avec et sans versement libre, elle est respectivement de 1,45 et 1,40. Cette variable ne parait a priori pas discriminante. Il en est de même pour la variable top assuré IARD, 11% des assurés réalisant un versement libre ont au moins un contrat iard. Ce pourcentage est de 10% dans le cas contraire.

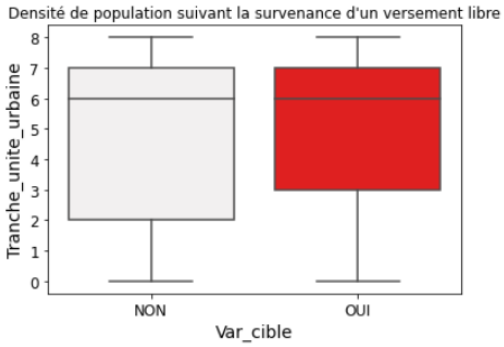
✓ Le salaire moyen des assurés en fonction de la commune de résidence



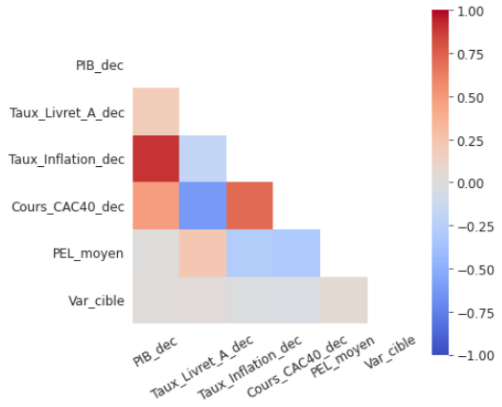
On constate des salaires légèrement plus élevés pour les assurés réalisant des versements libres. En moyenne, un assuré réalisant des versements libres réside dans une commune dont les revenus moyens sont de 35058€ tandis que celui qui n'en réalise pas est à 33787€.

✓ Densité de la commune de l'assuré

Les assurés résidant dans des unités urbaines de plus 10000 habitants ont davantage tendance à réaliser des versements libres. En effet, on constate que le *boxplot* est plus large pour les assurés ne réalisant pas de versements libres. Ainsi, les assurés des unités urbaines comprenant entre 5000 et 9999 habitants feraient a priori moins de versements libres.



✓ Les données financières



Les données financières semblent peu corrélées à la variable cible. Pour autant, cela paraît cohérent car les assurés sensibles aux fluctuations du marché sont plus rares. Les variables financières visent à prédire les versements conjoncturels. On peut supposer que ces variables ressortiront dans le cas où ce type de comportement serait perçu.

On constate également que les données financières sont très corrélées entre elles. Cela parait naturel car ces différents indicateurs interagissent conjointement.

3.3.3 Focus sur les variables cibles étudiées

La partie modélisation de ce mémoire va explorer deux pistes de prédiction des versements libres. Une première approche visera à modéliser une variable binaire de survenance ou non d'un versement libre. La seconde s'orientera sur un modèle actuariel de fréquence. Dans ce cas, la variable cible correspondra au nombre de versements libres par année réalisé par l'assuré.

✓ **La variable cible #1 : décision de versement libre (oui ou non) dans l'année**

La variable cible est déséquilibrée comme en témoigne la figure 3.14, la classe minoritaire représente seulement 3,7% des valeurs. En effet, seules 3,7% des lignes du dataframe comptabilisent un versement libre dans le portefeuille retraite individuelle. La base de données contient ainsi 19874 versements libres. Un point d'attention est néanmoins à noter car le nombre de lignes correspond à la granularité assuré, nombre d'années en cours et support. Ainsi, si on s'intéresse au nombre d'assurés concernés par les versements libres, on obtient que 9412 assurés ont réalisé au moins un versement libre entre 2015 et 2021 sur les supports euro ou UC. A l'inverse, 118096 assurés n'ont pas réalisé de versements libres dans tout ou une partie de l'historique sur les supports euro et UC. Les deux effectifs cités peuvent évidemment se rejoindre. En effet, un assuré peut avoir versé en 2015 puis plus en 2016. Dans ce cas, il se retrouvait dans les deux groupes.

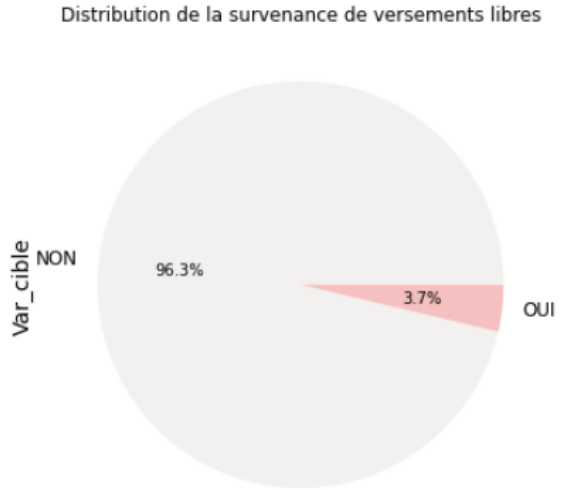


FIGURE 3.14 – Distribution de la première variable cible

✓ **La variable cible #2 : nombre de versements libres dans l'année**

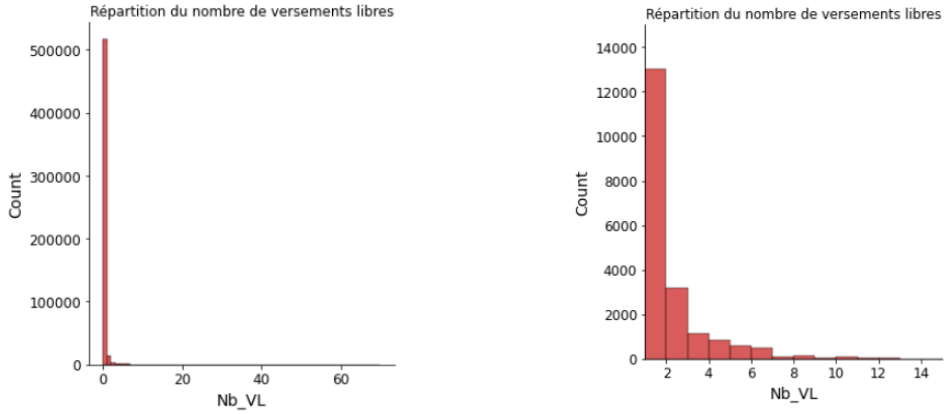


FIGURE 3.15 – Distribution de la seconde variable cible

La variable nombre de versements libres regroupe 31 modalités allant de 0 à 70. Comme cela a été évoqué dans le paragraphe précédent, la modalité 0 est la plus représentée. Le second graphique de la figure 3.15 permet de se concentrer sur les modalités avec au moins un versement libre. Au-delà de quinze versements par an, les effectifs sont plus faibles avec six cas au maximum. La problématique réside dans le fait que les modalités avec au moins un versement sont très peu représentées comme en témoigne le tableau ci-dessous :

Modalités	0	1	2	3	4	5
Effectifs concernés en %	96.3%	2.4%	0.6%	0.2%	0.2%	0.1%

Lors de la modélisation, nous tiendrons compte de l'exposition annuelle. Cela permettra ainsi de pouvoir réaliser une distinction entre les assurés réalisant deux versements sur six mois ou un an.

Chapitre 4

Projection des versements libres par une approche simplifiée

Cette partie vise à introduire une première approche de modélisation simple et facilement implémentable dans les outils internes afin de comprendre les comportements de versements libres.

4.1 Les étapes de construction des taux de versements libres

Dans un premier temps, l'objectif de cette étude a été de répliquer des travaux déjà réalisés en Epargne et de les adapter aux spécificités de la Retraite. L'idée est de construire des lois de versements libres à partir de l'historique fourni par nos bases de données avec des méthodes de calcul simples. La loi de versements vise à approcher le comportement futur des assurés en termes d'investissements via des versements libres non programmés et non modélisés par ailleurs. L'historique sélectionné comprend les versements libres effectués entre 2015 et 2021.
L'objectif de la construction des taux de versements libres est leur implémentation dans Prophet. Brièvement, Prophet est un logiciel permettant de projeter des flux sur un horizon de temps établi. Chez Generali, l'outil est séparé en deux parties entre Prophet Passif et Prophet ALS. Prophet Passif permet de projeter la partie déterministe à savoir principalement les primes, les provisions et les frais. Prophet ALS permet de projeter des indicateurs nécessaires à la construction du bilan comptable et du compte de résultat. Sa particularité par rapport à Prophet Passif réside dans le fait qu'une interaction est prise en compte dans le bilan entre l'actif et le passif par exemple. Cette partie de Prophet est stochastique.
L'objectif de cette partie est d'alimenter les tables d'hypothèses de Prophet afin de quantifier les impacts de l'intégration des versements libres de la retraite individuelle sur le Best Estimate et la PVFP.

4.2 Statistiques descriptives et définition du périmètre

Avant de débuter la construction des lois de versements libres, il était nécessaire de se baser sur quelques statistiques descriptives afin de s'orienter vers les périmètres les plus significatifs en termes d'investissements par versements libres.

Les travaux ont commencé sur le périmètre de la retraite individuelle. En effet, comme cela été évoqué à plusieurs reprises, le périmètre GB2000 est le système regroupant le plus de versements libres. Les produits collectifs ont été retirés de ce périmètre car les montants sont négligeables et de ce fait ne permettent pas de construire une loi.
Ensuite, certains produits sont plus sujet à réaliser des versements libres que d'autres au sein de ce système. Il est possible de réaliser des versements libres sur tous les produits ; pour autant certains comptabilisent peu ou pas de versements libres dans les mouvements des contrats. Cela dépend des avantages et des objectifs des différents produits. En effet, le versement libre, s'il est faisable pour tous les produits, reste une option. L'intérêt de l'assuré va donc influer sur ses investissements. Des montants supplémentaires peuvent ainsi être versés du fait d'incitations fiscales par exemple. D'autres produits vont encourager à verser un montant conséquent à la souscription et n'inciteront pas à investir au cours de la vie du contrat.

C'est donc en ce sens que des statistiques descriptives ont été réalisées pour déterminer si des lois devaient être construites pour tous les produits individuels de GB2000.
Le graphique 4.1 présente la répartition des versements libres réalisés entre 2015 et 2021 par produit :

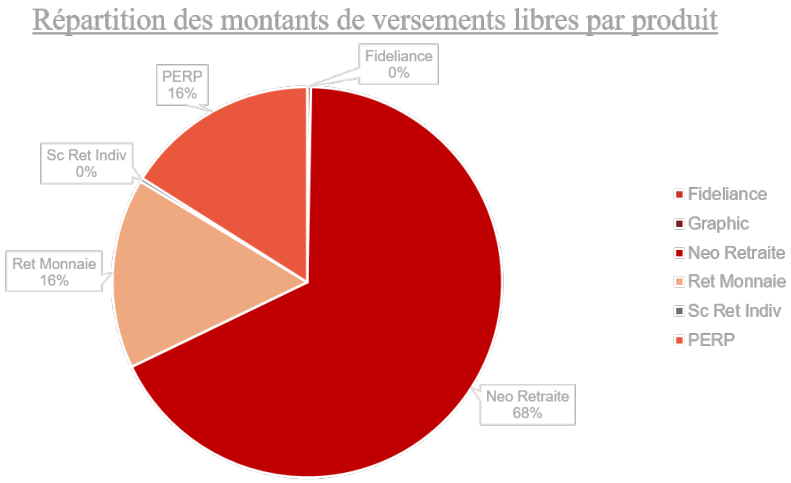


FIGURE 4.1 – Répartition des montants de versements libres par produit

Cette répartition a permis de constater que les volumes les plus significatifs en termes de versements libres sont : la Néo Retraite, la Retraite Monnaie et les PERP. Ces trois regroupements de produits contiennent presque 100% des versements libres du système GB2000. Seuls ces produits ont donc été présentés tout au long de ce mémoire. Les autres produits n'ont pas été considérés dans notre étude car les volumes sont trop faibles pour permettre de construire des lois.

4.3 Méthodologie de calcul

Le périmètre étant désormais connu, les taux de versements libres peuvent être construits. Les taux de versements libres sont calculés à partir du montant de versements libres et de provisions mathématiques. Les données doivent donc être récupérées en ce sens afin de comparer ces montants. Les calculs sont réalisés à partir de l'historique. Ainsi, l'étude comprend tous les versements et provisions mathématiques compris entre 2015 et 2021. Il est alors question de créer deux bases de données selon une maille qui sera discutée ultérieurement : la première est relative aux provisions mathématiques et la seconde est relative aux primes.

La méthodologie de calcul a été construite en suivant les travaux réalisés en épargne. Pour chaque regroupement, la logique de construction de la loi de versement libre est similaire. Les versements libres non programmés ainsi que les provisions mathématiques d'ouverture sont donc agrégés de la même manière.

Ainsi, le taux de versements libres se calcule en faisant le rapport entre les versements libres et les provisions mathématiques d'ouverture. Les formules implémentées sont les suivantes :

$$\forall\ N \neq 0\ alors\ VL_N\ =\ \frac{Montant\_Versement\_Libre_N}{Montant\_PM\_Ouverture_N} \tag{4.1}$$

$$Pour\ N = 0\ alors\ VL_0\ =\ \frac{Montant\_Versement\_Libre_0}{Montant\_PM\_Ouverture_1 - Montant\_Versement\_Libre_0} \tag{4.2}$$

où N est l'ancienneté du contrat
où le montant de PM d'ouverture N correspond au montant de PM de clôture N-1

Un traitement particulier est réalisé pour les affaires nouvelles car généralement la provision mathématique d'ouverture des contrats est nulle. Un proxy est effectué pour palier ce sujet.

FIGURE 4.2 – Détail des approximations sur les taux de versements libres

La table dans laquelle les taux de versements libres sont renseignés dans Prophet impose d'avoir des taux sur 35 années d'ancienneté. De ce fait, il a été nécessaire d'extrapoler les taux de versements libres comme l'explique la figure 4.2. L'hypothèse retenue a consisté à calculer une moyenne mobile sur les trois dernières années d'ancienneté disponibles. Cela a permis d'éviter de reporter des taux trop élevés sur les dernières anciennetés à disposition. Pour le cas particulier de la R20, seulement deux années d'historique étaient disponibles. Ainsi, ce sont les taux calculés pour les produits R08, R13 et R15 qui ont été retenus pour prolonger ceux de la R20. Néanmoins, pour les prochaines années, des anciennetés vont s'ajouter pour étoffer les données disponibles.

De plus, des taux à 0% ont été imposés sur certaines des lois pour les premières années d'ancienneté, comme cela est montré sur la figure 4.2. Les produits La Retraite fonctionnent par génération. De ce fait, lorsque la commercialisation du produit la Retraite 20 a débuté, il n'était plus possible de souscrire pour les anciennes générations de produits. Ainsi, il n'est aujourd'hui plus possible de constater des anciennetés 0 sur La Retraite 15 en portefeuille par exemple. Pour éviter de projeter des versements libres sur des anciennetés qui n'existent plus, des taux à 0% ont ainsi été utilisés. Il en est de même pour les PERP car Graphic n'est plus commercialisé depuis 2010 et Anthologie depuis fin 2019.

4.4 Choix de la maille d'agrégation des données

Une réflexion a été réalisée concernant la maille d'agrégation des lois de versements libres. Des contraintes outils limitent le choix de l'agrégation car l'objectif est de pouvoir projeter les versements libres dans Prophet afin d'analyser leurs impacts dans le bilan et le compte de résultat. En effet, la structure des modèles internes ne permet pas de modifier la maille de travail actuelle. Cette idée a été envisagée mais n'a pas été jugée

réalisable pour des contraintes de temps. En effet, la réalisation de ces modifications nécessiterait des travaux longs et le besoin de contrôler et de justifier les écarts entre le modèle officiel et le modèle modifié.
Dans Prophet, les taux de versements libres sont renseignés par ancienneté et code Prophet. La variable ancienneté est calculée en faisant la différence entre la date du versement libre et la date d'émission du contrat. Le code Prophet correspond à des regroupements de produits qui ont des caractéristiques similaires. Chez Generali, les caractéristiques sont la génération du produit, le support, le réseau de distribution et le fonds d'investissement. Les travaux effectués en épargne avaient suivi cette logique. Les lois avaient ainsi été construites par ancienneté, système et support. Cette étude avait considéré tous les systèmes de l'épargne car ce sont les périmètres avec le plus de versements libres.

Concernant les travaux en Retraite, il était possible de statuer sur la partie système, produit et support. L'objectif était de savoir si l'implémentation d'une loi plus fine au niveau produit ou regroupement de produits pouvait entraîner de meilleures prédictions. Comme évoqué précédemment, il n'était pas envisageable de modifier la structure des outils internes par ancienneté et code Prophet des lois de versements libres. La seule différence pouvait résider sur la gestion de la maille Code Prophet et de la construction des lois de versements libres. En effet, la maille code Prophet correspond à un produit et à un support en particulier. L'idée était donc d'observer les comportements des taux pour chaque produit et de constater si certains d'entre eux agissaient de la même manière ou non. Ensuite, certains regroupements pourraient être envisagés. Chaque loi serait affectée à un ou plusieurs codes Prophet.

Une réflexion a donc été menée pour déterminer comment adapter au mieux les travaux de l'Epargne à ceux de la Retraite. En Epargne, une loi commune a été créée pour chaque système. Pour la Retraite, des statistiques descriptives ont été mises en place afin de constater si des disparités de comportements étaient visibles entre les produits.

L'idée était de savoir si une segmentation plus fine pouvait améliorer les prévisions et les résultats issus de Prophet. Les produits suivant les mêmes dynamiques de taux de versements libres ont ainsi été regroupés. Cette étape a également permis de comprendre davantage les comportements des versements libres pour chacun des produits.

Les graphiques 4.3 suivant représentent tous les taux de versements libres obtenus pour chacun des produits retenus. On constate des disparités et des similitudes en fonction des regroupements de produits analysés.

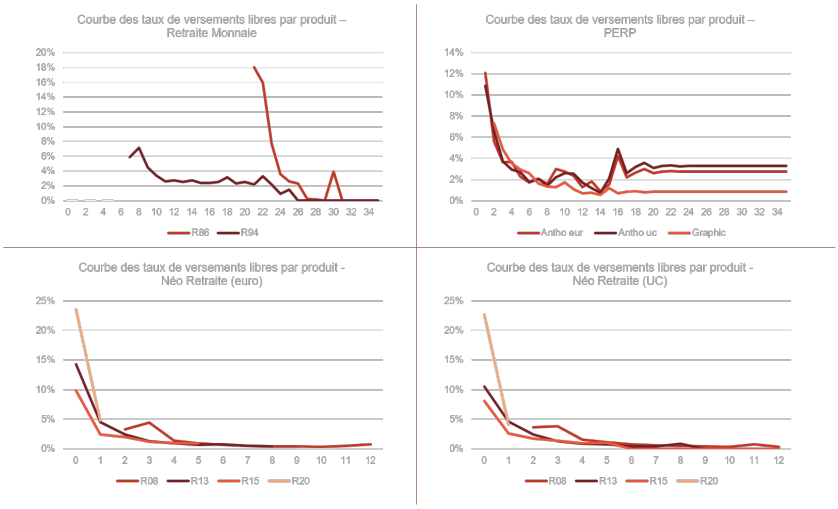


FIGURE 4.3 – Courbe des taux de versements libres par produit

Les représentations graphiques de la figure 4.3 ont permis de constater les points suivants :

— Le produit La Retraite 80 ne comptabilise pas suffisamment de versements libres pour construire une loi. Ce produit a donc été écarté de l'étude.

— Les autres produits La Retraite Monnaie ont des comportements très différents en termes de versements libres. Comme cela avait été évoqué dans la partie 2 concernant la présentation des produits Generali, la R86 garantit toujours des taux d'intérêt technique avantageux, ce qui n'est plus le cas pour la R94. Les taux de versements libres de la R86 sont ainsi nettement supérieurs à ceux de la R94.

— Les PERP ont des dynamiques relativement similaires. Pour autant, on constate un décrochage au niveau des tendances pour les anciennetés supérieures à 18 années. Les anciennetés avec peu de volumétrie n'ont pas été conservées dans l'extrapolation des taux. Cet écart a donc conduit à considérer des lois distinctes pour ces deux produits.

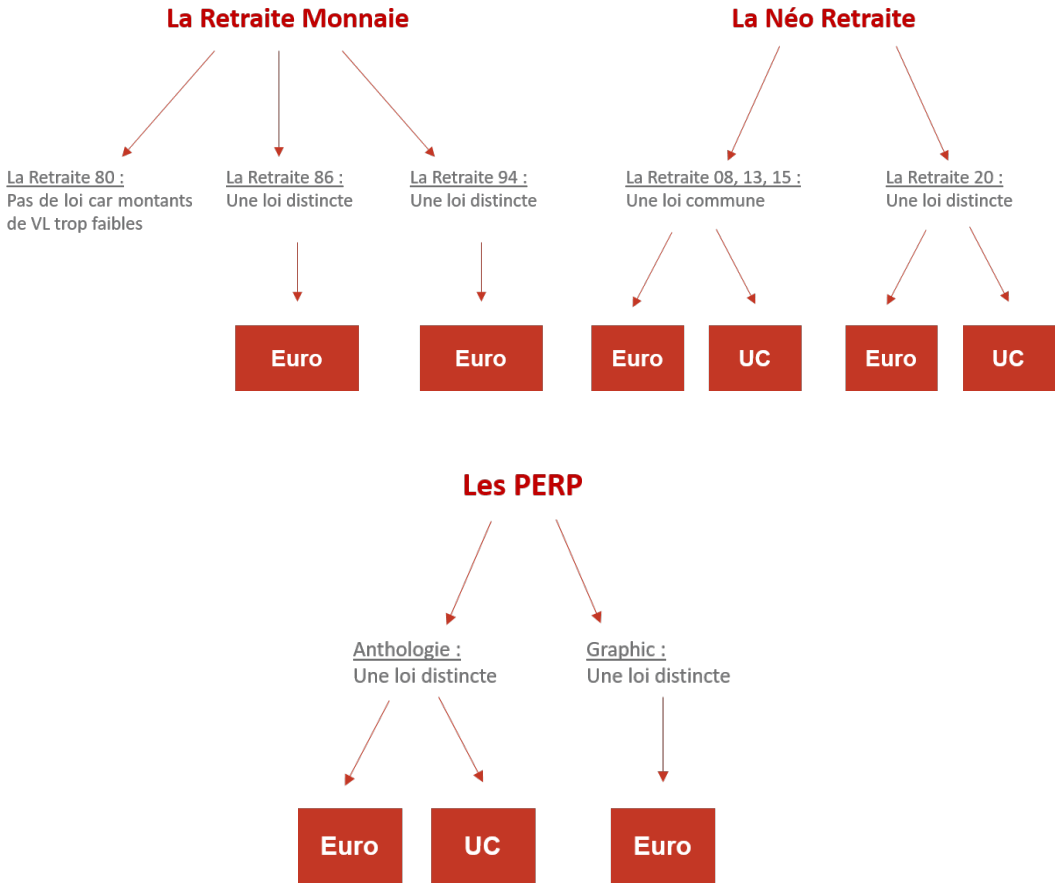
— Pour les produits Néo Retraite, les spécificités évoquées pour la R20 ressortent ici clairement. Les taux de versements libres observés sont nettement supérieurs à ceux visibles pour les anciennes générations. Néanmoins, les tendances des trois autres produits (R08, R13 et R15) sont globalement similaires.

— Concernant les supports euro et UC, un choix a été fait de créer une loi distincte car il est attendu que les comportements évoluent dans le temps. Les lois vont être mises à jour chaque année en ajoutant une année d'historique. C'est donc en ce sens que ce choix a été réalisé. Toutefois, les produits Néo Retraite enregistrent bien des écarts d'investissements entre l'euro et l'UC. Les taux de versements

libres investis sur des supports euro sont supérieurs à ceux de l'UC. Une étude interne à Generali a permis d'observer que ce constat est inversé pour les produits de l'épargne. Ainsi, cet aspect peut être intégré comme une première compréhension du comportement des assurés. Les supports d'investissements sont les mêmes entre l'épargne et la retraite mais il semble que les comportements des assurés diffèrent. Les assurés investissant sur leur contrat retraite semblent avoir une approche plus prudente que les assurés en épargne. Néanmoins, des incitations, sur la R20 notamment, encouragent à réaliser une part plus importante de versements sur des fonds UC du fait du contexte actuel de taux bas.

Finalement, cette analyse a permis de constater que des disparités étaient visibles entre les différentes générations de produits au sein des regroupements évoqués dans la partie précédente : Retraite Monnaie, Néo Retraite et PERP. Lorsque ce n'était pas le cas, une loi commune à tous les produits a été paramétrée.

Un choix a donc été fait d'ajouter la maille regroupement de produits pour fiabiliser nos lois. Les agrégations suivantes ont ainsi été retenues :



La maille d'agrégation choisie a donc généré la construction de neuf lois de versements libres dont la méthodologie de calcul est détaillée dans la partie précédente. Des lois brutes ont ainsi été construites et ont donné les résultats suivants :

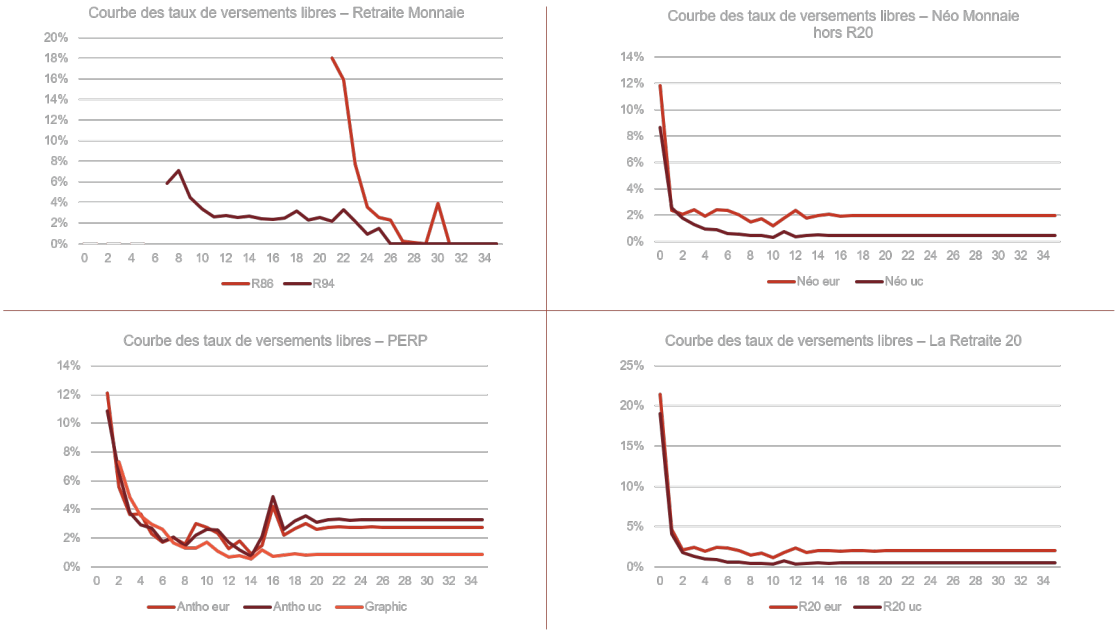


FIGURE 4.4 – Courbes des taux de versements libres retenues

4.5 Premiers constats plus détaillés sur les comportements des assurés réalisant des versements libres

Les dynamiques de taux de versements libres sont décroissantes. Plus l'ancienneté des contrats est élevée moins les assurés sont amenés à réaliser de versements libres. Ce phénomène est également motivé par le fonctionnement générationnel des produits. En effet, on constate une baisse des provisions mathématiques et des versements libres d'année en année sur les différents contrats. Cela peut s'expliquer par la réalisation de transferts ou le passage en rente.

Un second phénomène est également observable au sujet des supports euro et UC. Les assurés ont tendance à effectuer davantage de versements libres sur des supports en euro. Les résultats obtenus pour l'épargne montraient une dynamique inverse. Il est donc possible de supposer que les assurés détenant des produits retraite souhaitent privilégier des placements prudents et limiter leur exposition au risque. D'autant plus qu'en moyenne, les assurés souscrivent les produits retraite entre 40 et 50 ans. Au-delà de 20 années, très

peu d'UC sont constatés alors que pour un produit Epargne (souscrit plus tôt) des UC sont réalisés sur ces anciennetés.

Le dernier point remarquable concerne la Retraite 20. En effet, les taux de versements libres sont nettement supérieurs aux taux observés sur les anciennes générations. Ce point a d'ailleurs déjà motivé les choix d'agrégation de la partie précédente. Les versements libres considérés ne tenant pas compte des transferts internes ou externes, il est possible de constater dès à présent sur les deux années d'historique à disposition que les incitations fiscales liées à la loi PACTE ont porté à la hausse les investissements par versements libres sur les contrats retraite.
En effet, comme cela avait été présenté dans le choix de la maille d'agrégation, les produits La Retraite 08, 13 et 15 ont formé une seule et même loi car les dynamiques observées étaient semblables. Ainsi, le produit La Retaite 20 se distingue de toutes les anciennes générations de produits comptabilisées jusqu'à présent.

4.6 La projection des versements libres dans les outils internes

Une fois les lois de versements libres construites, il a fallu quantifier leur impact dans le bilan et le compte de résultat. Pour cela, les taux devaient être ajoutés aux hypothèses des modèles internes. Pour autant, les variables nécessaires à la projection des versements libres pour le périmètre de la retraite n'étaient pas créées dans Prophet Passif. Il a donc été nécessaire de les intégrer dans le code des structures existantes. L'objectif initial était de reprendre les codes réalisés en Epargne et de les adapter aux spécificités de la Retraite.

La modification des codes a été réalisée sur la partie Prophet Passif. La partie ALS n'a pas nécessité de modifications. Notons ici que les impacts des versements libres sont toujours obtenus en faisant la différence entre les résultats des *workspace* modifiés et le *workspace* officiels. En effet, il n'existe pas de variable permettant d'extraire les montants de versements libres à proprement parlé. La partie passif produit également des tables d'hypothèses qui sont nécessaires pour lancer la partie ALS.

Avant de développer les méthodes et les calculs utilisés pour ces travaux, l'environnement de travail Prophet va être explicité. Un environnement de travail ou *workspace* a été fourni par les équipes ayant travaillées sur le sujet en épargne. Ce *workspace* a servi de support dans le sens où tout ce qui avait été fait a été conservé. L'objectif était seulement d'ajouter le code relatif à la Retraite permettant la projection des versements libres. En d'autres termes, toutes les hypothèses et les codes de l'Epargne ont été gardés. La variable projetant les versements libres en Epargne ayant déjà été créée, elle a servi de support pour les travaux en Retraite.
Dans Prophet, un *run* est la compilation des codes permettant de projeter les flux à partir des tables d'hypothèses. Nous avons donc dû alimenter ces tables tout en s'as-

surant de la cohérence des codes ajoutés. En effet, la table d'hypothèses contenant les taux de versements libres existait déjà. Tous les codes produits de l'Epargne et de la Retraite étaient renseignés dans cette table. Pour autant, les taux relatifs au périmètre de la Retraite étaient égaux à zéro puisqu'ils n'avaient pas encore fait l'objet d'une étude. Les étapes précédentes de cette partie ont alors permis d'alimenter ces tables. Ainsi, les lois calculées ont été rattachées au code Prophet associé. L'étape suivante a donc été de modifier le code existant dans le modèle pour permettre la projection des versements libres de la Retraite à l'image de l'étude de l'Epargne.

La modification des codes existants dans les outils internes est complexe et nécessite une compréhension générale des dépendances entre toutes les variables du modèle. Finalement, les spécificités des produits la Retraite n'ont pas permis une simple recopie des codes implémentés en Epargne. Les travaux se sont peu à peu détachés de l'existant côté Epargne. En effet, les produits d'Epargne fonctionnent essentiellement en prime unique alors que dans le cas de la retraite, il faut également considérer les primes périodiques. La gestion mensuelle imposée par l'outil a donc rendu difficile la prise en compte de produits de périodicité trimestrielle par exemple.

Les versements libres ont donc été calculés annuellement et ont été répartis dans l'année de projection suivant la périodicité du produit. Le montant de versements libres a ensuite été ajouté à la prime annuelle déjà calculée dans l'outil. Un point d'attention a été porté au sujet des écarts de récurrence. En effet, lors de la construction des codes, nous nous sommes aperçus que les versements libres étaient pris en compte dans la provision de clôture mais pas dans la provision d'ouverture. En principe, le calcul réalisé est le suivant :

$$\begin{gathered}\text{PM Ouverture} + \text{Prime Périodique brute} + \text{Intérêt Technique brut} + \text{Participation} \\ \text{aux bénéfices} - \text{Prestations} = \\ \text{PM Ouverture} + \text{Versement Libre net} + \text{Chargement sur Versement Libre} \\ + \text{Prime Périodique nette} + \text{Chargement sur Prime Périodique} + \text{Intérêt Technique} \\ \text{net} + \text{Participation aux bénéfices} - \text{Prestations}\end{gathered} \tag{4.3}$$

Concrètement ici, les versements libres nets n'étaient pas renseignés dans la vision nette de la récurrence ci-dessus. Les versements libres modélisés ne respectaient pas l'égalité de récurrence de la formule et cela a créé un résultat technique inexpliqué égal au montant de versements libres observé dans le compte de résultat.

Ensuite, du calcul de la prime annuelle découle ensuite la prime brute, la prime d'inventaire et la provision mathématique. Des écarts de récurrence ont été observés car les montants de versements libres étaient correctement ajoutés à la prime annuelle mais pas à la provision mathématique.

Ainsi, de nombreux tests ont été produits jusqu'à obtenir un univers de projection juste.

Cette partie a nécessité beaucoup de travail et de compréhension de l'environnement Prophet ainsi que des variables. De nombreuses solutions ont été testées avant d'aboutir à la solution finale. Ainsi, une fois le code modifié, il a été possible d'entrer les tables d'hypothèses modifiées contenant les taux de versements libres de la retraite.

Dans un premier temps, la partie Passif a été fiabilisée car les sorties de ce *run* sont nécessaires pour lancer le *run* sous ALS. Dès lors, l'interaction actif et passif a été analysée via le bilan et le compte de résultat après le lancement du run ALS.

Une fois les codes sous Prophet validés, les *runs* sont gérés dans des classeurs excel pour analyser les sorties et les projections. Des *templates* officiels ont donc été alimentés une fois les *runs* du périmètre de la Retraite exécutés.

4.7 Analyse des résultats

Dans cette partie, les chiffres ont été modifiés pour des besoins de confidentialité sans induire de changements sur les analyses faites.

Comme évoqué précédemment, la première partie de cette étude a visé à modifier le code de la partie Passif de Prophet afin de pouvoir projeter les versements libres. Les tables d'hypothèses ont été renseignées avec les taux de versements libres construits. L'idée désormais est d'analyser les impacts de l'ajout des versements libres de la retraite individuelle.
L'historique de versements libres ayant servi à construire les taux a été reproduit dans le tableau 4.5 afin d'analyser les projections réalisées par Prophet sur l'année 2021 dans la première colonne.

Produits retenus	Prévisions 2021	Observations 2021	Observations 2020	Observations 2019	Observations 2018	Observations 2017	Observations 2016	Observations 2015
Anthologie	1 745K€	11 237K€	7 824K€	2 979K€	3 622K€	4 921K€	4 676K€	5 340K€
Graphic	147K€	180K€	199K€	157K€	159K€	209K€	253K€	411K€
R08	5 881K€	2 959K€	2 976K€	3 796K€	4 433K€	1 252K€	789K€	1 738K€
R13	3 359K€	1 312K€	1 477K€	1 791K€	2 174K€	641K€	60K€	1 160K€
R15	4 255K€	2 330K€	3 020K€	3 475K€	4 114K€	3 531K€	1 265K€	373K€
R20	2 674K€	9 046K€	3 547K€	204K€				
R80	0K€				1K€			
R86	144K€	124K€	138K€	296K€	268K€			
R94	12 894K€	4 746K€	5 568K€	5 933K€	7 241K€			
	31 100K€	31 933K€	24 749K€	18 630K€	22 012K€	10 555K€	7 044K€	9 022K€

FIGURE 4.5 – Historique des versements libres ainsi que les prédictions obtenues sur 2021

La dernière ligne du tableau 4.5 correspond à la somme des versements libres tous

produits confondus. Tout d'abord, les projections réalisées sur 2021 semblent cohérentes au global avec un écart de seulement 3%. Plus précisément, 31 933 143,29€ ont réellement été versés en 2021 et Prophet a prédit 31 110 282,71€. Donc sur le total, les résultats semblent très bons.

Néanmoins, lorsque l'on regarde le détail des prédictions par produit, des écarts parfois importants sont constatés. Finalement, la compensation de ces écarts permet d'obtenir de bons résultats.
Les produits Anthologie et La Retraite 20 sont largement sous-estimés. A l'inverse, les produits La Retraite 94, 08, 13 et 15 sont surestimés. Les produits Graphic et La Retraite 86 sont néanmoins correctement estimés.

D'où proviennent ces écarts ?

- Anthologie : la fin de commercialisation de ce produit a eu lieu en fin d'année 2019. Lors de cette annonce, de nombreux assurés ont soit versé beaucoup plus qu'à l'accoutumée soit souscrit ce produit pour, dans les deux cas, anticiper le transfert vers les nouveaux PER.
  Ce phénomène, de par son aspect exceptionnel, ne pouvait pas être anticipé par nos lois brutes construites sur la base d'une moyenne.
- Les produits La Retraite 08, 13 et 15 : les montants prédits par Prophet se rapprochent de ceux de l'année 2018. On constate que finalement les taux de versements libres construits sont proches d'une moyenne des versements libres de l'historique. Or, ces produits ne sont plus commercialisés et voient leur nombre ainsi que leur montant de versements libres décroître dans le temps. Cette baisse n'est pas anticipée par l'outil. La figure 4.6 témoigne de la baisse des taux de versements libres dans le temps pour les différents produits qui ne sont plus commercialisés. En effet, jusqu'en 2020, malgré l'enchaînement des générations de produits R8, R13, R15, on constate un niveau de taux de versements libres plutôt à la baisse sur notre période de construction des lois. Le niveau moyen de taux de versements libres projeté en 2021 par le modèle Prophet se rapproche donc, toutes choses égales par ailleurs, du taux de versements libres moyen observé sur notre période de construction des lois (correspondant au niveau moyen observé en 2018/2019). Cela a pour effet une surestimation des versements libres pour le périmètre R8, R13, R15 sur notre projection 2021.
- La Retraite 20 : les montants sont largement sous-estimés mais cela s'explique par le fait que l'historique disponible est très faible (seulement deux années). Cela ne permet pas une bonne prédiction des montants réels.
- La Retraite 94 : les versements libres sont fortement surestimés. Les arguments sont proches de ceux développés pour les produits La Retraite 08, 13 et 15. En effet, l'historique a dû porter à la hausse les montants prédits. Cela vient s'ajouter au fait que les provisions mathématiques de ce produit sont plus importantes du fait de l'ancienneté du produit. Ainsi, les montants ont été largement surestimés.

FIGURE 4.6 – Évolution des taux de versements libres par année

Cette baisse des taux de versements libres dans le temps est l'hypothèse privilégiée pour justifier les écarts observés. Anthologie relève davantage d'un phénomène exceptionnel qui n'était donc pas prévisible. Cela peut davantage relevé de versements libres conjoncturels incités par la mise en place de la loi PACTE et de la commercialisation de la R20. Nous avons souhaité vérifier cette hypothèse, pour cela nous réalisons un test de sensibilité des prédictions de l'année 2021 du modèle Prophet par rapport à la période de construction des modèles. Pour cela un historique réduit a été conservé en enlevant les années 2015 à 2018. Nous nous attendons ainsi à voir les montants prédits baisser.
De ce fait, les taux de versements libres des produits La Retraite et Néo Retraite (hors La Retraite 20) ont été reconstruits avec un historique réduit de 2019 à 2021. Les résultats suivants ont ainsi été obtenus :

Produits retenus	Prévisions 2021	Observations 2021	Observations 2020	Observations 2019	Observations 2018	Observations 2017	Observations 2016	Observations 2015
Anthologie	1 745K€	11 237K€	7 824K€	2 979K€	3 622K€	4 921K€	4 676K€	5 340K€
Graphic	147K€	180K€	199K€	157K€	159K€	209K€	253K€	411K€
R08	5 722K€	2 959K€	2 976K€	3 796K€	4 433K€	1 252K€	789K€	1 738K€
R13	2 876K€	1 312K€	1 477K€	1 791K€	2 174K€	641K€	60K€	1 160K€
R15	3 294K€	2 330K€	3 020K€	3 475K€	4 114K€	3 531K€	1 265K€	373K€
R20	2 674K€	9 046K€	3 547K€	204K€				
R80	0K€				1K€			
R86	133K€	124K€	138K€	296K€	268K€			
R94	11 003K€	4 746K€	5 568K€	5 933K€	7 241K€			
	27 595K€	31 933K€	24 749K€	18 630K€	22 012K€	10 555K€	7 044K€	9 022K€

FIGURE 4.7 – Présentation des résultats après réduction de l'historique considéré dans la construction des lois

Comme attendu, les montants prédits sont plus bas avec cette nouvelle hypothèse. Cette méthode a permis de baisser les versements libres prédits pour ces produits des proportions ci-dessous :

La Retraite 08	La Retraite 13	La Retraite 15	La Retraite 86	La Retraite 94
-2,7%	-14,4%	-22,6%	-7,7%	-14,7%

Évolution des montants de versements libres prédits avec cette nouvelle hypothèse

Les prédictions ont donc été globalement améliorées même si des écarts sont toujours présents. Des limites avaient déjà été démontrées quant à la gestion des versements libres par ancienneté. Ces résultats ont permis de confirmer ce constat. Ce point a motivé la suite de ce mémoire pour chercher d'autres variables pouvant influer sur les investissements par versements libres. Ce point sera abordé plus en détails ultérieurement.

Ces résultats ont été validés par les équipes IFRS 17 et ont conduit à la projection dans l'outil ALS. L'objectif était de pouvoir analyser les interactions actif et passif et d'extraire des indicateurs tels que le Best Estimate et la PVFP notamment.
Pour rappel, le Best Estimate correspond à la valeur actuelle de l'ensemble des flux de trésorerie futurs. La PVFP correspond à la Valeur Actuelle des Profits Futurs.

Après avoir renseigné les sorties Passif présentées précédemment, les résultats ALS suivant ont été obtenus :

TOTAL + VL_RET - TOTAL			
Portefeuille IFRS 17	BE	PVFP	MV
Produits La Retraite	-40 587 125	40 407 430	-179 695
PERP Grafic	310 690	-310 690	0
PERP Anthologie	-1 050 229	1 048 507	-1 722
TOTAL La Retraite	-40 595 220	40 415 526	-179 695
TOTAL PERP	-739 341	919 035	179 695
TOTAL	-41 334 561	41 334 561	0

FIGURE 4.8 – Présentation des impacts de l'intégration des versements libres dans Prophet ALS

L'impact des versements libres de la retraite individuelle est obtenu en faisant la différence entre le *run* incluant les versements libres et le *run* officiel sans la gestion des versements libres. Finalement, deux tableaux ont permis d'extraire celui présenté dans la figure 4.8. Seules les lignes modifiées ont été présentées. Nous nous sommes néanmoins assurés que les modifications n'aient pas créées d'écarts sur les autres postes. Seuls quelques frottements jugés négligeables ont été observés dans les résultats. Ces montants ont été validés par les équipes.

L'impact des produits La Retraite sur le Best Estimate est de -40 595 220€. Il est de 40 415 526€ pour la PVFP.
Les PERP sont représentés dans les *pools* PERP Graphic et PERP Anthologie. L'impact des versements libres hors produits La Retraite est de -739 341€ sur le Best Estimate et 919 035€ pour la PVFP. Au global les versements libres ont engendré -41 334 561€ pour le Best Estimate et 41 334 561€ pour la PVFP.

En termes de montants de versements libres par système, GB2000 retraite représente 2,5% des versements. L'ajout des versements libres sur la retraite individuelle génère un supplément de résultats futurs de +41,3 millions d'€, cela représente un delta de +0,06% pour le Best Estimate et de +1,12% sur la PVFP, comme en témoigne le tableau 4.7.

Best Estimate	PVFP	MV
+0,06%	+1,12%	0%

TABLEAU 4.7 - Impacts de l'ajout des versements libres dans les indicateurs analysés

Le gain obtenu est relativement faible mais nous nous attendions à un tel résultat. En effet, les produits de retraite représentent une part relativement faible par rapport aux produits d'épargne. Néanmoins, les montants obtenus auraient pu être légèrement plus élevés si les montants créés par Graphic n'étaient pas négatifs. Après avoir approfondi ce point, il apparaît que ce produit n'est pas rentable à cause de la manière dont sont modélisés les frais sur ce produit, l'apport de versements libres (donc de prime) n'a fait qu'augmenter la perte projetée sur ce périmètre. Les résultats sont ainsi cohérents avec les sorties du run officiel.

4.8 Validation des résultats obtenus et des ordres de grandeur

Lorsque tous les *runs* ont été tournés, les différents résultats produits ont dû être validés. Une réflexion a donc débuté pour trouver des indicateurs de validation des ordres de grandeur. Au total, trois tests ont été mis en place.

Vérification de l'adéquation entre le run officiel et le run modifié sans projection de versements libres

Initialement, le *workspace* de l'Epargne avait été fourni pour nous servir de support pour nos travaux. La première vérification a constitué à s'assurer de l'adéquation du *workspace* officiel de l'Epargne et de celui modifié avec des taux de versements libres nuls pour la Retraite.
Plus concrètement, l'idée était de s'assurer qu'en figeant les taux de versements libres du périmètre de la retraite à zéro, les montants de primes obtenus étaient équivalents à ceux observés dans le *run* officiel.

Schématiquement, le *template* utilisé prenait la forme de la figure 4.9.

	Workspace officiel	Workspace modifié	
Code Prophet	PP + VL Epargne	PP + VL Epargne	Delta
FCAEFA	530 000	530 000	-
FCAEPI	720 000	720 000	-
FCAFQM	1 800 000	1 800 000	-
FRBP94	-	-	-
FRBP08	-	-	-
FRBP13	-	-	-
FCBPEE	2 900 000	2 900 000	-
FCBPER	100 000	100 000	-
FCBPFE	6 000 000	6 000 000	-
FRBP20	-	-	-

FIGURE 4.9 – Validation des montants entre le *workspace* officiel et le *workspace* modifié sans versement libre

Par soucis de confidentialité, les montants ont été modifiés. Concrètement, dans le tableau 4.9, les montants non vides représentent les primes additionnées des versements libres projetés sur les périmètre Assurance vie et Epargne. Les montants vides correspondent aux produits retraite. Comme les taux de versements libres en Retraite sont nuls, les montants associés le sont également. Dans ce premier test, les versements libres en Epargne ne sont pas impactés et ceux de la retraite sont bien nuls. On peut donc en conclure que l'ajout de la variable simulant les versements libres de la retraite dans Prophet n'a pas entraîné de régression sur le *workspace* Prophet.

Vérification de la cohérence des taux de chargement

La seconde vérification a constitué à s'assurer de la cohérence des taux de chargement projetés par Prophet Passif sur le périmètre Retraite dans le compte de résultat. Cette vérification a à nouveau été effectuée en comparaison du *workspace* officiel et du *workspace* modifié avec les projections de versements libres en Retraite cette fois-ci.
En effet, les taux de chargements appliqués sont toujours les mêmes quel que soit le niveau de prime car ils sont fixés dans les tables d'hypothèses. Il était donc intéressant de regarder si les taux de chargement étaient restés les mêmes avant et après les traitements réalisés dans le code.
Les taux de chargement ont été calculés comme suit :

$$\text{Taux de chargement} = \frac{\text{Chargements d'acquisition}}{\text{Prime brute}}$$

Les taux de chargement ont été calculés via cette formule sur le *run* officiel dans un premier temps. Les chargements d'acquisition ainsi que les primes brutes sont des informations récupérées dans les sorties excel de Prophet Passif. Ensuite, ce même calcul a

été fait sur le *run* modifié. L'adéquation des taux de chargement d'acquisition a également été validée. Des graphiques ont été réalisés afin de s'en assurer sur un historique de projection suffisamment grand. Ainsi, sur un horizon de 30 ans, les taux de chargement d'acquisition ont été comparés à l'aide représentation graphiques et de tableaux prenant la forme 4.10 pour s'assurer qu'il n'y ait pas d'écarts.

	2021	2022	2023	2024	2025	...
Run officiel	0,07%	0,07%	0,07%	0,07%	0,07%	...
Run modifié	0,07%	0,07%	0,07%	0,07%	0,07%	...
Test validé ?	OUI	OUI	OUI	OUI	OUI	...

FIGURE 4.10 – Validation des taux de chargement pour un produit fictif

Vérification de la cohérence entre les versements libres futurs simulés et les lois entrées dans Prophet

Ce dernier test visait à vérifier les ordres de grandeur des taux de versements libres entre les lois entrées et les lois sorties de Prophet. L'idée est de s'assurer que les ordres de grandeur concordent et que Prophet applique des taux cohérents sur les projections par rapport à nos hypothèses.

Cette dernière vérification a nécessité plus de travail car les indicateurs nécessaires n'étaient pas récupérables directement. Les taux de versements libres projetés ont dû être recalculés à partir des sorties Prophet. Les montants de versements libres ont été obtenus par différence entre le *workspace* officiel et le *workspace* modifié comme précédemment. Les provisions mathématiques ont été récupérées par code produit. Pour autant, comme cela avait été évoqué, les provisions mathématiques utilisées pour projeter des versements libres sont celles pour lesquelles le taux minimum garanti est négatif ou nul. Or l'assiette obtenue par la méthode précédente considère toutes les provisions mathématiques. De ce fait, la même assiette de provisions devait être récupérée.

Ainsi, les calculs ont été réalisés à la maille SPCode. Les identifiants SPCode permettent en effet d'obtenir des informations sur les produits de manière plus fine et d'identifier les contrats avec des taux garantis négatifs ou nuls. Ensuite, les montants de provisions mathématiques et de versements libres ont été obtenus et sélectionnés via ces codes. Les montants de versements libres en sortie ont ainsi pu être calculés.

Pour conclure sur cette dernière vérification, des graphiques comme le 4.11 ont été construits. Il permet d'analyser les taux moyens de versements libres projetés par rapport à ceux observés dans nos lois brutes. Il nous a permis de valider globalement la cohérence entre l'*input* du modèle et l'*output*. En effet, on constate que les tendances sont très proches. Les quelques écarts observés proviennent principalement de la méthode de comparaison des taux car les taux moyens en entrée ne sont pas pondérés par le poids

des anciennetés dans la projection.

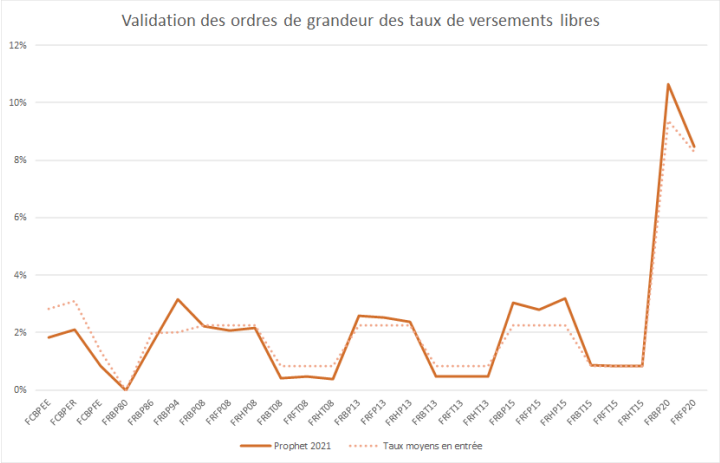


FIGURE 4.11 – Taux de versements moyens en entrée et sortie de Prophet

Au sein de ce même test, une dernière vérification a été abordée. Pour le test précédent, les horizons de projection étaient les mêmes, c'est-à-dire que les montants comparés correspondaient à l'année 2021 dans les projections. Un deuxième test a donc visé à confirmer que lorsque l'horizon de projection s'agrandit, les taux de versements libres sortis de Prophet diminuent. En effet, plus l'ancienneté augmente plus les taux de versements libres sont faibles. Un second visuel a donc été produit pour s'en assurer.
Les horizons de projection retenus ont été 2022, 2025, 2030 et 2040 comme en témoigne la figure 4.12.

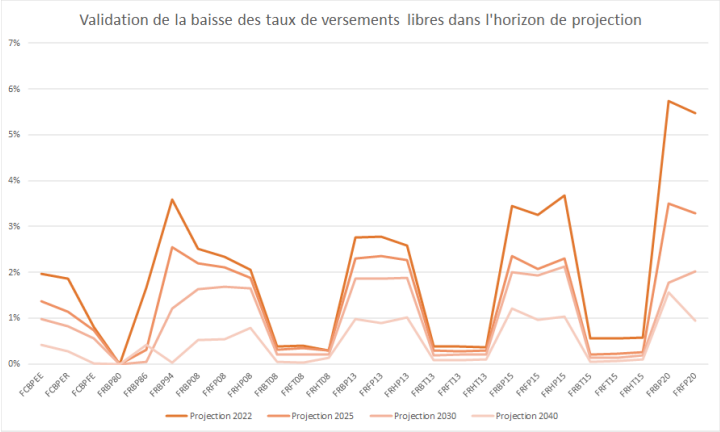


FIGURE 4.12 – Taux de versements en sortie de Prophet sur plusieurs années de projection

Le graphique 4.12 a donc permis de confirmer que les taux de versements libres baissaient dans le temps de projection. De ce fait, les différentes vérifications réalisées ont bien permis de valider les travaux produits dans la partie Prophet Passif. Ces tests ont clôturé cette partie de l'étude.

Chapitre 5

Présentation des méthodes de machine learning utilisées

Cette partie va s'attacher à la présentation des méthodes de *machine learning* retenues pour la modélisation des versements libres. Quatre méthodes ont été retenues à savoir : un GLM pénalisé, des arbres de décision de type CART, une forêt aléatoire et un Histogram Gradient Boosting. Nous avons choisi d'aller des modèles plus simples et plus lisibles vers des modèles plus complexes, plus performants et moins lisibles. Néanmoins, ces différents algorithmes seront couplés avec une méthode robuste d'analyse du comportement de ces modèles : SHAP.

5.1 Présentation des méthodes de machine learning

5.1.1 GLM pénalisé

Brièvement, un rappel va être réalisé concernant le principe d'un GLM [15]. L'objectif de cette méthode est de prédire des variables $Y_1, ..., Y_n$ définies sur un espace probabilisé $(\Omega, \mathcal{F}, P)$ mutuellement indépendantes à partir d'autres variables explicatives X_k définies sur ce même espace probabilisé. Cela revient à écrire le modèle sous la forme suivante :

$$g(\mathbb{E}[Y|x_1, ..., x_n]) = \sum_{k=1}^{n} \beta_k x_k$$

où on applique une fonction de lien strictement monotone et dérivable g et où l'on a n variables explicatives x_i pour $i = 1, ..., n$.

Dans cette étude, un modèle GLM de Poisson est appliqué à Y $\in \mathbb{N}$. Ainsi, la fonction log a été utilisée comme fonction lien dans ce modèle. [16]

Les modèles mis en oeuvre ont également utilisé une variable offset. Une variable offset est une variable pour laquelle on associe un coefficient de 1 dans le GLM. Son utilisation

dans les modèles est très fréquente dans le domaine de l'assurance et plus particulièrement dans les problématiques de prédiction de fréquence. L'objet de cette étude porte sur la modélisation des versements libres. L'une des approches a donc visé à créer un modèle de fréquence puis un modèle de survenance du versement.

Pourquoi utilise-t-on une variable offset ? L'utilisation d'une variable offset permet de passer d'un modèle de comptage à un modèle de fréquence. La distribution de la variable relative au nombre de versements libres suit une dynamique de Poisson. La démonstration ci-dessous va permettre de comprendre davantage l'intégration d'une variable offset :

$$\mathbb{E}[Y|x_1, ..., x_n, e] = e \times \exp\left(\sum_{k=1}^{n} \beta_k x_k\right)$$

où e correspond à la variable offset, c'est-à-dire dans ce mémoire à l'exposition. Cela revient donc à ajouter une variable explicative au modèle tout en fixant son coefficient β à 1. [16]

Finalement, l'idée est de modéliser la fréquence de versements libres. Ainsi, une distinction sera opérée entre un assuré réalisant un versement libre en étant présent pendant six mois et un second réalisant un versement libre en étant présent une année complète dans le portefeuille.

GLM Pénalisé L'objectif d'un GLM Pénalisé [8] est de sélectionner ou de réduire le nombre de variables pertinentes des modèles. Ces modèles sont très intéressants pour analyser la causalité entre les variables d'un modèle de machine learning.

Les régressions pénalisées peuvent être utilisées dans les cas suivants : soit le modèle présente un nombre conséquent de variables explicatives, soit nous souhaitons restreindre le nombre de variables n pour ne conserver que les plus pertinentes. Ainsi, deux possibilités sont envisageables : soit le nombre de variables est réduit passant ainsi de n à p soit nous cherchons à trouver la meilleure fonction f telle que $f(x) = y$ avec le moins de variables explicatives possible afin d'obtenir la meilleure approximation de la variable cible Y_i.

L'idée est donc de contrôler le bruit des coefficients en évitant les corrélations entre les variables et de développer le meilleur modèle possible.

A la différence des modèles d'un GLM classiques, le GLM pénalisé vise à minimiser deux composantes :

- la somme des écarts quadratiques entre les valeurs prédites et observées comme à l'accoutumée à laquelle la contrainte suivante est ajoutée selon le modèle utilisé (Ridge ou Lasso) :
- le paramètre de réglage λ multiplié par la somme des coefficients au carré en cas de Ridge Regression. Mathématiquement, cela revient à minimiser la quantité suivante :

$$\sum_{k=1}^{n}(\hat{y}_k - y_k)^2 + \lambda \sum_{l=1}^{p} |\beta_l|^2$$

- le paramètre de réglage λ multiplié par la somme de la valeur absolue des coefficients en cas de Lasso Regression. Mathématiquement, cela revient à minimiser la quantité suivante :

$$\sum_{k=1}^{n}(\hat{y}_k - y_k)^2 + \lambda \sum_{l=1}^{p} |\beta_l|$$

où p est le nombre de prédicteurs et β les coefficients de la régression.

Ainsi les coefficients de la régression sont pénalisés de telle sorte à ne pas être trop grands ou que leur nombre soit réduit. En d'autres termes, la somme des coefficients β_k est pénalisée afin qu'elle soit inférieure à une certaine valeur fixée par la contrainte. Cette pénalité est réglée et optimisée par le paramètre λ.

Quelles sont les différences entre les méthodes Lasso et Ridge ? La contrainte appliquée pour les modèles Lasso est plus restrictive du fait de la valeur absolue par rapport à l'élévation au carré de la méthode Ridge. Dans les deux méthodes, les coefficients β_k sont testés afin de savoir s'ils seront conservés ou non. La méthode Ridge aura plus facilement tendance à faire tendre les coefficients vers des valeurs proches de zéro quand la méthode Lasso optera plus radicalement pour une valeur de zéro. Les contraintes de la méthode Lasso sont donc plus fortes.

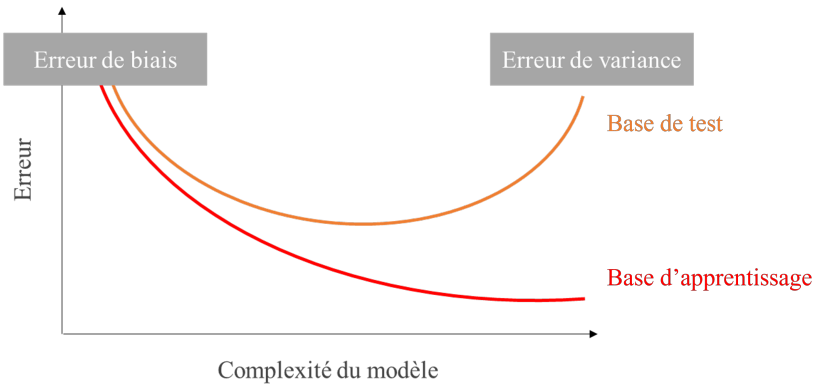


FIGURE 5.1 – Phénomène de sous-apprentissage et de sur-apprentissage

En quoi les régressions pénalisées sont-elles plus efficaces que les régressions classiques dans notre cas ? En machine learning, il existe deux types d'erreur : l'erreur de biais pour le sous-apprentissage des modèles et l'erreur de variance pour le sur-apprentissage. Ces deux erreurs sont schématisées sur la figure 5.1.
Les modèles classiques visent à réduire seulement l'erreur de biais. Les modèles pénalisés sont plus efficaces car ils réduisent à la fois l'erreur de biais et l'erreur de variance. Cet algorithme permet surtout essentiellement de trancher davantage dans la sélection des variables les plus significatives. Cet aspect réduit les deux types d'erreur connus. Pour autant, cette approche ne s'adapte pas à toutes les problématiques.

5.1.2 Arbre CART

Les arbres de décision [5] sont prisés du fait de leur facilité d'interprétation sous la forme d'un arbre. Cette seconde approche permet d'enlever la contrainte de la linéarité du modèle tout en conservant un modèle facilement interprétable.
Cet algorithme relève d'un apprentissage automatique. Il consiste à isoler au maximum une population homogène par rapport à la valeur cible en sélectionnant les variables les plus discriminantes. Pour rappeler rapidement le fonctionnement de cette méthode, nous illustrons ce propos avec une base de données en dehors du sujet de mémoire, mais qui est très souvent utilisée dans une optique pédagogique car il présente peu de variables explicatives. Ainsi, l'exemple ci-dessous vise à identifier les populations ayant survécues ou non au naufrage du Titanic [1]. Pour cela, trois variables ont été retenues : le sexe, l'âge et le nombre de frères et soeurs. Ces trois variables sont celles qui permettent de séparer au mieux les populations ayant survécues ou non.

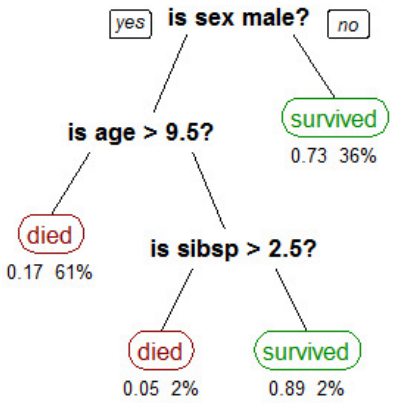


FIGURE 5.2 – Exemple d'un arbre de décision sur la survie des passagers du Titanic[1]

1. Source : https ://cedric.cnam.fr/vertigo/Cours/ml2/coursArbresDecision.html

Un arbre de décision est composé de noeuds et de branches. Un noeud permet de scinder les échantillons d'individus en deux branches distinctes en choisissant la variable la plus discriminante. Il existe trois types de noeuds : **le noeud racine** qui est le premier noeud de l'arbre ; **le noeud interne** : qui est poursuivi par un autre noeud, il est dans le coeur de l'arbre de décision et **le noeud terminal** qui clôture l'arbre et affiche la prédiction de l'échantillon concerné par les différentes segmentations opérées.
De plus, un arbre de décision implique une lecture de haut en bas. En d'autres termes, il utilise une **représentation hiérarchique**.

Les arbres de décision se distinguent en fonction du type de modélisation. On retrouve ainsi les arbres de classification et de régression [18]. Ces deux modèles vont être développés davantage dans les prochains paragraphes.

Les arbres de classification Les arbres de classification visent à prédire les classes de la variable cible. Cet algorithme se rapproche de méthodes plus classiques telles que l'analyse discriminante. Dans le cas de l'exemple 5.3, cette méthode permet d'obtenir un score de réalisation d'un versement en fonction des variables ancienneté du contrat, montant de provisions mathématiques et réalisation d'un versement l'année précédente. Cet arbre est utilisé à titre d'exemple et les variables ont été considérées aléatoirement. Les prédictions finales de l'arbre de décision sont réalisées en fonction de la classe majoritairement représentée dans le noeud terminal.

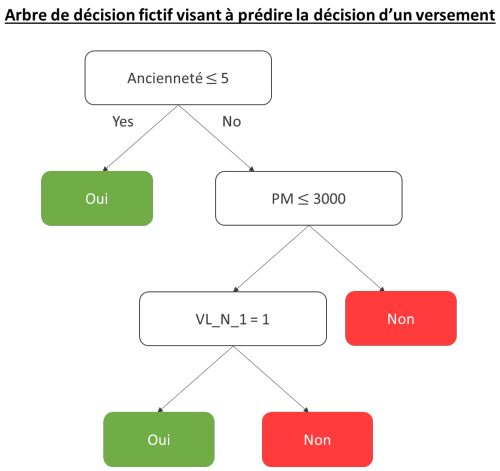


FIGURE 5.3 – Exemple fictif d'un arbre de décision dans le cadre de cette étude

Pour en revenir aux arbres de décision, contrairement aux autres algorithmes prédictifs, ils étudient chaque variable tour à tour afin d'opter pour la suite de variables la plus pertinente au sens de la prédiction. Ces divisions univariées peuvent être effectuées pour tous les types de variables explicatives qu'elles soient numériques, catégorielles etc.

Afin de réaliser la sélection de la variable du noeud en question, des tests statistiques sont mis en place. Chaque noeud est optimisé par l'algorithme selon des critères établis (indice de Gini, entropie ou khi deux). L'indice de Gini est utilisé pour calculer l'impureté d'un noeud. L'impureté correspond au fait que les échantillons soient clairement séparés dans les feuilles terminales. L'indice de Gini est donc calculé comme suit :

$$G_i = 1 - \sum_{k=1}^{n} (p_{i,k}^2)$$

où $p_{i,k}$ correspond au ratio du nombre d'individus de la classe k parmi la population du i^{eme} noeud.
Un noeud est considéré comme étant pur lorsque l'indice de Gini est égal à 0. Ainsi, l'indice de Gini permet de quantifier le gain de pureté du choix d'une variable dans un noeud. L'objectif est de connaître le coût de la décision. Deux indices de Gini sont calculés. On mesure tout d'abord l'impureté du noeud de gauche puis celle du noeud de droite. Ces mesures sont ensuite pondérées afin de comparer le gain avec l'impureté du noeud d'origine.
Pour ce faire, l'algorithme parcourt toutes les variables ainsi que les valeurs de segmentation pour optimiser cette évolution du gain de pureté.

Les arbres de régression Les arbres de régression sont utilisés dans le cas d'une variable numérique. Ainsi, ils visent à prédire une valeur numérique et non une classe comme c'était le cas pour les arbres de classification. Contrairement aux algorithmes de classification, la variable prédictive est une quantité (continue ou discrète). En considérant un exemple en lien avec cette étude, lorsqu'une feuille terminale prédit deux versements libres réalisés dans l'année, il est possible de comparer et d'ordonner ce nombre par rapport à un autre. L'exemple 5.4 ci-dessous permet d'illustrer plus simplement l'algorithme :

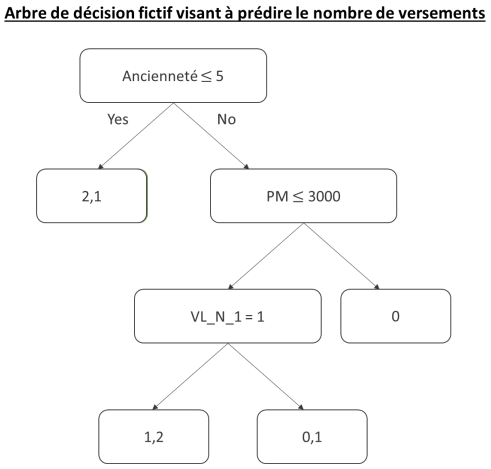


FIGURE 5.4 – Exemple fictif d'un arbre de décision dans le cadre de cette étude

A nouveau dans cet exemple, les résultats sont purement inventés et visent seulement à expliciter le fonctionnement d'un arbre de régression. L'objectif ici est de prédire le nombre de versements libres. On constate donc aisément que l'algorithme prédit un nombre et donc qu'il est possible d'ordonner les résultats obtenus.

Les tests statistiques réalisés pour séparer les échantillons des données sont également différents de ceux utilisés pour les méthodes de classification. Le critère d'erreur retenu est l'erreur quadratique moyenne. Cette erreur est calculée comme suit :

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(\hat{y}_k - y_k)^2$$

où $\hat{y}_k$ correspond aux valeurs prédites par l'algorithme et y_k aux valeurs observées dans le jeu de données.
Le noeud est ensuite créé en minimisant la moyenne pondérée des erreurs quadratiques calculées. L'objectif est le même que pour le modèle de classification à savoir de maximiser la pureté des feuilles terminales de l'arbre de décision.

5.1.3 Forêts Aléatoires

La méthode de forêts aléatoires [4] est un algorithme d'apprentissage supervisé. Elle est basée sur la modélisation de plusieurs arbres de décision CART formés par la méthode *bagging*. Les forêts aléatoires combinent plusieurs modèles d'apprentissage afin d'obtenir de meilleures performances. Ils résolvent ainsi le problème d'instabilité des arbres de décision en réduisant l'erreur de variance. En effet, généralement les arbres de décision sont *overfittés* et s'adaptent mal aux nouvelles données. Les forêts aléatoires représentent une amélioration du *bagging* pour les arbres de décision CART.

Les notions essentielles pour comprendre les forêts aléatoires :

— Les méthodes d'*ensemble learning* : sont des méthodes combinant plusieurs méthodes de *machine learning* afin d'améliorer les performances de prédiction. Parmi ces méthodes, on retrouve les méthodes de *bagging*, *boosting* et de *stacking*. Les prédictions réalisées par ces modèles sont agrégées et permettent d'obtenir une unique prédiction retenue par l'algorithme.

— Le *bagging* : utilise des procédures indépendantes au cours desquelles le modèle est entraîné. Ainsi, les données utilisées sont sans cesse revues par tirage aléatoire avec remise afin que chaque modèle s'entraîne avec un jeu de données différent. La variable cible est prédite avec ces variables et ces individus par le modèle d'apprentissage. Généralement, les arbres de décision sont profonds et possèdent une erreur de variance très élevée. Une fois toutes les procédures réalisées, l'instabilité des arbres de décision est réduite. L'erreur de variance, et donc par définition le sur-apprentissage, est réduite.

Comment fonctionne l'algorithme de forêts aléatoires ? La forêt aléatoire est un peu plus qu'un simple bagging. En effet, c'est là tout son intérêt. Des arbres de décision classiques sont entraînés. Mais à chaque création de noeud, on tire aléatoirement une nouvelle liste de variables explicatives disponibles. C'est ce qui fait que les arbres générés sont très peu corrélés entre eux et que la variance du modèle global diminue considérablement par rapport à un arbre CART classique. Finalement, on calcule la moyenne des prévisions de chaque arbre de décision.

Le schéma 5.5 permet de synthétiser le fonctionnement d'une forêt aléatoire :

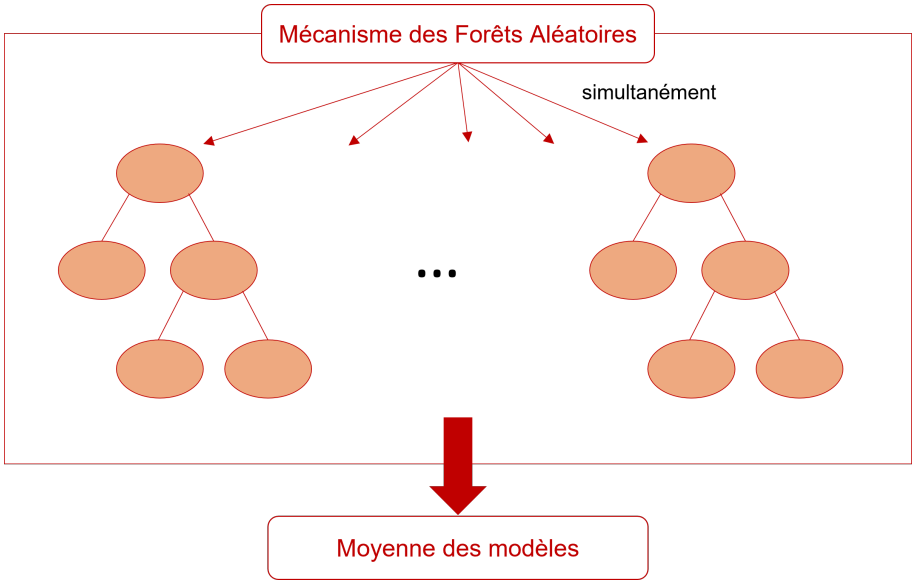


FIGURE 5.5 – Fonctionnement d'une forêt aléatoire

Les forêts aléatoires en quelques formules : Comme évoqué précédemment, le *bagging* vise à construire plusieurs arbres de classification sur des échantillons *bootstrap*. Désormais, considérons une base d'apprentissage décrite par les conditions suivantes :

- Soit Y la variable à expliquer et $X_1, ..., X_p$ les variables explicatives ;
- Les données d'apprentissage sont décrites par les couples (x_k, y_k), $x_k \in \mathbb{R}^p, y_k \in \mathbb{R}, k = 1, ..., N$;
- La variable explicative y_k peut être continues ou discrètes ;

Soit $\Phi(x)$ un modèle de prédiction appris sur un échantillon de données où z est donné par les couples $\{(x_k, y_k)\}_{k=1}^n$.
Une fois la base de données détaillées, les étapes du bagging sont les suivantes :

1. j tirages aléatoires avec remise sont effectués dans la base d'apprentissage, ils sont appelés échantillons bootstrap et notés z_k avec $k = 1, ..., j$;

2. Pour chaque échantillon k on calcule le modèle $\Phi_{z_k}(x)$
3. La variable explicative Y est prédite en agrégeant les différentes décisions sur chacun des z_k par :
   (a) $\hat{\Phi}(x) = \frac{1}{j}\sum_{k=1}^{j} \Phi_{z_k}(x)$ dans le cas d'un modèle de régression ;
   (b) $\hat{\Phi}(x) =$ Vote majoritaire parmi les $\Phi_{z_k}(x)$ dans le cas d'un modèle de classification.

Les modèles sont estimés à partir de l'erreur OOB (Out Of Bag). Des individus sont isolés de manière aléatoire dans la phase d'apprentissage pour la validation du modèle. Chaque modèle construit propose une prédiction pour ces individus, la solution la plus représentée ou la moyenne est finalement retenue. L'erreur OOB correspond donc au nombre de bonnes prédictions.

Quels sont les avantages et les faiblesses de cet algorithme ? Ce modèle est très utilisé car il est assez simple de par sa conception et permet d'obtenir de bonnes performances aux vues de l'erreur de biais et de variance. Il a cependant parfois été remis en question par les méthodes de boosting.

5.1.4 Histogram Gradient Boosting

La méthode *Histogram Gradient Boosting* [20] est une technologie d'apprentissage automatique. C'est un algorithme d'apprentissage et d'amplification de gradient. Contrairement aux algorithmes de forêts aléatoires, l'*Histogram Gradient Boosting* relève d'une approche de boosting.

Quel est le principe du boosting ? Le boosting est une procédure étape par étape qui consiste à créer des règles au fur et à mesure. Ainsi, chaque nouveau modèle cherche à améliorer les prédictions du précédent. Pour ce faire, des poids plus importants sont attribués aux individus mal classés. Le modèle final est donc plus robuste du fait de l'ajout systématique de règles qui finissent par n'en constituer qu'une seule. Les modèles de boosting cherchent à réduire l'erreur de biais. Ainsi, l'algorithme cherche à éviter le problème de sous-apprentissage des modèles, c'est-à-dire que les hypothèses fixées ne sont pas suffisamment contraignantes pour permettre de prédire correctement la variable cible.

La principale difficulté rencontrée sur les modèles de boosting est la vitesse d'exécution. Contrairement aux forêts aléatoires, les modèles de boosting sont entraînés de manière séquentielle. Les temps de traitement sont donc plus longs. Pour résoudre cette problématique, des modèles de gradient boosting optimisés ont été développés.

Comment comprendre le passage du boosting au gradient boosting ? Concrètement, les méthodes de gradient boosting sont des algorithmes basés sur une fonction de perte convexe, comme cela est le cas pour l'algorithme *Adaboost*. Ainsi, les modèles sont

construits de manière séquentielle afin d'améliorer les performances de l'algorithme au fur et à mesure des pas ajoutés. Les améliorations des performances sont réalisées grâce au gradient de la fonction de perte. L'idée est de chercher le meilleur pas de la descente de gradient permettant d'obtenir les meilleures performances pour l'algorithme.

En revenant à la méthode présentée dans ce paragraphe, l'*Histogram Gradient Boosting* est également basé sur des arbres de décision. Il cherche à trouver le meilleur *split* possible pour réduire l'erreur de prédiction. Pour cela, des arbres sont modélisés à nouveau de manière séquentielle. Si les performances sont améliorées alors le modèle est conservé. Cette caractéristique définit les algorithmes de boosting.

La particularité de cet algorithme de *machine learning* réside dans le fait qu'il soit plus rapide que les méthodes classiques de gradient boosting. La méthodologie est largement accélérée du fait de la discrétisation des variables sous la forme d'histogrammes. Cette discrétisation des variables est la particularité des algorithmes Light GBM et Histogram Gradient Boosting. Cela permet d'accélérer significativement les temps de calcul. Ce mécanisme est illustré ci-dessous :

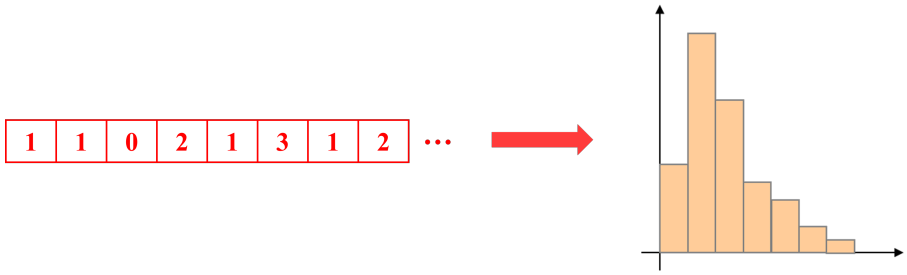


FIGURE 5.6 – Exemple fictif de la discrétisation d'une variable en histogramme

Ainsi, les valeurs des variables en entrée sont regroupées dans des compartiments permettant de réduire considérablement le nombre de valeurs. Cette discrétisation va permettre de calculer la fréquence de chaque variable dans une période établie. Ces périodes sont appelées *bins*.
L'idée est de minimiser l'erreur de prédiction du modèle lorsqu'il est additionné avec les modèles précédents. Il faut donc fixer les résultats attendus pour chaque modèle en réduisant l'erreur de prédiction totale. On parle d'amplification de gradient car chaque nouveau modèle contribue à réduire l'erreur de prédiction.

Cet algorithme se rapproche beaucoup d'un Light GBM. En effet, cet algorithme cherche également à réduire le temps de calcul en réduisant le nombre de modalités des variables. Ces dernières sont donc discrétisées par compartiments. L'avantage de cette méthode est double car il réduit le temps de calcul et n'impacte pas les performances du modèle.

Le Gradient Boosting en quelques formules : L'objectif de cet algorithme est de minimiser la fonction de perte. Cette fonction peut être définie différemment suivant les modèles, dans ce mémoire c'est la fonction de loss-ratio qui a été retenue et qui est définit comme suit : $L = \frac{1}{n}\sum_{k=1}^{n}(\hat{y}_k - y_k)^2$ où $\hat{y}_k$ correspond aux valeurs prédites par l'algorithme et y_k aux valeurs observées dans le jeu de données. L'objectif est donc de minimiser cette quantité en trouvant la meilleure prédiction des $\hat{y}_k$.
Cette étape peut être résumée ainsi :

$$Step_1 = argmin_{\hat{y}_k} \sum_{k=1}^{n} L(\hat{y}_k, y_k)$$

L'objectif est donc de trouver les $\hat{y}_k$ qui minimisent la fonction de loss-ratio. Pour cela, il faut étudier la dérivée de la fonction $\phi(\hat{y}) = \sum_{k=1}^{n} L(\hat{y_k}, y_k)$.

5.2 Présentation des méthodes de rééchantillonnage

⇒ Le SMOTE

Le Synthetic Minority Oversampling TEchnique, SMOTE [6], est une méthode de sur-échantillonnage des données. Elle est fréquemment utilisée dans le cadre de données déséquilibrées. Cette méthode va principalement être introduite dans le cas de classification binaire. En effet, cela permettra de simplifier la compréhension. Des données sont considérées comme étant déséquilibrées dès lors que les deux classes de la variable cible n'ont pas la même proportion. Dans la pratique, on parle de données déséquilibrées lorsque la classe minoritaire est inférieure à 10% de la variable cible. Ce constat est très fréquent dans les problématiques liées à l'assurance et peut parfois nuire aux prédictions des modèles utilisés.

C'est pourquoi SMOTE est très régulièrement employé dans ces problématiques. SMOTE relève d'une stratégie d'*oversampling*. L'idée de cet algorithme est d'augmenter le nombre d'individus présents dans la classe minoritaire. Les nouveaux individus générés sont créés à partir de ceux existants tout en étant légèrement modifiés. Ils ne sont donc pas une simple copie des individus déjà présents dans la base. La base de données est alors rééquilibrée et les prédictions des modèles de *machine learning* sont théoriquement améliorées du fait de l'ajout de ces individus. La classe minoritaire représente de ce fait une part plus importante des données.

Pour ce faire, l'algorithme procède selon les étapes suivantes :

1. Un individu de la classe minoritaire est sélectionné de manière aléatoire
2. Une droite est tracée entre ce point et l'un de ces k plus proches voisins de façon aléatoire

3. La longueur de cette droite est multipliée par un coefficient aléatoire compris entre 0 et 1
4. Un nouveau point est construit à cet endroit

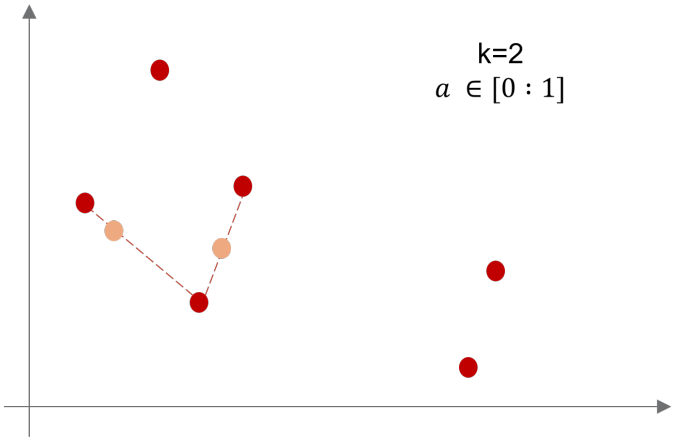


FIGURE 5.7 – Création de nouveaux individus avec SMOTE [2]

L'exemple 5.7 permet d'illustrer ce mécanisme [2]. Finalement, les étapes sont répétées jusqu'à atteindre le nombre d'individus supplémentaires de la classe minoritaire escompté.

Intégration de SMOTE dans les données : Dans le cadre de ce mémoire, le package *imbalanced learn* a été sélectionné afin d'utiliser la fonction SMOTENC. SMOTENC signifie *Synthetic Minority Over-sampling Technique for Nominal and Continuous*. En effet, cette fonction a permis d'utiliser les méthodes d'*oversampling* via SMOTE tout en permettant de gérer les variables catégorielles et nominales. Cette fonction procède exactement comme la méthode SMOTE classique à l'exception qu'il faut spécifier les variables catégorielles dans les paramètres.

Dans ce mémoire, c'est cette approche qui a été conservée. Pour autant, il existe bien d'autres méthodes d'*oversampling* des données. On peut retrouver dans le package *imbalanced learn* de Python les fonctions suivantes :

— SMOTE : cette méthode a été détaillée précédemment.
— SMOTEN : est une méthode de data augmentation basée sur SMOTE utilisée exclusivement dans le cas où toutes les variables du jeu de données sont catégorielles.
— SVMSMOTE : est une variante des méthodes de data augmentation qui est basée sur les SVM. Brièvement, la différence principale est que les individus sont

2. Source : https ://kobia.fr/imbalanced-data-smote/

rééchantillonnés par rapport à des individus sélectionnés par un SVM.

— BorderlineSMOTE : est une méthode de rééchantillonnage basé sur SMOTE. Toutefois, les k plus proches voisins sont déterminés grâce au tracé d'une frontière (et non plus d'une droite) entre les individus de la classe majoritaire et ceux de la classe minoritaire. Concrètement, cette idée est résumée dans le graphe 5.8 :

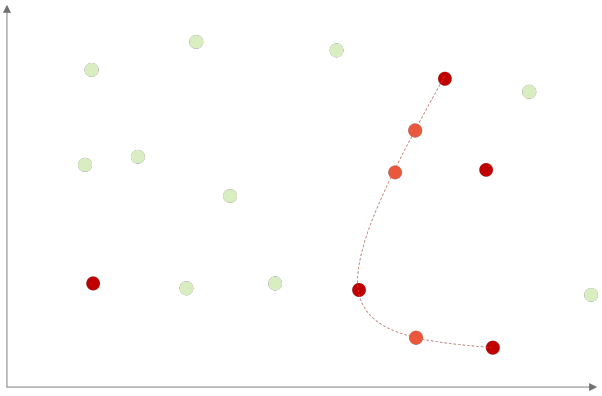


FIGURE 5.8 – Création de nouveaux individus avec Borderline SMOTE

— ADASYN : fonctionne comme SMOTE à la différence que les k plus proches voisins sont établis en fonction d'un cercle autour des individus tirés aléatoirement. A nouveau, le fonctionnement est détaillé dans la figure 5.9 :

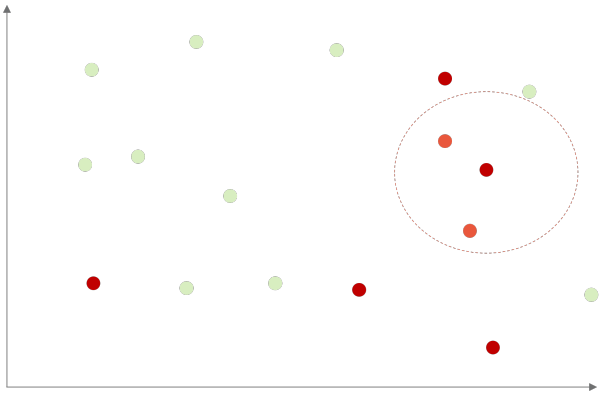


FIGURE 5.9 – Création de nouveaux individus avec ADASYN

— KMeansSMOTE : est une méthode d'*oversampling* basée sur les *clusterings*. Elle regroupe les données dans des *clusters* et sélectionne les groupes avec beaucoup d'individus minoritaires. Ainsi, plus d'individus sont attribués dans ce *cluster*.

⇒ **L'*undersampling***

Dans le cadre de ce mémoire, une méthode d'*undersampling* a été également utilisée. Cette méthode vise à réduire le nombre d'individus de la classe majoritaire afin de donner davantage d'importance aux données de la classe minoritaire. Les individus de la classe majoritaire sont ainsi retirés aléatoirement.

Pour cette approche, la fonction *RandomUnderSampler* du package *imbalanced learn* a été utilisée. En effet, de nombreuses informations de la classe majoritaire étaient disponibles et pouvaient potentiellement nuire aux prédictions de la classe minoritaire.

Pour conclure, les méthodes de rééchantillonnage ont été utilisées afin d'essayer d'améliorer les performances de nos modèles en donnant plus d'importance aux assurés réalisant des versements libres.

5.3 Présentation de la méthode SHAP d'analyse des variables importantes

SHAP [13] est une méthode permettant de mieux comprendre les sorties de nos modèles de *machine learning*. En effet, les modèles les plus performants ne sont pas nécessairement les plus simples à décrypter. Un réel enjeu s'articule autour de cette problématique d'interprétabilité des modèles. Cet enjeu s'inscrit encore plus dans le coeur de ce mémoire car nous cherchons à connaître les variables les plus importantes de notre modèle. L'avantage de SHAP est que cette méthode peut être utilisée pour tous les modèles de machine learning.

Plusieurs analyses sont réalisées dans le cadre de la méthode SHAP :

- Une première approche générale : nous informe sur les variables importantes dans la globalité du modèle.
- Une seconde approche plus locale : permet d'interpréter les variables importantes sur un ou quelques individus.

Comment sont déterminées les variables importantes ? Cette méthode est basée sur les valeurs de Shapley Additive Explanation (Shap). Concrètement, cette valeur est calculée pour chaque combinaison de variables. Ces combinaisons sont ensuite synthétisées grâce à une moyenne.

Cette valeur est utilisée en théorie des jeux. Elle se présente dans le cas de jeux coopératifs où l'objectif est de répartir équitablement les gains d'un jeu aux différents participants. L'idée est de comprendre les interactions entre les joueurs afin de maximiser le gain.

Comme évoqué précédemment, ces valeurs sont calculées pour toutes les variables.

Les valeurs de Shapley sont calculées comme suit :

$$Shapley(V_1) = \sum_{E \subset N,\ i \in E} \frac{(k - |E|)!(|E| - 1)!}{k!} \times [V_1(E) - V_1(E \backslash \{k\})] \tag{5.1}$$

où k est un individu de l'ensemble E et E est une partie de la totalité des individus notée N.

Plus généralement, les valeurs de Shap sont utilisées dans de nombreuses représentations graphiques afin d'analyser les impacts de chaque variable. Les graphiques les plus fréquemment utilisés dans le cadre de ce mémoire sont présentés dans la figure 5.10 [3].

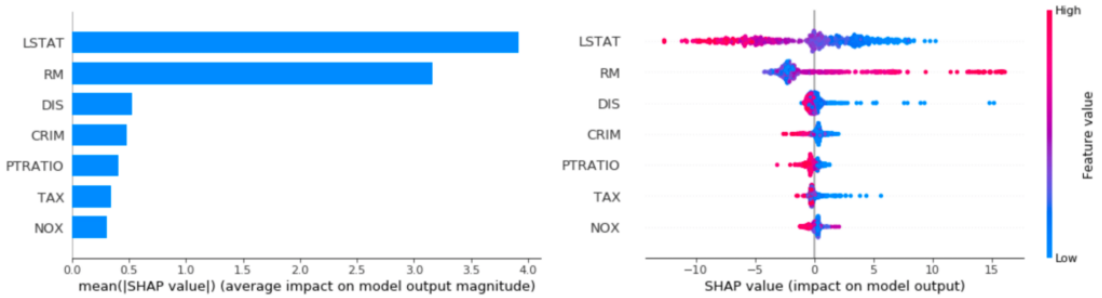


FIGURE 5.10 – Exemples de visuels utilisés pour représenter l'importance des variables d'un modèle [3]

A gauche de la figure 5.10, ce sont les valeurs absolues de Shap qui sont moyennées. Les variables les plus importantes sont logiquement les variables avec le coefficient le plus élevé. Le second graphique permet de visualiser toutes les valeurs de Shap calculées. Les valeurs de SHAP sont visibles sur l'axe des abscisses. Des valeurs de SHAP élevées sont ainsi visibles sur la droite et les basses sur la gauche. Le degré de couleurs du bleu au rouge correspond aux niveaux de la variable étudiée. Ces éléments permettent ainsi d'analyser le sens dans lequel la variable influence la prédiction.

Les analyses Shap permettent également d'analyser les dépendances entre les variables explicatives du modèle comme en témoigne la figure 5.11 [4].

3. Source : `https://www.aquiladata.fr/insights/shap-mieux-comprendre-linterpretation-de-modeles/#:~:text=Gr%C3%A2ce%20%C3%A0%20la%20valeur%20de,l'article%20%5B3%5D`

4. Source : `https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html`

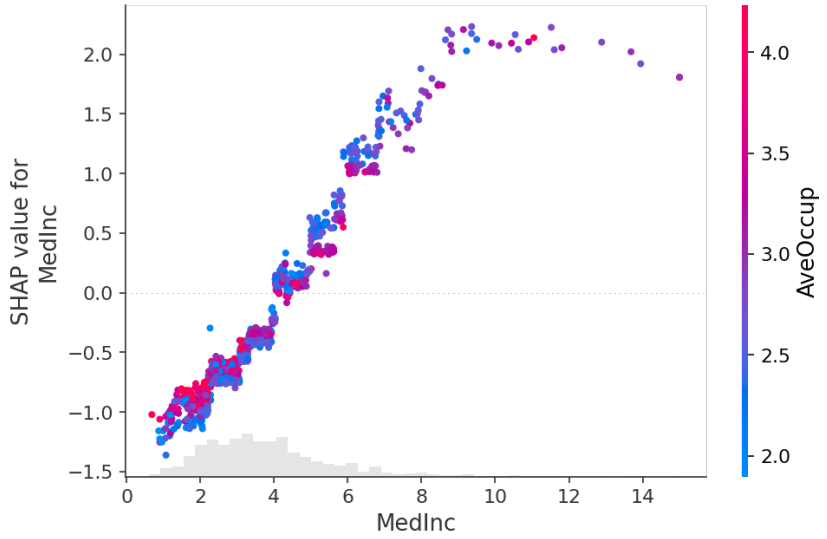


FIGURE 5.11 – Exemple de visuel utilisé pour représenter la dépendance entre deux variables [4]

Dans le cadre de ce mémoire, c'est le package SHAP sous Python qui a été utilisé. Il a ainsi permis de réaliser des graphes similaires à ceux présentés dans cette introduction à la méthode.

Y a-t-il des limites à l'utilisation de cette méthode ? Cette méthode nécessite un temps important de calcul. C'est pour cette raison que l'on travaille sur un échantillon représentatif. Néanmoins, cet inconvénient a été contré par la création de *TreeExplainer*. En effet, cet algorithme a permis de réduire la complexité des arbres de décision utilisés.

Chapitre 6

Analyse des résultats de modélisation

Cette partie va présenter les résultats des modèles testés pour appréhender les comportements de versements libres. Deux approches de modélisation ont été utilisées comme cela a été évoqué dans la partie précédente. Pour rappel, la première est une classification binaire de la survenance de versements libres. La seconde est un modèle actuariel classique de fréquence généralement utilisé dans les tarifications. L'idée est de modéliser le nombre de versements libres par assuré, année et support. L'exposition sera utilisée en variable offset.
Les différents modèles explicités dans la partie précédente ont été testés pour les deux approches. Cette partie vise donc à présenter le procédé de modélisation mis en place ainsi que les résultats obtenus et retenus. Une fois le modèle final construit, un SMOTE sera appliqué aux données afin d'essayer d'améliorer les prédictions. Finalement, un travail d'explicabilité du modèle final sélectionné a été réalisé à l'aide de l'algorithme SHAP.

6.1 Construction d'un modèle

6.1.1 Étapes générales de modélisation

Tout d'abord, la base de données a été séparée entre une base d'apprentissage et une base de test. La base de test représente 20% de la base totale. Le *dataset* test a été considéré de manière aléatoire car l'objet de cette étude porte davantage sur la compréhension des versements libres. Une autre approche avait été envisagée en considérant les derniers mois de notre base de données comme *dataset* de test mais cela se rapproche plus d'une problématique de prédiction et non de compréhension.
La suite de cette partie va s'attacher à détailler le déroulé des modèles mis en place.

Les modèles présentés dans la partie précédente ont été implémentés à l'aide de

Python et plus particulièrement du package *scikit-learn*. L'utilisation d'un pipeline a permis de réaliser à la fois :

— L'encodage des variables catégorielles ordinales et nominales dans une étape de *preprocessing* ;
— La mise en place du modèle à tester dans une étape classification ou régression suivant l'approche envisagée.

Les différents modèles ont suivi les mêmes étapes citées ci-dessous :

1. Les modèles ont été testés avec les paramètres par défaut avec une pondération par l'exposition.
2. Les performances ont été analysées par validation croisée sur la base d'apprentissage.
3. Les hyperparamètres des modèles ont été optimisés à nouveau à partir d'une validation croisée. Ainsi, plusieurs hyperparamètres ont été testés afin d'améliorer les prédictions globales.

Le tableau ci-dessous récapitule les premiers paramètres optimisés pour les différents modèles :

GLM Ridge	Arbre CART et Forêt Aléatoire	Histogram Gradient Boosting
Le paramètre de réglage λ	*max_features* : le nombre maximal de variables considérées dans un noeud	*loss* : choix de la fonction de perte
	max_depth : la profondeur de l'arbre	*max_depth* : la profondeur maximale de chaque arbre
	min_samples_split : le nombre minimal d'individus requis pour construire un noeud interne	*max_bins* : le nombre maximal de compartiments dans la discrétisation des variables
	min_samples_leaf : le nombre minimal d'individus requis pour construire un noeud terminal	*min_samples_leaf* : le nombre minimal d'individus requis pour construire un noeud terminal
	ccp_alpha : le paramètre de complexité	

Ces hyperparamètres ont été testés à l'aide de la fonction *GridSearchCV()* du package *scikit-learn* dans Python. Cette fonction permet de tester une sélection de paramètres établis par validation croisée. L'erreur moyenne et l'écart-type des différents *folds* sont déterminés pour chaque combinaison de paramètres sur les données de validation. Finalement, la fonction nous informe sur les hyperparamètres donnant les meilleures prédictions pour notre modèle.

Une fois les modèles optimisés obtenus, il a été question de s'assurer que les modèles ne soient pas en sur-apprentissage. Dans le cas du modèle d'*histogram gradient boosting*, une option appelée *early_ stopping* permet déjà de parer cette problématique. Elle est paramétrée par défaut dans les modèles. Cette méthode d'optimisation des modèles vise à contrer les phénomènes de sur-apprentissage à partir d'algorithmes itératifs telle que la descente de gradient. En quelques mots, elle détermine si l'ajout d'arbres dans le modèle permet une amélioration significative des performances. La figure 5.1 montre une augmentation de l'erreur sur les données de validation à partir d'un certain seuil de complexité de l'algorithme apprenant toujours mieux sur les données d'entraînement. Cela revient à la définition du surapprentissage. La méthode d'*early stopping* détecte cette phase et arrête l'entraînement à partir de ce moment. Ainsi, l'erreur globale est minimisée et le modèle ne sur-apprend pas des données d'apprentissage.

Néanmoins, cette fonctionnalité n'est pas disponible pour les forêts aléatoires. Une fois les paramètres généraux définis, la profondeur du modèle a été questionnée par validation croisée. Les paramètres cités précédemment ont été testés avec une profondeur d'arbres relativement faible, c'est-à-dire entre 30 et 40 arbres, pour éviter un phénomène de sur-apprentissage. Finalement, une seconde validation croisée a été mise en place en fixant les paramètres optimaux et en faisant varier la profondeur des arbres de décision. Seul le paramètre *n_ estimators* a ainsi été challengé dans cette seconde partie. Des graphes ont permis de constater l'évolution de l'erreur en fonction du nombre d'arbres utilisés.

Une fois les hyperparamètres des modèles optimisés, les modèles ont été évalués sur la base de test selon plusieurs indicateurs qui seront détaillés dans la prochaine section 6.1.2. Tous les modèles ne seront pas détaillés de la même manière afin de simplifier la lecture.

6.1.2 Identification des métriques

Afin d'analyser les différents modèles, il est nécessaire de sélectionner les indicateurs adaptés à notre problématique. Cette partie va donc introduire les métriques retenues dans nos analyses.

Dans les modèles de classification, l'indicateur qui a été privilégié est le F1 score car il permet d'avoir une vision globale des performances du modèle. En effet, le F1 score est fréquemment utilisé dans le cadre de données déséquilibrées. Cet indicateur représente une synthèse des taux de rappel et de précision. Chaque optimisation des paramètres a donc été faite en fonction de ce critère.

Les calculs des taux de rappel, précision et du F1 score découlent de la matrice de confusion présentée dans la figure 6.1.

		Prédiction	
		VL=0	VL=1
Réel	VL=0	Vrai négatif : VN	Faux positif : FP
	VL=1	Faux négatif : FN	Vrai positif : VP

FIGURE 6.1 – Schéma d'une matrice de confusion

Cette matrice va donc générer plusieurs calculs de taux d'erreur :

$$\text{Taux de rappel } = \frac{VP}{VP+FN} \text{ et Taux de précision } = \frac{VP}{VP+FP}$$

$$\text{F1-score} = \frac{2}{\frac{1}{\text{Taux de précision}} + \frac{1}{\text{Taux de rappel}}} = \frac{VP}{VP + \frac{1}{2}(FN+FP)}$$

Dans le cadre de données déséquilibrées, il faut être vigilant aux métriques utilisées. L'*accuracy* fait partie de ces métriques. En effet, cette métrique donne une erreur globale du modèle. Pour autant, si nous considérons les données de ce mémoire, nous avons 96,3% de classe 0 et 3,7% de classe 1. Un modèle trivial prédisant systématiquement 0 aurait un *accuracy* de 96.3% mais la problématique ne sera pas résolue. Cette métrique ne sera pas conservée dans les analyses.

Dans les modèles de régression, l'indicateur qui a été privilégié est le RMSE pour *Root Mean Squared Error* et le MAE pour *Mean Absolute Error*. Ces indicateurs permettent de calculer les écarts entre les données observées et les donnés prédites. Ils sont donnés par les formules suivantes :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(\hat{y}_k - y_k)^2} \quad \text{et} \quad \text{MAE} = \frac{1}{n}\sum_{k=1}^{n}|\hat{y}_k - y_k|$$

Ces deux métriques se complètent. Le RMSE se distingue par le fait qu'il soit plus pénalisé sur les valeurs extrêmes.

6.1.3 Comparaison des performances des modèles avec les paramètres par défaut

Tous les modèles ont été testés avec les paramètres par défaut afin de comparer leurs performances. Dans cette partie, les performances ont été jugées à partir de validations croisées [1]. Le meilleur modèle sera ensuite conservé dans les analyses qui suivront. Cette

1. 10 folds ont été utilisés

section va donc s'articuler autour de deux parties présentant ainsi les deux approches de modélisation retenues.

L'approche de classification binaire de la survenance des versements libres :
Les différents modèles testés ont permis de fournir les résultats suivants par validation croisée :

	GLM Pénalisé	Arbre CART	Forêt Aléatoire	Histogram GB
Taux de précision	0,961	0,398	0,943	0,870
Taux de rappel	0,288	0,447	0,310	0,328
F1 score	0,443	0,437	0,466	0,476
Accuracy	0,973	0,954	0,974	0,973
AUC	0,837	0,711	0,864	0,874

Ces premiers résultats nous ont permis de privilégier la méthode d'Histogram Gradient Boosting car l'objectif était de maximiser le F1 score. De plus, c'est l'algorithme qui a obtenu l'AUC le plus élevé. C'est donc ce modèle qui a été retenu pour les prochaines étapes.

L'approche de modélisation de la fréquence de versements libres :
Les différents modèles testés ont permis de fournir les résultats suivants :

	GLM Pénalisé	Arbre CART	Forêt Aléatoire	Histogram GB
RMSE	0,296	0,408	0,282	0,299
MAE	0,085	0,082	0,08	0,078

Pour cette approche, les taux d'erreur semblent très proches. Pour rappel, l'objectif était de minimiser le RMSE ou le MAE. Le RMSE est plus élevé que le MAE car la méthode de calcul est plus pénalisante sur les valeurs extrêmes. En considérant le RMSE, c'est la forêt aléatoire qui est retenue. Mais si l'on privilégie le MAE, c'est l'Histogram Gradient Boosting qui affiche le meilleur score.

Pour la suite des travaux, c'est l'Histogram Gradient Boosting qui a été retenu car il affiche des performances très satisfaisantes. Cela sera également l'occasion de comparer les résultats des deux approches pour le même algorithme. De plus, l'Histogram Gradient Boosting présente des temps de compilation bien plus optimisés que la forêt aléatoire. En effet, cet algorithme a été conçu pour accélérer les temps de traitements des méthodes de boosting. Pour l'algorithme des forêts aléatoires, un choix a été fait de préparer les données en encodant en *one hot encoding*, ce qui n'a pas été nécessaire pour l'histogram gradient boosting. Cette distinction explique également la rapidité de l'histogram gradient boosting par rapport à la forêt aléatoire.

Modèle sélectionné pour les prochaines étapes :
Pour la suite des travaux, c'est la méthode d'Histogram Gradient Boosting pour les deux approches (classification et fréquence) qui a été retenue.

6.1.4 Optimisation des hyperparamètres des modèles retenus

Les hyperparamètres des modèles d'Histogram Gradient Boosting vont être optimisés afin d'améliorer les performances des deux approches. Cette partie a été réalisée à partir de la fonction *GridSearchCV* de Python.

✓ Approche de classification de la survenance de versements libres

La stratégie d'évaluation du *Grid Search* s'est orientée sur le F1 score pour ce modèle. Ainsi, plusieurs hyperparamètres ont été testés et mis en concurrence par validation croisée.

Le tableau ci-dessous permet de visualiser les hyperparamètres par défaut ainsi que ceux obtenus après optimisation.

Nom des paramètres	Paramètres par défaut	Paramètres après *Fine-Tuning*
max_ depth	**None**	5
max_ bins	225	225
min_ samples_ leaf	20	250

Ces paramètres ont permis d'améliorer les performances globales du modèle. Les résultats présentés ci-dessous ont ainsi été obtenus :

Taux de précision	0,84
Taux de rappel	0,35
F1 score	0,493
Accuracy	0,97
AUC	0,87

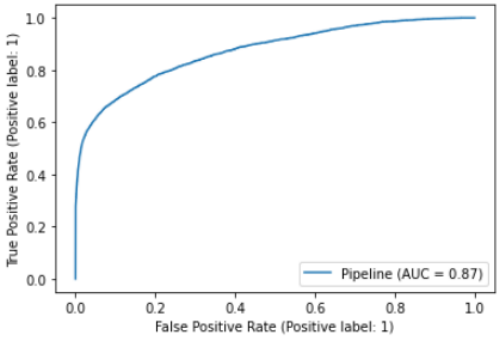


FIGURE 6.2 – Courbe ROC du modèle optimisé

Ce modèle optimisé a donc été conservé pour la réalisation des analyses SHAP.

✓ Approche de modélisation de la fréquence de versements libres

Pour cette approche, la stratégie d'évaluation du *Grid Search* s'est orientée sur le RMSE. Ainsi, plusieurs hyperparamètres ont été testés et mis en concurrence par validation croisée.

Le tableau ci-dessous permet de visualiser les hyperparamètres par défaut ainsi que ceux obtenus après optimisation.

Nom des paramètres	Paramètres par défaut	Paramètres après *Fine-Tuning*
max_ depth	**None**	20
max_ bins	225	225
min_ samples_ leaf	20	250

Ces paramètres ont permis d'améliorer les performances globales du modèle. Les résultats présentés ci-dessous ont ainsi été obtenus :

RMSE	0,28
MAE	0,075

Ce modèle optimisé a donc également été conservé pour la réalisation des analyses SHAP.

6.2 Analyse des variables influentes

Comme cela a été présenté dans la partie 5, des analyses SHAP ont été mises en place sur les différents modèles afin d'identifier les variables influençant la décision d'investissements par versements libres. Cette section va donc s'attacher à présenter les différents visuels produits à cet effet.

Tout d'abord, ce sont les résultats de la **méthode de classification** qui vont être présentés. Deux types de graphes, 6.3 et 6.4, ont été produits afin d'identifier les variables importantes.

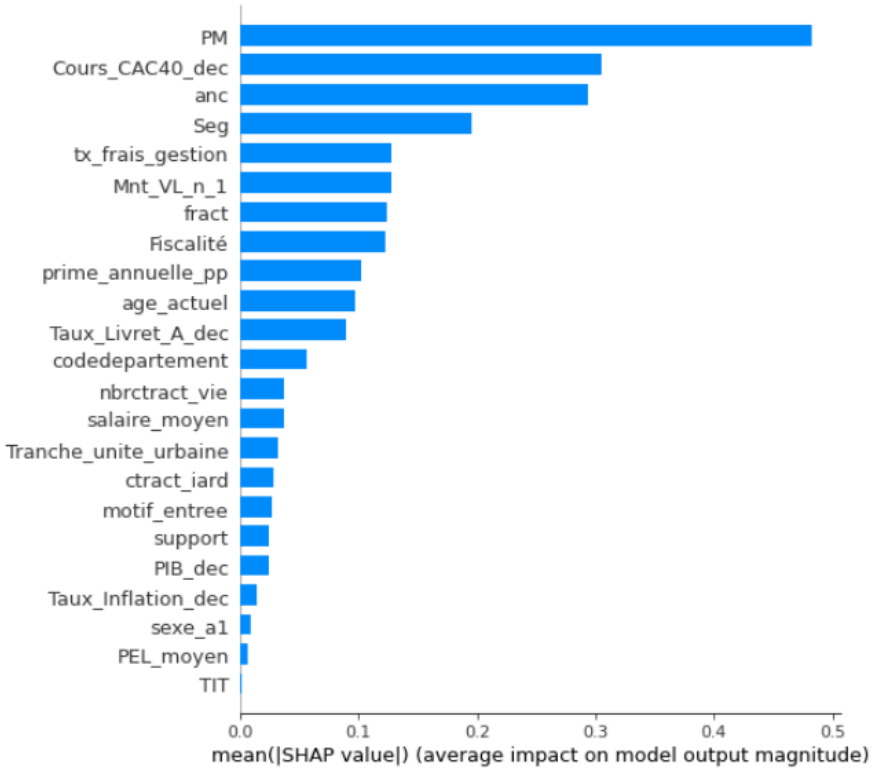


FIGURE 6.3 – Importance globale des variables avec une représentation des valeurs de Shap

Ce premier graphe nous permet de constater les variables influençant la décision d'un versement libre. Ainsi, on constate que quatre variables semblent se détacher du fait de valeurs moyennes de Shap plus élevées. On retrouve ainsi les variables PM, Cours du CAC 40, ancienneté et Segmentation.
Ensuite, quatre nouvelles variables ont également des valeurs moyennes de Shap significatives et proches. Parmi ces variables, nous retrouvons les taux de frais de gestion, le montant de versement libre de l'année précédente, le fractionnement des versements

périodiques et la fiscalité du produit.
Le prochain visuel 6.4 nous permettra de connaître le sens de l'influence de la décision de versements.

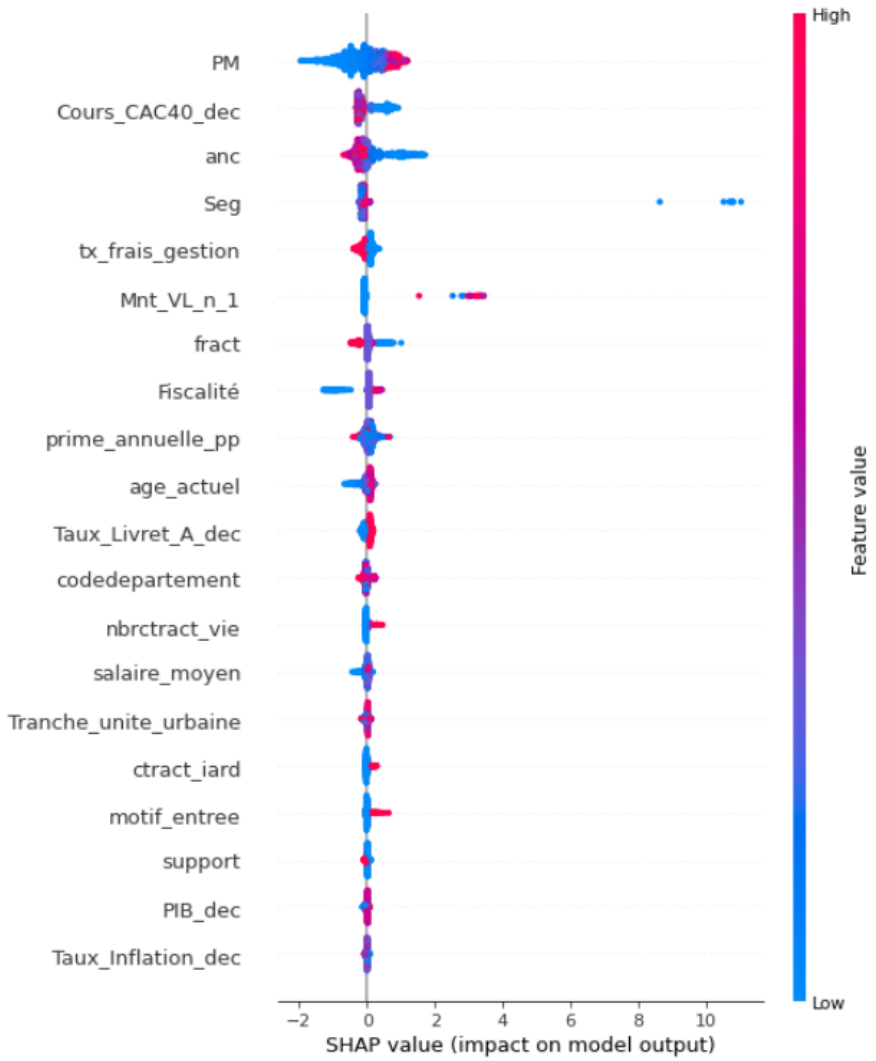


FIGURE 6.4 – Synthèse de l'importance globale des variables

On constate des relations plus importantes pour les variables PM, Cours du CAC 40 et ancienneté. Chacune des variables parmi les plus significatives va être détaillée et justifiée dans les prochains paragraphes.

⇒ **La provision mathématique**

La provision mathématique joue un rôle important dans la prédiction des versements libres. On pouvait en effet supposer que la provision mathématique influencerait nos modèles de fréquence car sous Prophet, les montants sont modélisés en fonction de cet indicateur.
De plus, la figure 6.4 nous permet d'ajouter que lorsque les assurés ont des provisions mathématiques importantes, ils sont plus sujet à effectuer des versements libres. Cela paraît une nouvelle fois cohérent car un assuré ayant peu capitalisé sera certainement moins enclin à faire des versements libres.

L'analyse de cette variable a été poursuivie à l'aide des graphes de dépendance proposés par SHAP. Ce graphe nous permet d'analyser l'impact d'une variable en particulier.

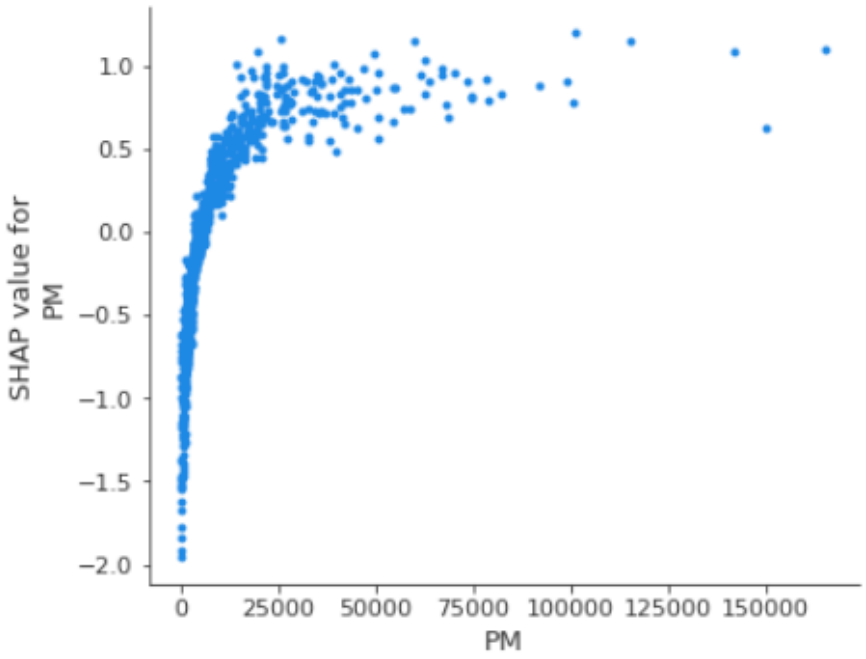


FIGURE 6.5 – Impact de la variable provision mathématique sur les prédictions

Ainsi, la figure 6.5 permet de confirmer que les contrats avec des montants de provisions mathématiques plus élevés sont ceux qui réalisent le plus de versements libres. Ce résultat paraît tout à fait cohérent car les assurés qui ont toujours beaucoup investi sur leur contrat continuent voire augmentent leurs investissements à l'approche de la retraite.

⇒ **Le cours du CAC 40**

La variable correspondant au cours du CAC 40 en fin d'année semble également jouer un rôle important dans la prédiction des versements libres. Ce résultat est très intéressant car cela signifie que notre modèle a identifié les assurés réalisant des versements libres conjoncturels. Cela signifie que les assurés ont investi en connaissance d'un contexte économique plus ou moins favorable aux investissements.

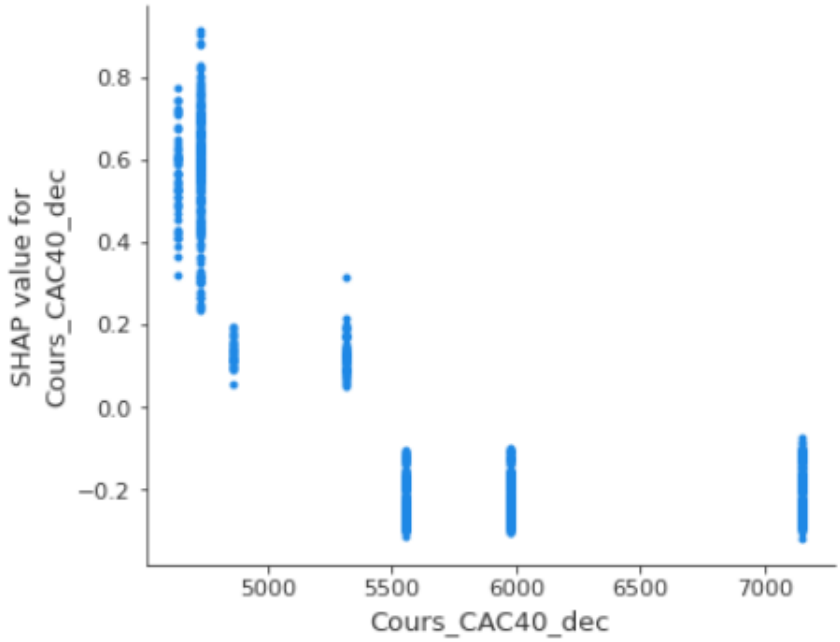


FIGURE 6.6 – Impact de la variable Cours_CAC40_dec sur les prédictions

Tout d'abord, on constate que plus le cours du CAC 40 est faible plus les assurés réalisent des versements libres. Ainsi, lorsque les cours sont bas, les assurés cherchent d'autres types de placements plus rémunérateurs. Ces assurés se sont donc orientés sur leur produit retraite. Une étude a permis d'identifier que les produits sur lesquels ces assurés ont investi sont principalement Anthologie, La Retraite 94, 08 et 13. Ce résultat est très intéressant car les cours du CAC 40 sont liés à l'année. Or, pour ces années en question, les taux garantis étaient positifs. Cette connaissance des produits permet donc de prouver le fait que les assurés ont réellement cherché des taux d'investissements plus favorables. Un autre effet potentiel peut venir du fait qu'un cours bas du CAC 40 peut induire de potentielles hausses futures des marchés. Les assurés ont également pu privilégier des investissements sur des supports UC qui pourraient se traduire par une plus-value futures.

⇒ **L'ancienneté**

Pour rappel, l'étude préliminaire de la partie 4 nous avait permis de constater une décroissance des taux de versements libres par ancienneté. Ce constat est confirmé par nos modèles. Nous pouvions déjà l'apercevoir sur la figure 6.4. Les anciennetés les plus faibles sont celles qui influencent le plus la réalisation de versements libres.
Le *dependence plot* 6.7 le démontre en effet.

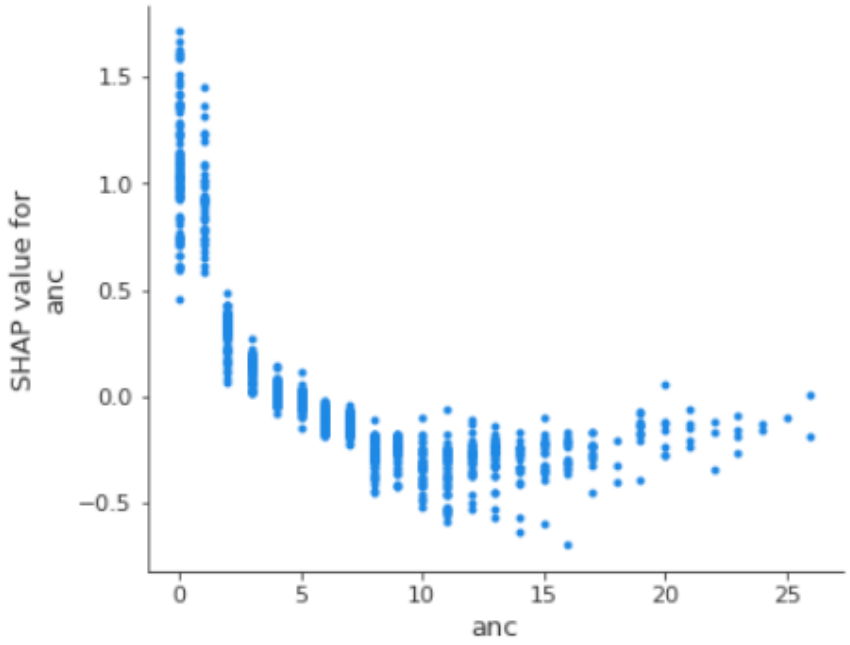


FIGURE 6.7 – Impact de la variable ancienneté sur les prédictions

La figure 6.7 nous indique également une décroissance de la variable ancienneté. Plus concrètement, les assurés ayant souscrits récemment sont plus amenés à effectuer des versements libres. Cela peut venir du fait qu'ils viennent de souscrire et souhaitent donc être actifs sur leur contrat. Les contrats les plus anciens sont donc davantage délaissés.

On constate également une légère hausse des valeurs de Shap pour les anciennetés dépassant 20 années. Comme Shap étudie l'impact des variables en décorrélant des autres variables, cela ne provient pas d'un effet produit. Ce phénomène peut venir du fait que les assurés soient proches de la retraite et réalisent davantage de versements libres.

⇒ **La segmentation des produits**

Les types de produits semblent également influencer la réalisation de versements libres de manière significative. Cette variable était pressentie lors de l'analyse descriptive des données. Nous avions identifié cette variable comme étant potentiellement discriminante. Cette hypothèse se confirme ici.

La figure 6.8 nous permet de voir que le produit Anthologie est clairement discriminant dans la modélisation. Comme cela avait été supposé dans la partie 3, le produit Anthologie connaît beaucoup d'investissements par versements libres.

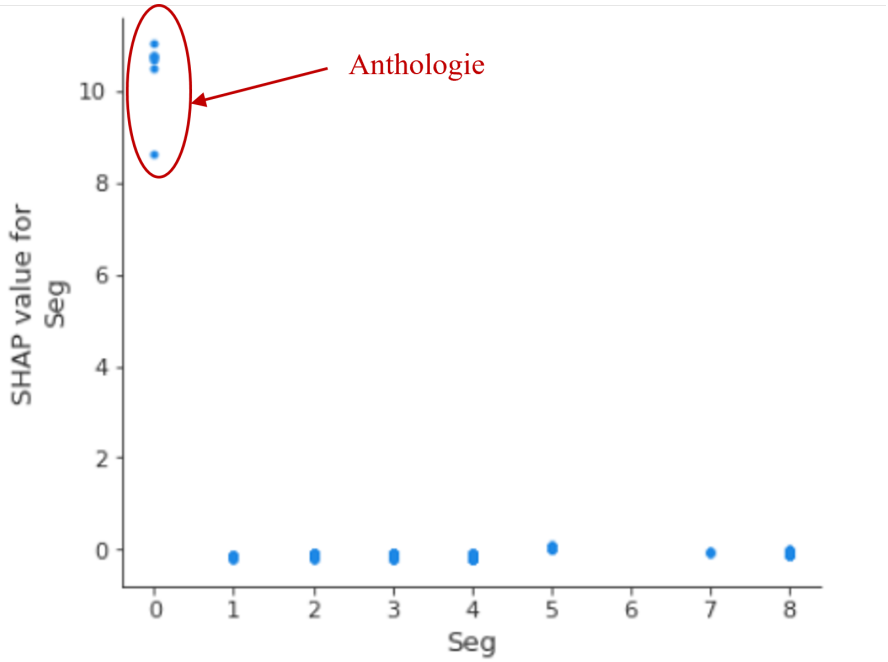


FIGURE 6.8 – Impact de la variable produit sur les prédictions

Nous l'avions déjà abordé dans la partie 4. L'annonce de la fin de commercialisation d'Anthologie a généré beaucoup de versements libres afin d'anticiper les transferts vers les PERP engendrés par la loi PACTE. Il est donc cohérent d'identifier ce produit dans les variables discriminantes.

Ainsi, le modèle a réussi à capter cet effet exceptionnel et temporaire lié à Anthologie. Cela n'est pas un problème dans notre étude car l'objectif est de comprendre le phénomène des versements libres. Néanmoins, cela pourrait être plus problématique dans le cadre d'un modèle de prédictions. Dans ce cas, le périmètre des PERP aurait dû être retraiter afin d'éviter d'apprendre sur cet évènement exceptionnel.

Ces quatre premiers paragraphes ont développé les variables qui se détachent significativement en terme d'impact sur la réalisation de versements libres. Néanmoins, un autre groupe apparaît également dans la figure 6.3 autour des variables :

— taux de frais de gestion
— montant de versements libres l'année précédente
— fractionnement des primes périodiques
— fiscalité
— montant annuel de prime périodique
— âge actuel de l'assuré
— taux de livret A

Certaines de ces variables vont être détaillées lors des prochains paragraphes.

⇒ Le taux de frais de gestion

La figure 6.3 nous montrait que lorsque le taux de frais de gestion était faible, les assurés effectuaient davantage de versements libres. Cette idée est davantage détaillée dans la figure 6.9 ci-dessous.

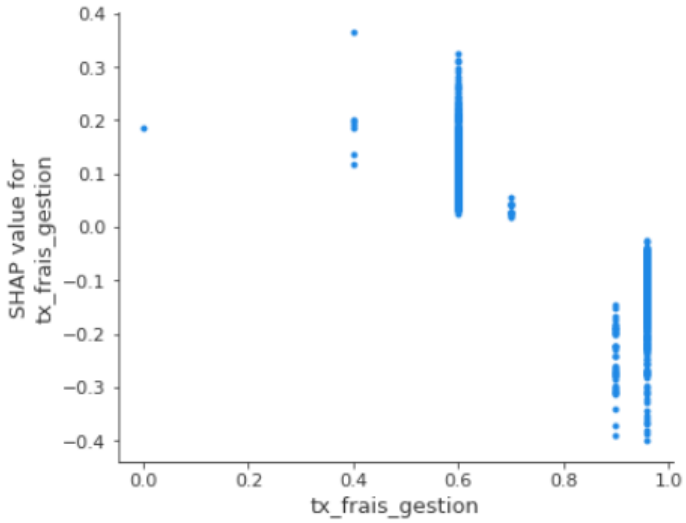


FIGURE 6.9 – Impact de la variable taux de frais de gestion sur les prédictions

La variable taux de frais de gestion présente peu de valeurs. Les principaux taux sont compris entre 0,6% et 1%. On constate néanmoins que plus le taux est faible plus les assurés réalisent de versements libres.

⇒ **Le montant de versement libre l'année précédente**

La variable montant de versement libre en N-1 que nous avions créée semble également jouer un rôle dans la prédiction. L'hypothèse que nous avions suggérée visant à dire que les assurés réalisant des versements le font chaque année apparaît ici validée.

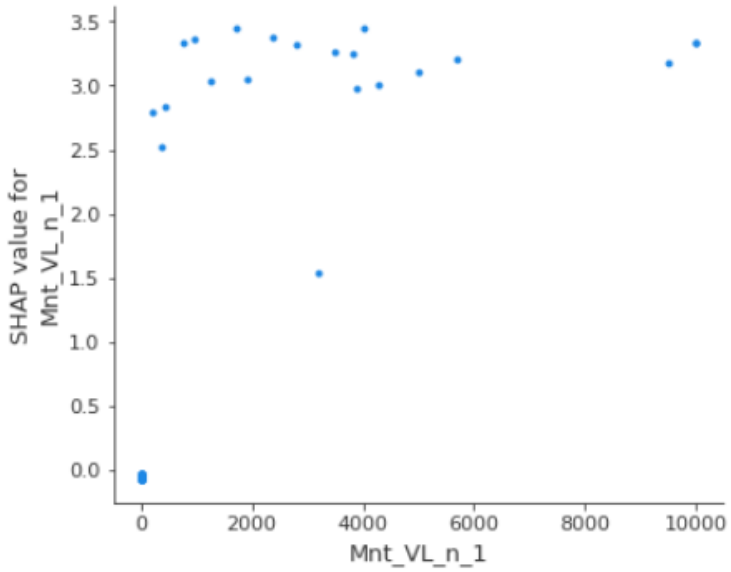


FIGURE 6.10 – Impact de la variable montant de versement libre l'année précédente sur les prédictions

La figure 6.10 nous confirme que les valeurs de Shap sont élevées lorsqu'un versement libre a eu lieu l'année précédente. On constate beaucoup de valeurs en 0 mais cela est tout à fait normal car 98% des valeurs de cette variable sont des 0.

⇒ **Le fiscalité des produits**

La fiscalité joue un rôle également dans les prédictions. Cette variable était également pressentie. Comme cela avait été évoqué précédemment, certaines fiscalités des produits encouragent les assurés à effectuer des versements libres.

Tout d'abord, les quatre groupes de valeurs de Shap de la figure 6.11 représentent l'assurance vie, les Madelin, le PER individuel et les PERP de gauche à droite. Ainsi, on constate que ce sont essentiellement les PER individuels et les PERP qui connaissent davantage de versements libres. Cette figure confirme donc les premiers constats réalisés pour la variable produit.

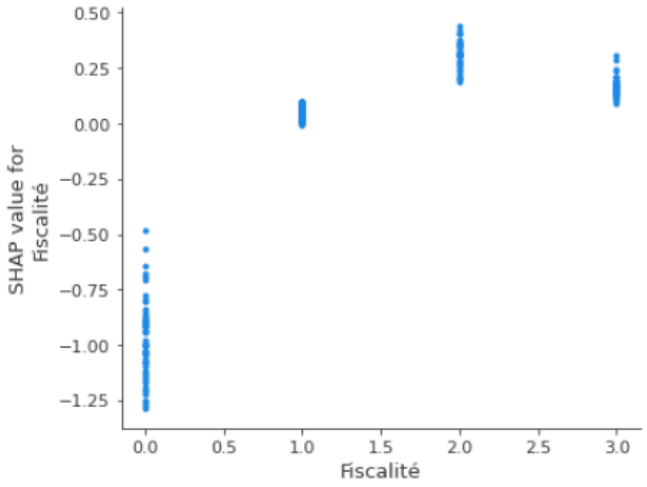


FIGURE 6.11 – Impact de la variable fiscalité sur les prédictions

⇒ **L'âge de l'assuré au moment du versement**

L'âge de l'assuré au moment du versement libre impacte également la prédiction, comme en témoigne la figure 6.12.

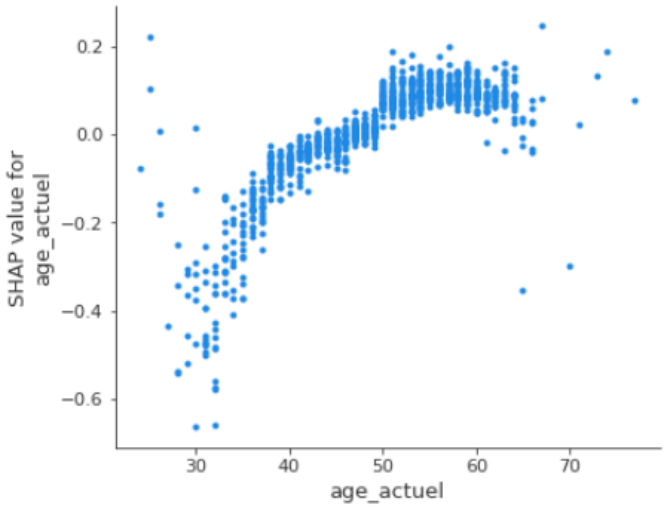


FIGURE 6.12 – Impact de la variable âge actuel sur les prédictions

On constate clairement que plus l'assuré est âgé plus il réalise de versements libres. En effet, il est naturel de penser que les assurés ont plus de revenus et moins de charges avec l'âge et donc préparent plus activement leur retraite.

Ensuite, ce sont les résultats de la **méthode de modélisation de la fréquence des versements libres** qui vont être présentés. Les mêmes analyses ont été effectuées. Pour plus de fluidité, seules les divergences avec les précédents résultats ont été détaillées.

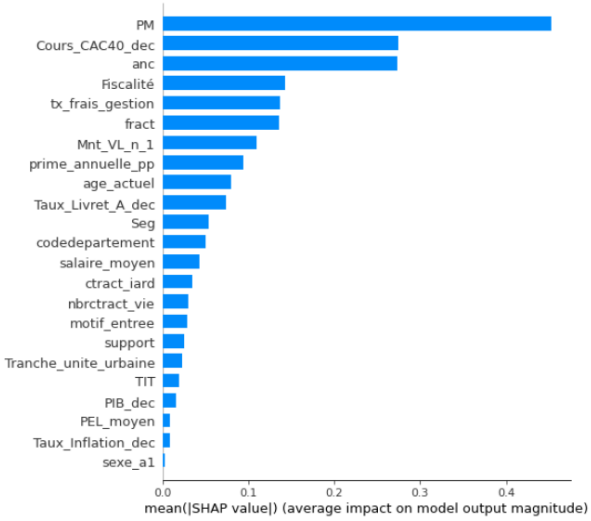


FIGURE 6.13 – Importances moyennes des variables du modèle

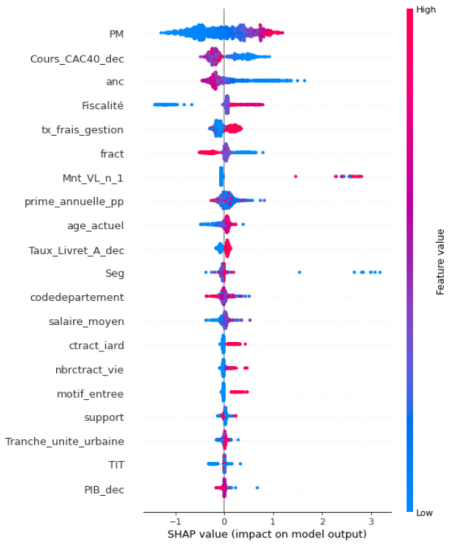


FIGURE 6.14 – Impact des variables sur les prédictions

Les trois variables les plus significatives dans ce modèle, comme en témoigne la figure 6.13, sont les mêmes que pour le modèle de classification. Les graphiques d'analyse de ces trois variables ont été mis dans la première partie de l'annexe à la page 106. Les constats sont les mêmes que pour la partie sur la classification, page 88.

Le second point remarquable est que les variables importantes qui suivent sont les mêmes qu'analysées précédemment à partir de la page 92. On retrouve ainsi les variables :

- Fiscalité
- Taux de frais de gestion
- Fractionnement des primes périodiques
- Montant de versement libre en N-1
- Montant annuel de primes périodiques
- Âge actuel
- Taux Livret A
- Segmentation

Les variables ne ressortent pas dans le même ordre mais on constate que globalement ce sont toujours les mêmes variables qui ressortent de nos modèles.

Notons que les démarches présentées pour l'Histogram Gradient Boosting ont été réalisées pour tous les modèles. Ainsi, nous avons pu constater que les variables importantes sur les modèles de GLM pénalisé, d'arbre de décision et de forêt aléatoire étaient globalement les mêmes. Les différentes informations à ce sujet ont été résumées dans la seconde section de l'annexe à la page 108.

Pour conclure sur cette partie, les modèles réalisés ont permis d'obtenir des résultats satisfaisants tout en permettant des analyses intéressantes quant aux impacts des différentes variables. Ainsi, certaines variables pressenties lors de la première approche simplifiée ont été retrouvées et certaines ont été découvertes. Ces dernières ont pu être justifiées par des analyses complémentaires. Ces différents points nous ont ainsi permis de mieux comprendre les versements libres.

6.3 Impact de la data augmentation

Comme cela a déjà été évoqué, les données sont déséquilibrées. Elles ont donc été rééquilibrées afin de voir si cela pouvait améliorer les performances de nos modèles.
Pour cette partie, seul le modèle d'Histogram Gradient Boosting pour l'approche de classification a été retenu. En effet, ce modèle a été jugé comme étant le plus simple en termes d'interprétations.

Des méthodes d'*oversampling* et d'*undersampling* ont été associées dans cette perspective.
La classe majoritaire étant conséquente, sa réduction a permis d'augmenter la proportion de la classe minoritaire dans les données. Seules les données de la base d'apprentissage ont été rééchantillonnées. Ainsi, cette première opération a permis de passer la classe minoritaire de 3,7% à 9,1% des données.
Les données ont par la suite été finalisées grâce à la méthode SMOTE, évoquée dans la section 5.2, page 73. De nouveaux individus ont été créés afin d'augmenter le nombre d'individus de la classe minoritaire. Cette dernière étape a fait passer la classe minoritaire de 9,1% à 16,7% des données.

Une fois les données rééchantillonnées, l'Histogram Gradient Boosting a été mis en place suivant les mêmes étapes que pour les autres modèles. Le modèle avec les paramètres par défaut a été entraîné sur les données rééchantillonnées. Les hyperparamètres ont ensuite été optimisés puis le modèle a été testé sur le *dataset* de test.

Les performances suivantes ont ainsi été obtenues :

	Performances initiales	Performances après rééchantillonnage
Taux de précision	0,84	0,57
Taux de rappel	0,35	0,49
F1 score	0,493	0,527
AUC	0,87	0,86

Le rééchantillonnage des données a permis d'améliorer les performances du modèle. Le F1 score a ainsi atteint 0,527 contre 0,493 précédemment. Le taux de rappel a été largement amélioré et ce au détriment du taux de précision. L'AUC est quant à lui resté presque inchangé.

Par la suite, les analyses SHAP ont également été réalisées afin de comparer les sorties des modèles.

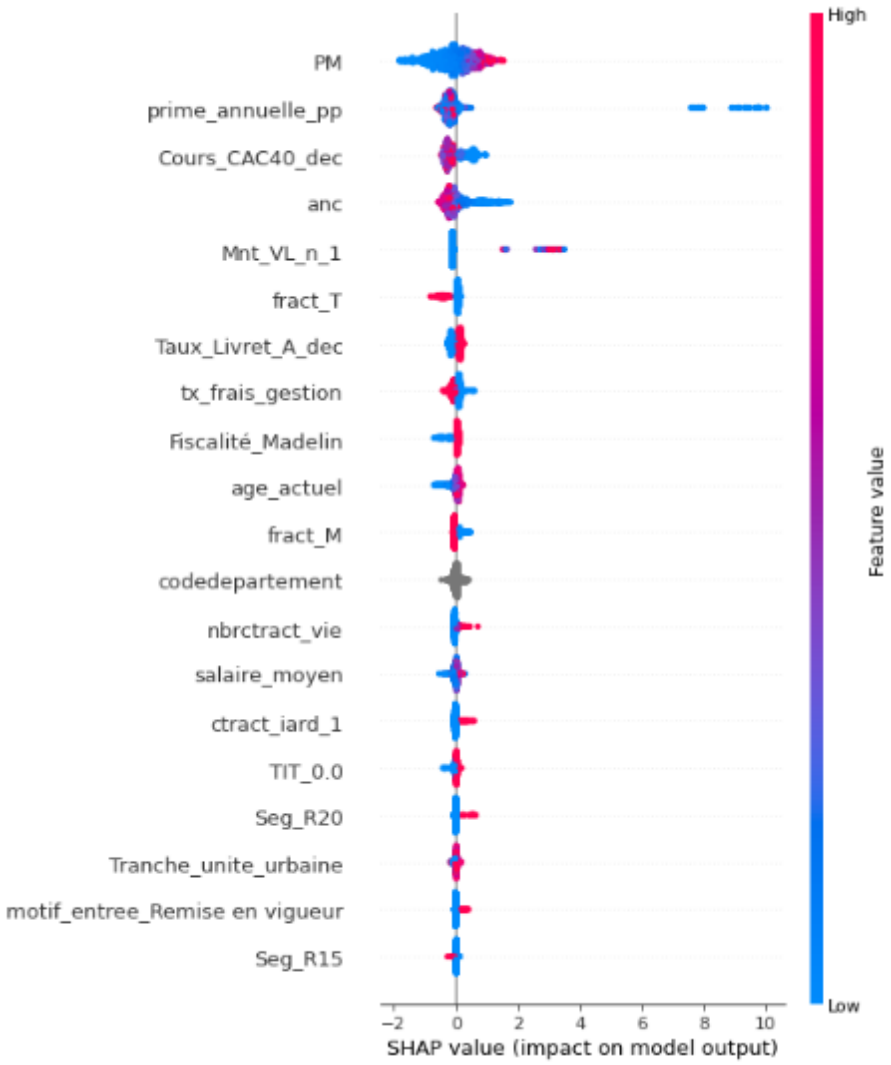


FIGURE 6.15 – Impact des variables sur les prédictions

Les variables ressorties de cette dernière approche ne font que confirmer la prédominance de certaines variables dans la prédiction des versements libres. Ainsi, on constate que les variables PM, Cours du CAC 40 et ancienneté sont à nouveau parmi les plus

importantes. La variable montant annuel de primes périodiques prend néanmoins davantage d'importance que pour les visuels précédents.

En effet, pour la majorité des modèles, des variables sont clairement ressorties. Parmi les 24 variables explicatives, neuf d'entre elles semblent se détacher :

— La provision mathématique
— Le cours du CAC 40
— L'ancienneté
— La segmentation et la fiscalité des produits
— Le fractionnement ainsi que les montants des primes périodiques
— Le montant de versement libre en N-1
— Le taux de frais de gestion

6.4 Préconisations pour la gestion des versements dans l'outil Prophet

L'outil Prophet gère actuellement la modélisation des versements libres en utilisant le montant de provision mathématique comme base de calcul. Ensuite, les taux de versements libres sont gérés dans des tables d'hypothèses fonctionnant par ancienneté et produit. De ce fait, les variables PM, ancienneté, segmentation des produits et fiscalité sont déjà considérées dans les modèles internes.

Nos analyses ont démontré que ces variables étaient clairement nécessaires à la prédiction des versements libres. En effet, elles ont toujours figuré dans les variables importantes des différents modèles considérés. Ce premier constat est rassurant quant au fonctionnement de l'outil Prophet.

Par ailleurs, d'autres variables ont également permis d'en apprendre davantage sur les versements libres. Ces différentes variables vont être détaillées. Nous allons tenter d'observer des différences de dynamiques dans les taux de versements libres afin de constater les impacts de ces variables.

Ainsi, le cours du CAC 40 est apparu comme une variable discriminante dans la modélisation. L'apport de données financières pourrait donc être une approche intéressante afin de capter d'autres types de comportements des assurés. Les versements libres conjoncturels seraient donc analysables.

De plus, les primes périodiques sont discriminantes dans la modélisation des versements libres. Que ce soit pour le montant annuel ou le fractionnement des versements, ces variables influencent la décision de réaliser un versement libre.
Des calculs ont été faits en ce sens en comparant les niveaux de taux de versements libres entre les versements périodiques annuels ou non. Ainsi, deux graphes ont été construits

pour le produit La Retraite 08 en fonction de ces deux hypothèses. Ce produit a été retenu car il dispose d'un nombre important de contrats et propose des investissements à la fois en euro et en UC. La distinction entre les périodicités annuelles ou non a été fixée grâce aux analyses des valeurs de Shap et à la partie de statistiques descriptives.

Ce constat paraît cohérent. Les assurés réalisant moins de versements périodiques (en nombre) peuvent être amenés à alimenter leur contrat par d'autres moyens. Les versements libres peuvent ainsi être envisagés. Ces résultats ont été confirmés par les graphes 6.16 et 6.17.

Les taux de versements libres sont supérieurs dans le cas où les versements périodiques sont annuels.

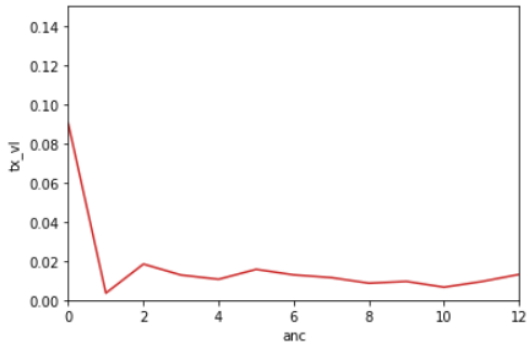


FIGURE 6.16 – Taux de versements libres pour les versements de périodicité mensuelle, trimestrielle ou semestrielle

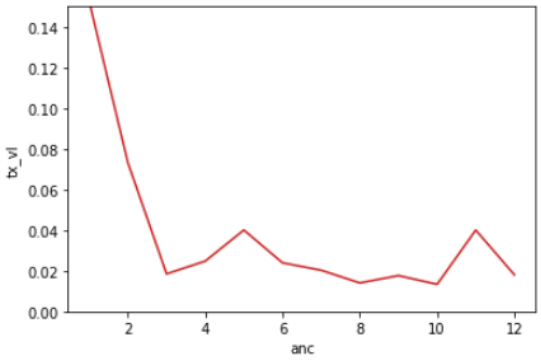


FIGURE 6.17 – Taux de versements libres pour les versements de périodicité annuelle

D'autre part, des taux de frais de gestion bas encouragent les versements libres.
A nouveau, deux graphes ont été construits pour le produit La Retraite 08 avec des taux de frais de gestion à plus et moins de 0,7%. Ce seuil a été fixé grâce aux analyses des valeurs de Shap de la figure 6.9. On constatait que les valeurs de Shap devenaient négatives en dépassant ce seuil de 0,7%. Les graphes 6.18 et 6.19 ont ainsi été obtenus.

Les niveaux de versements libres sont nettement supérieurs dans le cas où les taux de frais de gestion sont plus faibles.

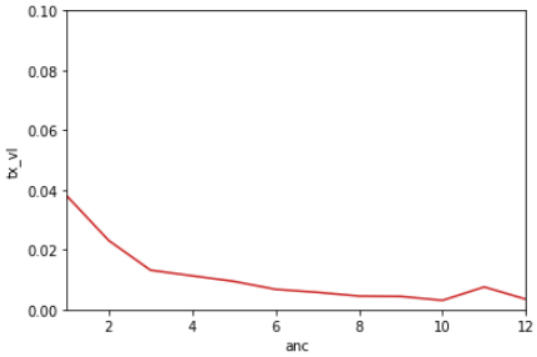


FIGURE 6.18 – Taux de versements libres pour les taux de frais de gestion de plus de 0,7%

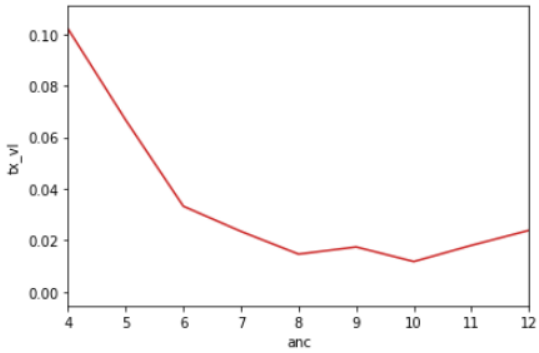


FIGURE 6.19 – Taux de versements libres pour les taux de frais de gestion de moins de 0,7%

Le dernier point concerne le fait que les assurés réalisant des versements libres soient généralement coutumier de le faire. Cette hypothèse a également été vérifiée par le biais des mêmes méthodes que précédemment pour le produit La Retraite 08. Les figures 6.20 et 6.21 ont ainsi été obtenues.

A nouveau, on constate que les taux de versements sont plus élevés dans le cas où un versement libre a été effectué l'année précédente. On confirme donc que les assurés réalisant des versements libres le font régulièrement.

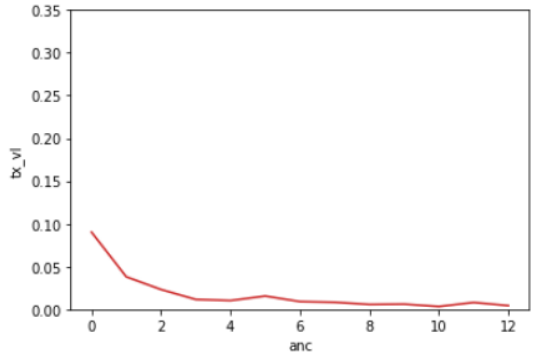


FIGURE 6.20 – Taux de versements libres pour les cas où un versement libre N-1 n'a pas eu lieu

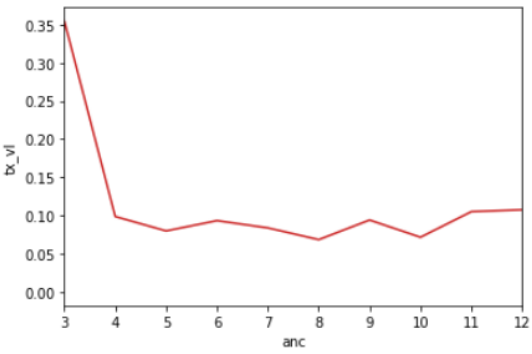


FIGURE 6.21 – Taux de versements libres pour les cas où un versement libre N-1 a eu lieu

Finalement, ces constats nous permettent d'affirmer que ces variables ont un impact sur les versements libres. IFRS 17 est une norme qui nécessite encore aujourd'hui des ajustements par les assureurs. Ce mémoire va donc permettre à Generali de mieux appréhender les versements libres. Certaines variables déjà utilisées dans les outils ont été identifiées par ces modélisations. La détection de nouvelles variables pourra ainsi permettre de proposer d'autres axes de modélisation des versements libres à l'avenir.

Conclusion

La norme IFRS 17 induit de nombreux changements dans la comptabilisation du bilan. Le souhait d'homogénéiser les pratiques des compagnies d'assurance pour plus de transparence impacte les éléments modélisés en actuariat. Ainsi, l'intégration des versements libres dans la frontière des contrats représente un réel enjeu. Ces versements sont aléatoires et complexes à appréhender. Pour ces raisons, actuellement les versements libres sont rarement considérés dans le bilan comptable des assureurs. Ce mémoire a donc cherché à comprendre et interpréter ce phénomène.

Cette compréhension est passée par plusieurs étapes importantes. Dans un premier temps, le contexte réglementaire et le périmètre de ce mémoire ont dû être assimilés. En effet, des éléments externes influencent les versements libres. Ainsi, la compréhension de la norme IFRS 17, de la loi PACTE et des produits retraite a été essentielle au bon déroulé de ce mémoire. Nous l'avons vu, la loi PACTE a impacté les habitudes de versements des assurés sur les PERP notamment en générant des niveaux exceptionnels de versements libres. Les produits ont également un rôle important car certains encouragent davantage la réalisation de versements libres.

Une fois l'environnement compris, une phase de modélisation a été développée. Deux approches ont été considérées afin de comprendre au mieux le phénomène des versements libres.

Tout d'abord, une approche simplifiée a permis d'appréhender à la fois les données de ce périmètre et la construction de taux de versements libres par Generali. Cette première approche a également été l'occasion de réaliser des travaux de modifications des codes dans l'outil Prophet. Cette partie, bien que complexe, a été très enrichissante. Il a été plus qu'essentiel de fiabiliser les modifications réalisées avant de pouvoir réaliser quelconque analyse. La partie visant à valider ces codes a ainsi été primordiale. Les lois construites et renseignées dans cet outil ont permis d'obtenir de très bonnes prédictions au global sur les montants annuels de versements libres. Nous l'avons vu ce bon résultat cachait néanmoins quelques compensations entre certains produits qui étaient soit surestimés soit sous-estimés. Pour autant, cette première méthode a permis de modéliser les versements libres de manière raisonnable. Ainsi, les impacts des indicateurs tels que le BE ou la PVFP quantifiés par la partie ALS ont été jugés cohérents par rapport aux

attentes sur ces produits.

Finalement, cette méthode a permis d'établir certaines limites quand à la modélisation des versements libres uniquement en fonction de l'ancienneté et des produits retraite. Ces limites ont motivé la partie faisant appel aux méthodes de *machine learning*. L'objectif était donc de chercher d'autres variables influençant la réalisation de versements libres par les assurés.

L'approche de modélisation par des méthodes de *machine learning* a donc eu pour objectif de challenger les hypothèses des modèles internes. A cet effet, deux modèles ont été envisagés :

— Un modèle de classification binaire visant à prédire la survenance de versements libres
— Un modèle de fréquence des versements libres

Pour ces deux modèles, les mêmes algorithmes ont été implémentés avec un GLM pénalisé, un arbre de décision, une forêt aléatoire et un histogram gradient boosting. Ces modèles ont été comparés et c'est finalement l'histogram gradient boosting qui nous a permis de prédire au mieux les versements libres.

Finalement, l'idée de ces modélisations était principalement de pouvoir extraire les variables les plus influentes afin de comprendre le fonctionnement des versements libres. Il a donc fallu interpréter les sorties des modèles. Cela n'est pas toujours simple dans le cadre de modèle complexe de *machine learning*. La méthode SHAP a ainsi permis de le faire. En effet, cette méthode offre la possibilité de connaître les variables les plus importantes ainsi que d'interpréter l'impact d'une variable isolée sur la prédiction. Des graphes ont rendu facilement interprétables les sorties et ont permis de dresser plusieurs analyses au sujet des versements libres.

Ainsi, la provision mathématique, le cours du CAC 40 et l'ancienneté sont les trois variables ressorties des modèles qui sont les plus importantes dans la prédiction des versements libres. Tout d'abord, le fait d'avoir obtenu la provision mathématique et l'ancienneté a permis de recouper avec l'architecture des outils internes. De plus, de nouvelles variables ont également permis d'en apprendre davantage. En effet, le cours du CAC 40 était une variable plus inattendue dans les résultats. Cela nous a permis de constater que les versements libres conjoncturels étaient pris en compte dans nos modèles. Cette variable offre ainsi des perspectives intéressantes d'hypothèses supplémentaires dans les outils internes.
D'autres variables sont également régulièrement apparues dans les analyses SHAP comme les produits, les taux de frais de gestion, la réalisation de versements libres en N-1, la fiscalité des produits, les primes périodiques ainsi que l'âge de l'assuré. Finalement, huit variables sont apparues comme nécessaires à la compréhension des versements libres. Chacune de ces variables ont été décryptées et ont permis de réaliser des analyses intéressantes.

Par la suite, des méthodes de *data augmentation* ont été mises en oeuvre en raison du caractère déséquilibré de la cible à prédire afin d'améliorer les performances des modèles. Des méthodes d'*oversampling* et d'*undersampling* ont ainsi permis d'améliorer les prédictions du modèle de survenance des versements libres. Les résultats de ce modèle ont confirmé les conclusions établies précédemment. Les trois variables principales sont en effet ressorties. La prime périodique a cependant pris davantage d'importance dans les prédictions. Après analyses, nous avons pu constater qu'effectivement les contrats avec des niveaux de primes périodiques bas étaient plus enclin à faire des versements libres.

Ce mémoire a donc permis de confirmer l'emploi de certaines variables et d'en proposer de nouvelles afin de comprendre le phénomène de versements libres. Bien qu'il paraisse complexe d'insuffler des modifications dans les outils internes, cette étude permet une compréhension plus large des versements libres.

Pour conclure, ce mémoire permet de motiver d'autres pistes d'études concernant les versements libres comme la modélisation des montants de ces derniers dans le cas où un versement libre a été effectué. Il pourrait également être intéressant de construire un outil simplifié de projection des versements libres afin d'analyser les impacts des différentes variables identifiées dans les prédictions des primes brutes et des indicateurs clés du bilan comptable. Une dernière piste serait également d'intégrer de nouvelles variables dans l'outil interne de Generali.

Listes des acronymes utilisés

Acronymes	Signification
AUC	Area Under the Curve
BE	Best Estimate
CAC	Cotation Assistée en Continu
CART	Classification And Regression Trees
CSM	Contractual Service Margin
EFRAG	European Financial Reporting Advisory Group
FRA	Full Retrospective Approach
FVA	Fair Value Approach
GLM	Generalized Linear Model
IARD	Incendie, Accidents et Risques Divers
IAS	International Accounting Standard
IASB	International Accounting Standards Board
IASC	International Accounting Standards Committee
IFRS	International Financial Reporting Standard
LAT	Liability Adequacy Test
MAE	Mean Absolute Error
MRA	Modified Retrospective Approach
PACTE	Loi relative à la croissance et la transformation des entreprises
PB	Participation aux Bénéfices
PEL	Plan Epargne Logement
PER	Plan Epargne Retraite
PERCO	Plan d'Epargne pour la Retraite COllectif
PERP	Produit Epargne Retraite Populaire
PIB	Produit Intérieur Brut
PM	Provision Mathématique
PP	Prime Périodique
PPNA	Provision pour Primes Non Acquises
PSAP	Provision pour Sinistre A Payer
PVFP	Present Value of Future Profits
RA	Risk Adjustment
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations
SII	Solvabilité 2
TA	Technique Assurance
TIT	Taux d'Intérêt Technique
TMG	Taux Minimum Garanti
UC	Unité de Compte
VL	Versements Libres

Annexe

Annexe 1 : Analyse des impacts des variables importantes de l'Histogram Gradient Boosting pour le modèle de fréquence

Les trois *dependence plot* (figures 6.22, 6.23 et 6.24) pour le modèle de fréquence des versements libres sont présentés ci-dessous :

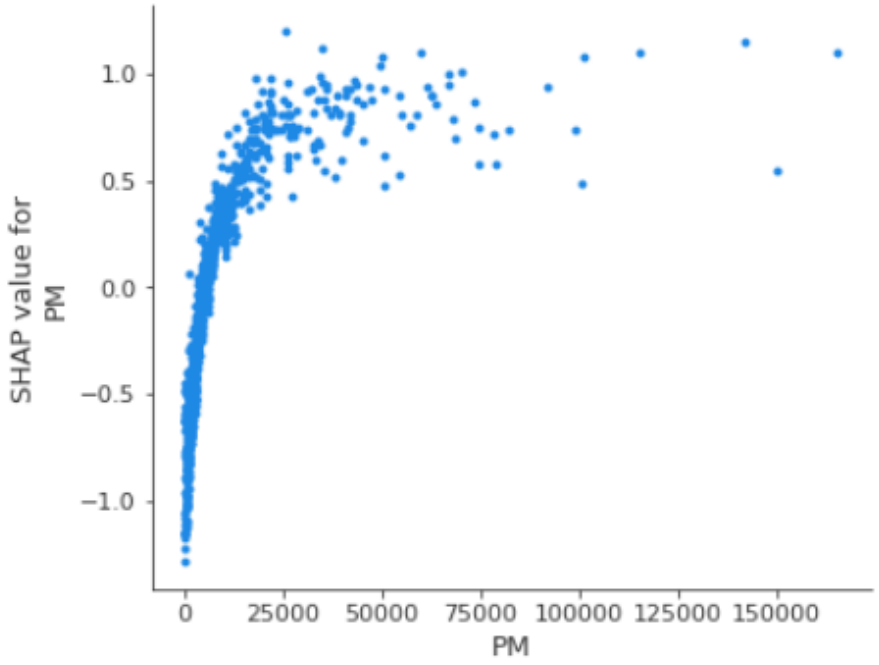


FIGURE 6.22 – Impact de la variable PM dans le modèle de fréquence

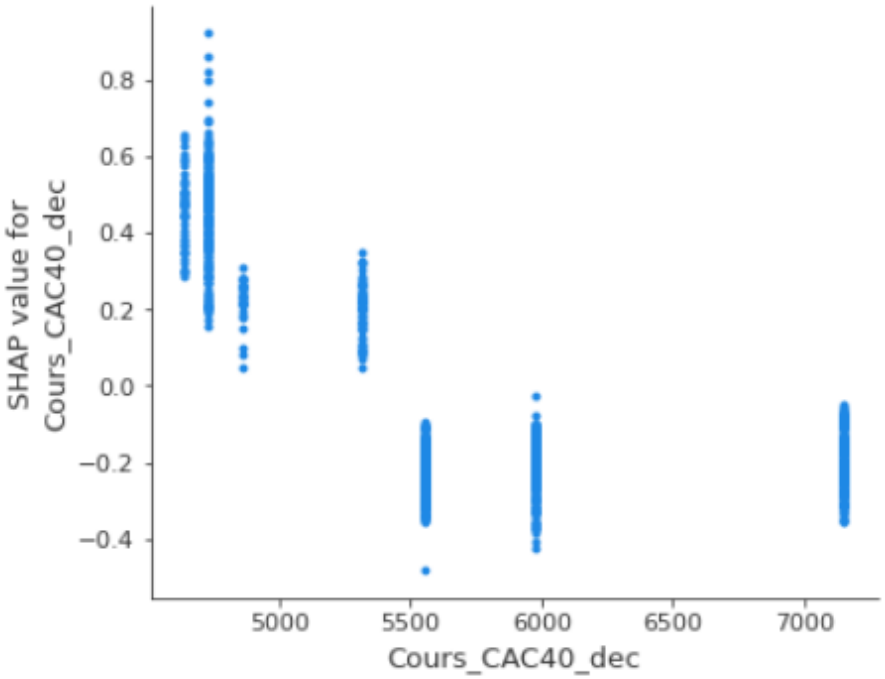


FIGURE 6.23 – Impact de la variable CAC 40 dans le modèle de fréquence

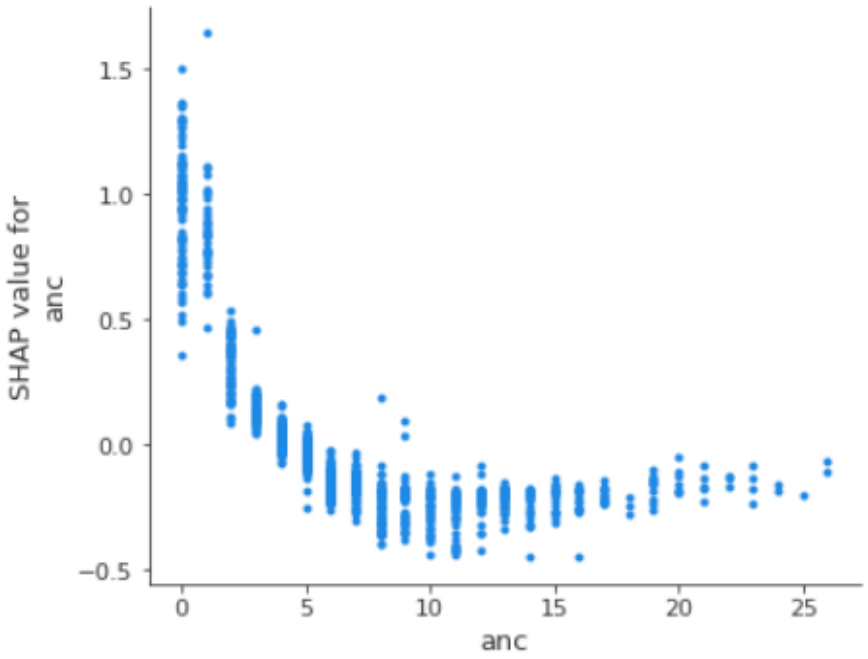


FIGURE 6.24 – Impact de la variable ancienneté dans le modèle de fréquence

Annexe 2 : Résultats des autres modèles pour la classification et le modèle de fréquence

Les résultats des autres modèles après optimisation des hyperparamètres sont présentés ci-dessous :

— Première approche : classification binaire de la survenance de versements libres :

	GLM Pénalisé	Arbre de décision	Forêt Aléatoire
Taux de précision	0,95	0,94	0,97
Taux de rappel	0,28	0,28	0,29
F1 score	0,437	0,434	0,442
AUC	0,84	0,79	0,88
Variables importantes	Segmentation, Fiscalité, Taux d'inflation, Taux de livret A, taux de PEL moyen ...	Montant annuel de primes périodiques, Montant de versement libre en N-1, Ancienneté, Âge actuel, Fiscalité ...	Fiscalité Segmentation, Support, Fractionnement, Montant versement libre en N-1 ...

— Seconde approche : modèle de fréquence de versements libres :

	GLM Pénalisé	Arbre de décision	Forêt Aléatoire
RMSE	0,28	0,39	0,27
MAE	0,083	0,076	0,075
Variables importantes	Taux d'inflation, Taux Livret A, Segmentation, Fiscalité, taux de frais de gestion	Segmentation, PM, Montant de versement libre en N-1, Cours CAC 40, Ancienneté, Montant annuel de primes périodiques	Segmentation, PM, Montant de versement libre en N-1, Cours CAC 40, Ancienneté, Montant annuel de primes périodiques

Bibliographie

[1] Sarah ANDRE : Comportement de versement libre en épargne individuelle : approche conceptuelle et modélisation. Mémoire de D.E.A., ISFA, 2019.

[2] Karine ASSARAF : Modélisation de versements libres sous ifrs 17 par des méthodes de machine learning. Mémoire de D.E.A., Dauphine, 2020.

[3] Fatima Zohra BENABDELKRIM : Modélisation des versements libres en assurance-vie : utilisation de méthodes de scoring. Mémoire de D.E.A., ISUP, 2017.

[4] L. Breiman : Random forests. machine learning 45, 5–32. 2001.

[5] L. Breiman, J. Friedman, R. Olshen et C. Stone : *CART : Classification and Regression Trees, Wadsworth International.* 1984.

[6] N. V. Chawla, K. W. Bowyer et W. P. Kegelmeyer : "smote : synthetic minority over-sampling technique," journal of artificial intelligence research, 321-357. 2002.

[7] Hastie, Tibshirani et Friedman : The elements of statistical learning (2nd edition). 2009.

[8] Arthur E Hoerl et Robert W Kennard : "ridge regression : Biased estimation for nonorthogonal problems." technometrics 12 (1). taylor francis group : 55–67. 1970.

[9] James, Witten, Hastie et Tibshirani : An introduction to statistical learning : with applications in r (2nd edition). 2021.

[10] Sory Ibrahima KABA : Ifrs 17 : Intégration des versements libres dans la frontière des contrats. Mémoire de D.E.A., ISFA, 2020.

[11] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye et Tie-Yan Liu : Lightgbm : A highly efficient gradient boosting decision tree. 2017.

[12] V. KERHAIGNON : *Cours EURIA : Comptabilité des assurances vie et non vie.* 2021-2022.

[13] S.M. Lundberg, G. Erion et H. Chen : From local explanations to global understanding with explainable ai for trees. nat mach intell 2, 56–67. 2020.

[14] Tristan MUSCAT : Étude du comportement client dans le cadre des versements exceptionnels en assurance vie. Mémoire de D.E.A., Dauphine, 2016.

[15] J. A. Nelder et R. W. M. Wedderburn : Generalized linear models. journal of the royal statistical society. series a (general), 135(3), 370–384. 1972.

[16] F. Planchet et A Miseray : Tarification iard : Introduction aux techniques avancées. Mars 2017.

[17] scikit-learn developers (BSD License) : Histogram-based gradient boosting classification tree. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html, 2007 - 2022.

[18] SERVAJEAN, LEVEAU et CHAILAN : Les arbres de régression et de classification. https://www.univ-montp3.fr/miap/ens/miashs/master/ues/test_maximilien/htmls/1_what_is_ml/2_regression_and_classification_trees.html, 2021.

[19] I. SOW : *Cours EURIA : IFRS 17 - Contrats d'assurance.* 2021-2022.

[20] Robert Tibshirani : "regression shrinkage and selection via the lasso." journal of the royal statistical society. series b (methodological). jstor, 267–88. 1996.

[21] F. VERMET : *Cours EURIA - Machine learning.* 2020-2022.