

Mémoire présenté devant l'ENSAE Paris  
pour l'obtention du diplôme de la filière Actuariat  
et l'admission à l'Institut des Actuaire  
le 14/03/2022

Par : **Maradonna CLINTON**

Titre : **Une nouvelle perspective de modélisation  
des arbitrages en assurance vie**

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

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

*Membres présents du jury de la filière*

*Nom : Caroline HILLAIRET*

*Membres présents du jury de l'Institut  
des Actuaire :*

*Entreprise :*

GENERALI

*Signature :*



*Directeur du mémoire en entreprise :*

*Nom : Adingra KOUAME*

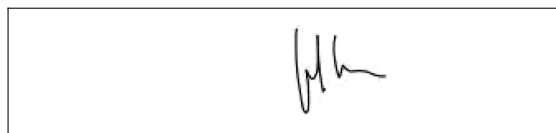
*Signature :*



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

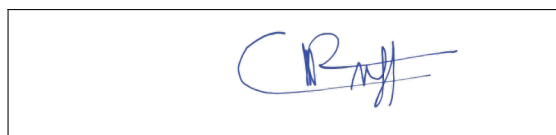
Secrétariat :

Signature du responsable entreprise



Bibliothèque :

Signature du candidat





# REMERCIEMENTS

Je tiens à remercier en premier lieu mon encadreur de stage Adingra KOUAME et le directeur de la Fonction Actuarielle de GENERALI Cédric OLLIVIER pour l'encadrement et pour m'avoir donné l'opportunité de faire un stage au sein l'équipe.

Ensuite, J'adresse mes sincères remerciements à Frederique FERRY et à Arnaud MEURIN pour le support métier et administratif tout au long de mon stage.

Je souhaite remercier également Caroline HILLAIRET qui est à la fois mon réfèrent pédagogique et responsable de la voie actuariat de l'ENSAE pour le suivi et les échanges qui ont toujours été constructifs.

Un grand merci à l'équipe Fonction Actuarielle et à l'équipe de la Valeur pour leur disponibilité, leur accueil et leur sympathie.

Enfin, Je remercie ma famille et mes amis pour leur soutien et pour leur encouragement.

# SOMMAIRE

<b>Remerciements</b> . . . . .	<b>i</b>
<b>Liste des Tableaux</b> . . . . .	<b>v</b>
<b>Liste des Figures</b> . . . . .	<b>vii</b>
<b>Résumé</b> . . . . .	<b>viii</b>
<b>Abstract</b> . . . . .	<b>ix</b>
<b>Note de synthèse</b> . . . . .	<b>x</b>
<b>Executive Summary</b> . . . . .	<b>xiv</b>
<b>Introduction</b> . . . . .	<b>1</b>
<b>1 Éléments de contexte et objectifs de l'étude</b> . . . . .	<b>3</b>
1.1 Contrat d'assurance vie . . . . .	3
1.1.1 Les différents modes de gestion en assurance vie . . . . .	5
1.1.2 Les options et garanties . . . . .	6
1.2 L'option et le risque d'arbitrage . . . . .	7
1.2.1 L'option d'arbitrage . . . . .	7
1.2.2 État des lieux des arbitrages . . . . .	9
1.2.3 Le risque d'arbitrage . . . . .	13
<b>2 Cadre théorique et une nouvelle perspective de modélisation des arbitrages</b> <b>15</b>	<b>15</b>
2.1 Revue de la littérature . . . . .	15
2.1.1 Approche microéconomique . . . . .	15
2.1.2 Approche statistique . . . . .	17
2.2 Une nouvelle perspective de modélisation des arbitrages . . . . .	18
2.2.1 Régression en deux étapes . . . . .	18
2.2.2 Une approche par l'espérance conditionnelle . . . . .	20
<b>3 Les modèles de machine learning</b> . . . . .	<b>22</b>
3.1 Le modèle de régression linéaire . . . . .	22
3.2 Le modèle GLM . . . . .	24
3.2.1 La régression logistique . . . . .	25
3.2.2 La régression tanh . . . . .	27
3.3 L'arbre de décision CART . . . . .	28
3.4 Le XGBoost . . . . .	32
3.5 Interprétation SHAP (Shapley Additive exPlanations) . . . . .	34

3.6	Intervalle de confiance . . . . .	35
3.6.1	Le percentile bootstrap . . . . .	35
3.6.2	La régression quantile . . . . .	36
<b>4</b>	<b>Construction et description du portefeuille . . . . .</b>	<b>38</b>
4.1	Construction de la base . . . . .	38
4.1.1	Définition du périmètre d'étude . . . . .	38
4.1.2	Choix et Justifications des variables explicatives . . . . .	39
4.1.3	Traitement des valeurs manquantes et des valeurs aberrantes . . . . .	43
4.2	Analyse exploratoire . . . . .	44
4.2.1	Analyse univariée . . . . .	44
4.2.2	Test d'indépendance . . . . .	45
4.2.3	Analyse de la distribution des variables cibles . . . . .	46
4.2.4	Analyse bivariée . . . . .	48
4.3	Découpage en échantillon d'entraînement et en échantillon test . . . . .	52
4.4	Choix des métriques d'évaluation . . . . .	53
4.4.1	Métrique pour la fréquence . . . . .	53
4.4.2	Métrique pour le taux d'arbitrage . . . . .	54
<b>5</b>	<b>Implémentation des modèles d'arbitrage structurel et conjoncturel . . . . .</b>	<b>56</b>
5.1	Modèle GLM fréquence . . . . .	57
5.2	Modèle XGBoost Fréquence . . . . .	62
5.3	Régression en deux étapes du taux d'arbitrage . . . . .	67
5.4	Modèle GLM taux d'arbitrage . . . . .	67
5.4.1	Modèle GLM taux d'arbitrage structurel . . . . .	67
5.4.2	Modèle GLM taux d'arbitrage conjoncturel . . . . .	70
5.5	Modèle XGBoost taux d'arbitrage . . . . .	72
5.5.1	Modèle XGBoost taux d'arbitrage structurel . . . . .	72
5.5.2	Modèle XGboost taux d'arbitrage conjoncturel . . . . .	74
<b>6</b>	<b>Analyse de l'impact d'introduction d'une loi d'arbitrage . . . . .</b>	<b>79</b>
6.1	Le modèle ALM . . . . .	79
6.1.1	Fonctionnement du modèle . . . . .	79
6.1.2	Le Cash-Flow du Passif . . . . .	80
6.1.3	Le Cash-flow de l'actif . . . . .	81
6.1.4	Le scénario d'équivalent certain . . . . .	81
6.1.5	Les scénarios stochastiques . . . . .	81
6.2	Le Best Estimate . . . . .	82
6.3	La PVFP . . . . .	83
6.4	Modélisation du taux d'arbitrage par groupe de contrat . . . . .	84
6.4.1	Arbitrages structurels . . . . .	85
6.4.2	Loi QIS5 . . . . .	85
6.4.3	Détermination des taux extrémaux . . . . .	87
6.4.4	Modèle GLM . . . . .	94
6.5	Impact sur le Best Estimate . . . . .	95

---

<b>Conclusion</b> .....	<b>97</b>
<b>Bibliographie</b> .....	<b>99</b>
<b>Annexes.</b> .....	<b>101</b>

# LISTE DES TABLEAUX

<b>Tableau 1</b> - Performances des modèles sur les données test . . . . .	xi
<b>Tableau 2</b> - Performances des modèles taux d'arbitrages . . . . .	xii
<b>Tableau 3</b> - RMSE de la moyenne mobile par ordre et $\alpha$ . . . . .	xii
<b>Tableau 4</b> - BE stochastique et PVFP stochastique par modèle . . . . .	xiii
<b>Tableau 5</b> - Model performance on test data . . . . .	xv
<b>Tableau 6</b> - Performance of arbitrage rate models . . . . .	xvi
<b>Tableau 7</b> - RMSE by order and $\alpha$ . . . . .	xvi
<b>Tableau 8</b> - Stochastic BE and stochastic PVFP . . . . .	xvii
<b>Tableau 4.1</b> - Liste des variables. . . . .	43
<b>Tableau 4.2</b> - Tableau des tests de Khi 2 . . . . .	46
<b>Tableau 4.3</b> - Matrice de confusion. . . . .	54
<b>Tableau 5.1</b> - Résultat GLM fréquence. . . . .	58
<b>Tableau 5.2</b> - Performances des modèles . . . . .	66
<b>Tableau 5.3</b> - Résultat GLM structurel . . . . .	69
<b>Tableau 5.4</b> - Résultat GLM conjoncturel . . . . .	71
<b>Tableau 5.5</b> - Performance des modèles taux d'arbitrages . . . . .	76
<b>Tableau 5.6</b> - Taux d'arbitrage annuel sur les données tests . . . . .	77
<b>Tableau 6.1</b> - Poids des fonds . . . . .	84
<b>Tableau 6.2</b> - RMSE par ordre et $\alpha$ . . . . .	85
<b>Tableau 6.3</b> - Test d'ADF et KPSS par fonds en euros . . . . .	91
<b>Tableau 6.4</b> - Paramètres par fonds en euros . . . . .	92
<b>Tableau 6.5</b> - Taux extremums obtenus par le modèle de Vasicek . . . . .	93
<b>Tableau 6.6</b> - Taux extremums dans l'historique . . . . .	93
<b>Tableau 6.7</b> - RMSE par méthode . . . . .	94
<b>Tableau 6.8</b> - Résultat modèle GLM avec écart des taux . . . . .	94
<b>Tableau 6.9</b> - BE déterministe et PVFP déterministe par modèle . . . . .	95
<b>Tableau 6.10</b> - BE stochastique et PVFP stochastique par modèle . . . . .	96
<b>Tableau 11</b> - Exemple tableau de contingence . . . . .	107

# LISTE DES FIGURES

<b>Figure 1.1</b> - Nombre d'arbitrages entre 2013 et 2020 .....	10
<b>Figure 1.2</b> - Répartition de la PM .....	11
<b>Figure 1.3</b> - Taux d'arbitrages entre 2013 et 2020 .....	12
<b>Figure 3.1</b> - Illustration de la restriction de l'espace d'optimisation par le Lasso et le Ridge .	24
<b>Figure 3.2</b> - La fonction tanh .....	27
<b>Figure 4.1</b> - Histogramme des variables quantitatives .....	44
<b>Figure 4.2</b> - Diagramme en bâtons des variables qualitatives.....	45
<b>Figure 4.3</b> - Histogramme du taux d'arbitrage.....	47
<b>Figure 4.4</b> - Histogramme du taux d'arbitrage par année.....	47
<b>Figure 4.5</b> - Fréquence d'arbitrage .....	48
<b>Figure 4.6</b> - Fréquence d'arbitrage par année.....	48
<b>Figure 4.7</b> - Matrice de corrélation des variables continues .....	49
<b>Figure 4.8</b> - Matrice de V de Cramer des variables qualitatives.....	49
<b>Figure 5.1</b> - Effets marginaux pour la fréquence GLM.....	61
<b>Figure 5.2</b> - Importance des variables dans le modèle GLM Fréquence .....	62
<b>Figure 5.3</b> - Importance des variables dans le modèle XGBoost Fréquence .....	65
<b>Figure 5.4</b> - Importance des variables par la méthode SHAP.....	66
<b>Figure 5.5</b> - Importance des variables dans le modèle XGboost structurel .....	73
<b>Figure 5.6</b> - Importance des variables dans le modèle XGboost structurel par la méthode SHAP .....	74
<b>Figure 5.7</b> - Importance des variables dans le modèle XGboost conjoncturel .....	75
<b>Figure 5.8</b> - Importance des variables dans le modèle XGboost conjoncturel par la méthode SHAP .....	76
<b>Figure 5.9</b> - Comparaison des taux d'arbitrage annuels .....	77
<b>Figure 6.1</b> - Regroupement des contrats.....	81
<b>Figure 6.2</b> - Architecture d'un modèle de projection Actif/Passif .....	84
<b>Figure 6.3</b> - Loi QIS .....	86
<b>Figure 6.4</b> - Evolution du taux d'arbitrage mensuel du fonds en euros .....	87
<b>Figure 6.5</b> - ACF et PACF du fonds en euros.....	88
<b>Figure 6.8</b> - ACF des fonds en euros.....	90
<b>Figure 6.11</b> - PACF des fonds en euros.....	91



---

<b>Figure 12</b> - Arbitrage par PM .....	101
<b>Figure 13</b> - Arbitrage par part en UC .....	101
<b>Figure 14</b> - Arbitrage par nombre de fonds en UC .....	101
<b>Figure 15</b> - Arbitrage par âge .....	102
<b>Figure 16</b> - Arbitrage par ancienneté .....	102
<b>Figure 17</b> - Arbitrage par sexe .....	102
<b>Figure 18</b> - Arbitrage par mode de gestion .....	103
<b>Figure 19</b> - Arbitrage par indice de périodicité des primes .....	103
<b>Figure 20</b> - Arbitrage par TMG .....	103
<b>Figure 21</b> - Arbitrage par Taux de PB .....	104
<b>Figure 22</b> - Evolution des arbitrages et de la cotation du CAC 40.....	104
<b>Figure 23</b> - Principe générale de la validation croisée .....	105
<b>Figure 24</b> - Exemples de corrélation .....	106
<b>Figure 25</b> - Force plot SHAP .....	110

## RÉSUMÉ

Dans le contexte économique actuel où le taux de rendement du fonds en euros baisse et la performance des fonds en unités de compte est volatile à cause de la survenance des récentes crises financière et sanitaire, les assurés se retrouvent dans une situation d'incertitude entre garder leur placement dans un fonds en euros dont le rendement ne fait que baisser ou de réorienter vers un fonds en unités de compte avec une espérance de rendement élevée, mais très volatile (ou inversement).

Une réorientation massive du placement des assurés vers un fonds en euros ou un fonds en UC peut avoir des impacts sur la solvabilité de l'assureur. Ainsi, une meilleure perception du comportement d'arbitrage des assurés permettra à l'assureur d'améliorer sa solvabilité et la prédiction de la valeur de ses engagements envers les assurés à l'avenir (Best Estimate).

Ainsi, dans ce contexte, les travaux porteront sur la modélisation du comportement des arbitrages structurels et conjoncturels des assurés par des méthodes statistiques (GLM) et des modèles de machine learning à la maille contrat et à la maille groupe de contrats ainsi que l'analyse de l'impact de la prise en compte de cette loi sur le Best Estimate.

**Mots clés** : arbitrage, Best Estimate, fonds en euros, fonds en UC, GLM, machine learning.

## ABSTRACT

In the current economic context, where the rate of return on the euro fund is falling and the performance of unit-linked funds is volatile due to the recent financial and health crises, policyholders find themselves in a situation of uncertainty between keeping their investment in a euro fund with a declining return or switching to a unit-linked fund with a high, but very volatile, expected return (or vice versa).

A massive reorientation of the policyholders' investment towards the euro fund or the unit-linked fund can have impacts on the insurer's solvency, thus, a better perception of the policyholders' arbitrage behaviour will improve its solvency and the prediction of the value of its commitments towards the policyholders in the future (Best Estimate).

Thus, in this context, the work will focus on modelling the behaviour of structural and dynamic arbitrages of policyholders using statistical methods (GLM) and machine learning models at the contract and group level, as well as analysing the impact of taking this law into account on the Best Estimate.

**Key words** : arbitrage, Best Estimate, euro funds, unit-linked funds, GLM, machine learning.

## NOTE DE SYNTHÈSE

Dans le contexte économique actuel, caractérisé par une baisse du taux de rendement du fonds en euros et la volatilité de la performance des fonds en unités de compte causée par la survenance des récentes crises financière et sanitaire, les assurés se retrouvent dans une situation d'incertitude entre garder leur placement sur un fonds en euros dont le rendement ne fait que baisser ou se réorienter vers un fonds en unités de compte avec une espérance de rendement élevée, mais très volatile (ou inversement).

Une réorientation massive du placement des assurés vers le fonds en euros peut faire baisser le rendement global des fonds en euros à cause de l'achat des nouvelles obligations à taux bas, ce qui implique qu'il sera plus difficile pour l'assureur de rémunérer les anciens contrats avec un taux minimum garanti élevé, ce qui augmente le capital à immobiliser et dégrade sa solvabilité. D'un autre côté, un arbitrage important vers le fonds unités de compte peut obliger l'assureur à vendre des obligations dans une situation de moins-value latente. La non-prise en compte du comportement d'arbitrage peut donc avoir des impacts sur la solvabilité de l'assureur. Ainsi, une meilleure perception du comportement d'arbitrage des assurés permettra d'améliorer sa solvabilité et la prédiction de la valeur de ses engagements à l'avenir.

Le comportement d'arbitrage des assurés dépend de plusieurs facteurs qui sont liés à l'assuré ou qui lui sont exogènes. Donc une séparation des arbitrages structurels qui sont dus aux facteurs structurels (caractéristiques liées à l'assuré et au contrat) et des arbitrages conjoncturels qui dépendent de la conjoncture économique (marché financier, inflation, chômage...) s'avère nécessaire. Par la suite, ce sont les indices financiers qui sont retenus comme variables conjoncturelles dans la partie modélisation, plus précisément le taux de rendement du fonds en euros et le taux de rendement du fonds en UC.

Dans un premier temps, la modélisation considérée sera une modélisation à la maille contrat. Ce modèle a l'avantage d'être plus précis et plus exhaustif. Cependant, il a l'inconvénient d'être très difficile à mettre en œuvre dans l'outil de gestion d'actif/passif de GENERALI. D'où la nécessité de considérer une modélisation par groupe de contrats dans un second temps. La modélisation individuelle permet de prédire le taux d'arbitrage à court terme avec plus de précision (1 an) alors que la modélisation par groupe de contrats permet de réaliser une projection de long terme afin d'évaluer l'impact de la prise en compte du comportement d'arbitrage des assurés sur la valeur de l'engagement de l'assureur (Best Estimate).

Le pas de temps du modèle sera annuel pour des raisons de volumétrie des données et de simplicité. C'est l'approche fréquence/févérité qui sera adoptée pour la modélisation à la maille contrat, c'est-à-dire prédire à la fois la probabilité qu'un assuré réalise un arbitrage ou non et prédire la valeur de son taux d'arbitrage (taux d'arbitrage net euros vers UC). Comme la variable réponse décision à arbitrer est binaire, nous utiliserons une régression logistique et un modèle XGBoost pour l'approcher. Comme le nombre d'assurés ayant effectué des arbitrages dans le portefeuille est très faible par rapport au nombre d'assurés qui n'ont pas arbitré, une méthode d'oversampling (ramener le nombre de contrats de la classe minoritaire au même nombre que la classe majoritaire) et une méthode d'undersampling (ramener le nombre de contrats de la classe majoritaire au même nombre que la classe minoritaire) ont été réalisées sur la base des données, ce qu'on appelle une méthode mélange. La distribution passe donc de  $\{86\%, 14\%\}$  à  $\{70\%, 30\%\}$ . Les deux bases seront testées sur les deux modèles.

TABLEAU 1 – Performances des modèles sur les données test

Modèles	GLM mélange	GLM normal	XGBoost normal	XGBoost mélange
Taux de bien classés	82.24 %	86.35 %	85.89 %	83.35 %
Sensitivité	87.36 %	98.55%	96.53 %	91.57 %
Spécificité	57.98 %	27.28%	34.80 %	43.91 %
LogLoss	0.3792	0.3467	0.3489	0.3727

Le tableau 1 montre que c'est le modèle XGBoost avec une base mélange qui performe le mieux sur notre jeu de données si nous regardons les 4 indicateurs en même temps. La spécificité étant le taux des clients ayant réalisés un arbitrage et classé comme tel, et la sensibilité est le taux de bon classement des clients qui n'ont pas réalisé un arbitrage.

Pour séparer la partie conjoncturelle de la partie structurelle du taux d'arbitrage, une modélisation en deux étapes a été faite. Nous supposons que le taux d'arbitrage de l'assuré est la somme du taux d'arbitrage structurel ( $Y_s$ ) et du taux d'arbitrage dynamique ( $Y_c$ ), que l'espérance du taux d'arbitrage conjoncturel est nulle et que le taux d'arbitrage dynamique est indépendant des facteurs structurels autrement dit tous les assurés réagissent de la même manière suite à des changements de la conjoncture économique (rendement des fonds). Ces deux dernières hypothèses sont très fortes mais nécessaires pour estimer séparément les deux types d'arbitrage.

On a montré que :

$$\begin{aligned}\mathbb{E}(Y | X_s) &= \mathbb{E}(Y_s | X_s) + \mathbb{E}(Y_c | X_s) = \mathbb{E}(Y_s | X_s) + \mathbb{E}(Y_c) \quad \text{car } Y_c \perp\!\!\!\perp X_s \\ &= \mathbb{E}(Y_s | X_s) \quad \text{car } \mathbb{E}(Y_c) = 0\end{aligned}$$

Cette espérance conditionnelle sera approchée par un modèle GLM et un modèle XGBoost

en utilisant les variables structurelles pour modéliser le taux d'arbitrage structurel. Le résidu du premier modèle sera donc considéré comme le taux d'arbitrage dynamique et sera approché par un modèle GLM et XGBoost en utilisant les variables conjoncturelles. Notons que le taux d'arbitrage varie entre -1 et 1 et possède une distribution quasi symétrique, donc nous avons défini un GLM avec une famille de loi normale et une fonction de lien  $\operatorname{arctanh}(x)$  que nous appelons la régression tanh.

Le tableau 2 montre que le XGBoost en deux étapes a une meilleure performance prédictive et explicative que le GLM tanh. L'analyse de l'importance des variables du modèle XGBoost a montré que ce sont les variables liées à l'aversion au risque (part en UC dans le contrat et montant de la provision mathématique) sont celles qui ont le plus d'influence sur le comportement d'arbitrage.

TABLEAU 2 – Performances des modèles taux d'arbitrages

Modèles	GLM tanh + LR	XGBoost
RMSE_train	0.4570	0.4027
RMSE_test	0.5958	0.5754
$R^2$	0.2787	0.5189

Pour la modélisation par groupe de contrats, une moyenne mobile pondérée a été adoptée pour modéliser le taux d'arbitrage structurel. Nous supposons que dans le calcul de la moyenne mobile, le poids des arbitrages de l'année  $n - 1$  est égal à 1 et le poids des autres arbitrages antérieurs décroît d'un facteur constant  $\alpha$  lors qu'on recule d'une année dans le passé afin de mieux représenter le comportement d'arbitrage des dernières années. Le choix des paramètres de la moyenne mobile est réalisé sur la base de la RMSE.

$$x_t = \sum_{i=1}^p \frac{\frac{1}{\alpha^{i-1}} x_{t-i}}{\sum_{j=1}^p \frac{1}{\alpha^{j-1}}}$$

TABLEAU 3 – RMSE de la moyenne mobile par ordre et  $\alpha$

$\alpha$	Ordre			
	2	3	4	5
2	0.0073	0.0069	0.0066	0.0074
3	0.0074	0.0071	0.0069	0.0069
4	0.0075	0.0073	0.0072	0.0069
5	0.0076	0.0074	0.0074	0.0074

Il en ressort que  $\alpha = 2$  et l'ordre de la moyenne mobile égal à 4 qui minimise la valeur de

la RMSE.

Le modèle standard utilisé pour modéliser les arbitrages dynamiques est la loi QIS 5 qui est préconisée pour la modélisation des rachats dynamiques. Elle suppose que le taux d'arbitrage dynamique est fonction de l'écart entre le taux de rendement du fonds en euros et le taux de rendement du fonds en UC. C'est une modélisation par une fonction linéaire par morceau, supposant que si cette différence de taux reste entre deux bornes, il n'y a pas d'arbitrage conjoncturel. En revanche, si elle sort de cette plage, les arbitrages dynamiques augmentent ou diminuent linéairement jusqu'à l'atteinte d'un taux maximum ou minimum. Les taux de déclenchement et les taux de stabilisations sont déterminés sur l'historique ou sur la base d'un modèle GLM (prise en compte des effets non linéaires). Les taux extrêmes sont obtenus sur l'historique ou sur la base d'un modèle de Vasicek.

Le deuxième modèle proposé est un modèle GLM (régression tanh). Une discrétisation de l'écart de rendement des taux a été réalisée pour prendre en compte son effet non-linéaire sur le taux d'arbitrage dynamique.

TABLEAU 4 – BE stochastique et PVFP stochastique par modèle

Modèle	Taux extrêmes	BE	PVFP
Sans Arbitrage		74 339 974 864	2 449 618 695
Loi QIS 5	Historique	74 621 081 315 (0.3781%)	2 168 579 806 (-11.473%)
Loi QIS 5	Modèle	74 927 713 511 (0.7906%)	1 859 739 086 (-24.080%)
GLM		74 643 142 923 (0.4078%)	2 145 520 527 (-12.414%)

Le tableau 4 montre que l'impact de la loi QIS 5 dépend très fortement de la nature des taux extrêmes utilisés dans le modèle. En plus d'avoir été calibré sur la base de l'historique, le modèle GLM présente un résultat robuste.

## EXECUTIVE SUMMARY

In the current economic context, characterised by a fall in the rate of return on the euro fund and the volatility of the performance of unit-linked funds caused by the recent financial and health crises, policyholders therefore find themselves in a situation of uncertainty between keeping their investment in a euro fund with a falling return or switching to a unit-linked fund with a high, but very volatile, return expectation (or vice versa).

A massive reorientation of the policyholders' investment towards the euro fund can lower the overall yield of the euro funds because of the purchase of new bonds at low rates, which implies that it will be more difficult for the insurer to remunerate the old contracts with a high minimum guaranteed rate, which increases the capital to be immobilised and degrades its solvency. On the other hand, a significant arbitrage towards the unit-linked fund can oblige the insurer to sell bonds in a situation of unrealised capital loss. Not taking arbitrage behaviour into account can therefore have an impact on the insurer's solvency. Thus, a better perception of the policyholders' arbitrage behaviour will improve its solvency and the prediction of the value of its commitments in the future.

The arbitration behaviour of policyholders depends on several factors which are linked to the policyholder and which are exogenous to him. Therefore, it is necessary to separate structural arbitrage, which is due to structural factors (characteristics linked to the policyholder and the contract), from cyclical arbitrage, which depends on the economic situation (financial market, inflation, unemployment, etc.). In the following, the financial indices are used as the economic variable in the modelling, more precisely the rate of return on the euro fund and the rate of return on the unit-linked fund.

Initially, the model considered will be a contract mesh model. This model has the advantage of being more precise and exhaustive. However, it has the disadvantage of being very difficult to implement in GENERALI's asset/liability management tool. Hence the need to consider modelling by group of contracts in a second step. Individual modelling allows the prediction of the short term arbitrage rate with more accuracy (1 year) whereas group modelling allows a long term projection in order to evaluate the impact of the policyholders' arbitrage behaviour on the insurer's commitment value (Best Estimate).

The time step of the model will be annual for reasons of data volume and simplicity. The Frequency/Severity approach will be adopted for modelling at the contract level, i.e. predicting



both the probability that a client will arbitrate or not and predicting the value of his arbitration rate (net euro to unit-linked arbitration rate). As the variable response to arbitrage is binary, we will use a logistic regression and an XGBoost model to approximate it. As the number of policyholders who have arbitrated in the portfolio is very small compared to the number of policyholders who have not arbitrated, an oversampling method (reducing the number of the majority class to the same number as the minority class) and an undersampling method (reducing the number of the minority class to the same number as the majority class) have been performed on the data, which is called the mixture method. The distribution thus changes from {86 % , 14 %} to {70 % , 30 %}. The two bases will be tested on the two models.

TABLEAU 5 – Model performance on test data

Model	GLM mix	normal GLM	normal XGBoost	XGBoost mix
Accuracy	82.24 %	86.35 %	85.89 %	83.35 %
Sensitivity	87.36 %	98.55%	96.53 %	91.57 %
Specificity	57.98 %	27.28%	34.80 %	43.91 %
LogLoss	0.3792	0.3467	0.3489	0.3727

Table 5 shows that the XGBoost model with a mixed base performs best on our dataset if we look at the four indicators at the same time. The specificity is the rate of well ranked clients who have have done arbitrage and the sensitivity is the rate of well ranked clients who haven't have done arbitrage.

In order to separate the cyclical from the structural part of the arbitrage rate, a two-step modelling has been done. We assume that the policyholder's arbitrage rate is the sum of the structural arbitrage rate ( $Y_s$ ) and the dynamic arbitrage rate ( $Y_c$ ), that the expectation of the cyclical arbitrage rate is zero and that the dynamic arbitrage rate is independent of structural factors, i.e. all policyholders react in the same way to changes in economic conditions (fund returns). These last two assumptions are very strong but necessary to estimate the two types of arbitrage separately.

$$\begin{aligned} \mathbb{E}(Y | X_s) &= \mathbb{E}(Y_s | X_s) + \mathbb{E}(Y_c | X_s) = \mathbb{E}(Y_s | X_s) + \mathbb{E}(Y_c) \quad , \quad Y_c \perp\!\!\!\perp X_s \\ &= \mathbb{E}(Y_s | X_s) \quad \text{because} \quad \mathbb{E}(Y_c) = 0 \end{aligned}$$

This conditional expectation will therefore be approximated by a GLM model and an XGBoost model using the structural variables to model the structural arbitrage rate. The residual of the first model will therefore be considered as the dynamic arbitrage rate and will be approximated by a GLM and XGBoost model using the cyclical variables. Note that the arbitrage rate varies between -1 and 1 and has a quasi symmetrical distribution, so we have defined a GLM

with a normal distribution family and a link function  $\operatorname{arctanh}(x)$  that we call tanh regression.

The table 6 shows that the two-step XGboost has a better predictive and explanatory performance than the tanh GLM. The analysis of the importance of the variables of the XGboost model showed that the variables linked to risk aversion (Unit-linked share in the contract and Amount of the mathematical provision) have the greatest influence on the arbitrage behaviour.

TABLEAU 6 – Performance of arbitrage rate models

Model	GLM tanh + LR	XGboost
RMSE_train	0.4570	0.4027
RMSE_test	0.5958	0.5754
$R^2$	0.2787	0.5189

For the modelling by contract group, a weighted moving average has been adopted to model the structural arbitrage rate. We assume that in the calculation of the moving average, the weight of arbitrages in year  $n - 1$  is equal to 1 and the weight of other previous arbitrages decreases by a constant factor  $\alpha$  when going back one year in the past in order to better represent the arbitration behaviour of the last years. The choice of the parameters of the moving average is made on the basis of the RMSE.

$$x_t = \sum_{i=1}^p \frac{\frac{1}{\alpha^{i-1}} x_{t-i}}{\sum_{j=1}^p \frac{1}{\alpha^{j-1}}}$$

TABLEAU 7 – RMSE by order and  $\alpha$

$\alpha$	order	2	3	4	5
	2	0.0073	0.0069	0.0066	0.0074
3	0.0074	0.0071	0.0069	0.0069	
4	0.0075	0.0073	0.0072	0.0069	
5	0.0076	0.0074	0.0074	0.0074	

This shows that  $\alpha = 2$  and the order of the moving average is 4 which minimises the value of the RMSE.

The standard model used to model dynamic arbitrage is the QIS 5 law which is recommended for modelling dynamic redemptions. It assumes that the dynamic arbitrage rate is a function of the difference between the rate of return on the euro fund and the rate of return on the unit-linked fund. It is a piecewise linear function model, assuming that if this rate difference remains between two bounds, there is no cyclical arbitrage. However, if it falls outside

this range, dynamic arbitrage increases or decreases linearly until a maximum or minimum rate is reached. Trigger rates and stabilisation rates are determined on the basis of historical data or a GLM model (taking into account non-linear effects). Extreme rates are obtained on the basis of historical data or on the basis of a Vasicek model.

The second model proposed is a GLM model (tanh regression). A discretization of the yield spread has been performed to take into account its non-linear effect on the dynamic arbitrage rate.

TABLEAU 8 – Stochastic BE and stochastic PVFP

Model	Extremum rate	BE	PVFP
Without arbitration		74 339 974 864	2 449 618 695
QIS 5 law	Historical	74 621 081 315 (0.3781%)	2 168 579 806 (-11.473%)
QIS 5 law	Model	74 927 713 511 (0.7906%)	1 859 739 086 (-24.080%)
GLM		74 643 142 923 (0.4078%)	2 145 520 527 (-12.414%)

The table 8 shows that the impact of the QIS 5 law depends very strongly on the nature of the extremum rates used in the model. In addition to having been calibrated on the basis of history, the GLM model presents a robust result.

## Introduction

L'environnement économique des assureurs est bouleversé depuis la survenance de la crise de 2008 et depuis la baisse du rendement des obligations. En France, les obligations sont les titres les plus détenus par les assureurs. En effet, les fonds en euros proposés par les assureurs aux assurés sont à dominance obligataires. De plus, l'instabilité des marchés financiers due aux récentes crises financière et crise sanitaire laisse beaucoup d'incertitude sur le rendement des placements en unités de compte.

Comme la rentabilité et le risque sont des notions très importantes en assurance-vie, cette situation met donc les assurés dans une situation de doute : placer son argent sur un fonds en euros à faible rendement, mais avec plus de sécurité ou dans un fonds en unités de compte avec une espérance de gain élevée, mais plus risqué. Certains assurés se verront donc obliger de changer la nature de leur support d'investissement pour des raisons économiques (arbitrages conjoncturels) ou pour des raisons propres à eux-mêmes (arbitrages structurels). Un changement brusque de la nature des placements des assurés peut avoir un impact sur la solvabilité et l'engagement des assureurs. Par exemple, un arbitrage massif du fonds en unités de compte vers le fonds en euros augmente le besoin en capital de l'assureur. La question se pose donc : comment l'assureur peut-il appréhender le comportement d'arbitrage pour mieux le prédire ? Quel aspect du comportement des assurés est le plus déterminant dans sa décision d'arbitrage ? Et dans quelle mesure le comportement d'arbitrage des assurés impact la valeur de l'engagement de l'assureur ?

L'assureur calcul la valeur de ses engagements à travers un outil de gestion d'actif/passif (le logiciel « Prophet » pour GENERALI). La projection des flux dans cet outil se fait en général par groupe de contrats. La modélisation contrat par contrat permet d'avoir une vision plus précise à court terme (1 an), mais très coûteuse en terme d'implémentation pour une projection de long terme. En revanche, une modélisation par groupe de contrats permet d'avoir une projection de long terme du comportement d'arbitrage des assurés, ce qui permet d'évaluer son impact sur la valeur de l'engagement de l'assureur.

L'objectif de cette étude est donc de modéliser le comportement d'arbitrage des assurés en assurance-vie, en utilisant l'approche fréquence/sévérité souvent utilisée en assurance non vie pour le modèle individuel et en utilisant une approche statistique pour le modèle par groupe de contrats.

Toutefois, il est important de distinguer la cause ou la source qui influe sur le comportement d'arbitrage des assurés. En effet, une meilleure compréhension du comportement des assurés permet à l'assureur d'avoir une bonne anticipation de ses engagements et lui permet donc d'améliorer sa solvabilité.

Pour la modélisation à la maille contrat, un modèle de fréquence (probabiliste) sera mis en place pour modéliser la décision à arbitrer ou non de l'assuré. Ensuite, une modélisation des arbitrages structurels et conjoncturels entre le fonds en euros et le fonds en UC à la maille

contrat sera mise en œuvre pour estimer le montant d'arbitrage de l'assuré. La modélisation en deux étapes sera adoptée dans le cadre de cette étude pour séparer les arbitrages conjoncturels des arbitrages structurels (première étape pour l'arbitrage structurel et deuxième étape pour l'arbitrage conjoncturel). Nous utiliserons la mesure de l'importance des variables pour l'interprétation des modèles. Ainsi, le modèle final permet à la fois d'estimer la probabilité qu'un assuré effectue un arbitrage ou non, le taux d'arbitrage et de séparer l'arbitrage structurel et l'arbitrage dynamique.

Pour la modélisation par groupe de contrats, les arbitrages structurels seront modélisés par une loi moyenne et les arbitrages conjoncturels par une loi statistique.

## ÉLÉMENTS DE CONTEXTE ET OBJECTIFS DE L'ÉTUDE

L'objectif de ce chapitre est de présenter les concepts généraux de l'assurance vie et la problématique liée au risque d'arbitrage. Il présente les différents types de contrat en assurance vie, les options et les risques d'arbitrage, ce qui permet de montrer la nécessité de les modéliser.

### 1.1 Contrat d'assurance vie

Le contrat d'assurance vie est une forme de placement qui permet à l'assuré de constituer un capital auprès de l'assureur dans le but de recevoir une prestation ou la transmettre à un ou plusieurs bénéficiaires lors de la survenance d'un événement lié à sa vie : décès, survie ou retraite. C'est un contrat qui se repose sur le principe d'épargne par capitalisation : l'assureur s'engage à verser un capital qui sera valorisé chaque année à l'assuré, moyennant le versement de primes.

Il existe deux types de contrats d'assurance vie : les contrats d'assurance en cas de vie qui consiste à verser le capital à l'échéance du contrat si l'assuré est toujours en vie, et les contrats d'assurance en cas de décès qui consiste à verser le capital à un bénéficiaire si l'assuré décède avant le terme du contrat. Du point de vue de l'objectif de placement, nous distinguons trois différents types de contrats d'assurance vie à savoir : les contrats d'épargne, les contrats de retraite et les contrats de prévoyance. Par la suite, l'étude se portera uniquement sur les contrats d'épargne.

Théoriquement, le capital placé par un assuré sur un contrat d'assurance vie ne lui appartient plus comme une somme d'argent sur un compte courant mais se transforme en une créance qu'il détient à l'égard de l'assureur. La créance peut être rachetée partiellement ou en totalité à tout moment (c'est la raison pour laquelle on ne parle pas du retrait d'un contrat d'assurance vie mais plutôt du rachat). L'assureur à son tour va adosser cette créance sur des actifs (actions, obligations...) inscrits à son bilan (actifs dont il est propriétaire), le contrat sera valorisé chaque année en fonction des rendements de ces actifs. Le choix des supports de placement reste primordial lors de la souscription à un contrat d'assurance vie car la rémunération du contrat en dépend. Les actifs peuvent être sélectionnés par l'assureur (généralement pour les fonds en euro) ou par l'assuré pour les fonds en unités de compte (ou UC).

Il existe 3 types de supports en assurance vie :

— Les fonds en euros : ce sont des fonds d'investissement sécuritaires, car le placement est

sans risque en échange d'un rendement modéré. Le capital et les intérêts sont garantis par l'assureur en contrepartie d'un taux de rendement plus faible. De plus, les fonds en euros disposent d'un effet cliquet, c'est-à-dire que les intérêts annuels versés sont définitivement acquis, s'ajoutent au capital garanti et génèrent à leur tour des intérêts pour les années qui s'ensuivent. Ce sont des fonds destinés à des assurés qui ont de l'aversion pour le risque, c'est l'assureur qui supporte tous les risques. Cependant, la gestion est effectuée par l'assureur uniquement donc l'assuré est passif. Les fonds en euros sont constitués en général par des obligations.

- Les fonds en unités de compte : contrairement aux fonds en euros, les fonds en unités de compte sont des fonds qui n'offrent pas des garantis en capital, mais l'espérance et la volatilité de leurs rendements sont supérieures à celles des fonds en euros. Ce sont des supports d'investissement financiers tels que les actions, les obligations ou l'immobilier. L'assureur ne garantit plus le capital, mais plutôt le nombre de parts en unités de compte détenu par ses clients. C'est donc l'assuré qui supporte tous les risques en cas d'une baisse de la valeur de ses unités de compte.
- Les fonds euro-croissance : moins risqués que les fonds en unités de compte, ces fonds offrent une espérance de rendement supérieur aux fonds en euros. Les fonds euro-croissance possèdent une garantie en capital à échéance, un pourcentage du capital investi sur ces fonds est garanti au terme d'une durée d'au moins 8 ans, qui sont définis au moment de la souscription du contrat. Les fonds euro-croissance ont été créés dans le but de faire un compromis entre espérance du rendement élevée et garantie en capital.

Il est possible de détenir un ou plusieurs fonds en euros, un ou plusieurs fonds en unités de compte et un ou plusieurs fonds euro-croissance sur un contrat d'assurance vie. En général, il existe deux types de contrat d'assurance vie :

- Contrat d'assurance vie monosupport : tout le capital de l'assuré est placé sur des supports financiers de même nature. Ce sont des contrats qui s'adressent généralement à des assurés qui ont un profil purement prudent ou un profil purement risqué. Dans ce type de contrat soit le capital est placé à 100% sur des fonds en euros ou à 100% sur des fonds en UC ou à 100% sur des fonds euro-croissance.
- Contrat d'assurance vie multisupport : contrairement au contrat monosupport, un contrat multisupport permet d'investir sur différents supports de nature différente, selon une répartition qui est laissée au choix de l'assuré. L'assuré va répartir son capital de sorte qu'une partie du capital soit sécurisée et assortie de garantie et le reste sera placé sur des fonds qui ont une perspective de rendement plus élevée à long terme. La fraction du

capital investie sur les unités de compte sera exposée directement aux fluctuations du marché financier, immobilier ou obligataire et l'assuré pourrait réaliser directement des pertes.

### 1.1.1 Les différents modes de gestion en assurance vie

Il est primordial de savoir si l'assuré a le recul nécessaire sur le marché (financier, immobilier, ...) pour définir sa stratégie patrimoniale par lui-même ou est-ce qu'il est plus raisonnable de la déléguer à l'assureur. Lors de la souscription à un contrat d'assurance vie, plusieurs modes de gestion sont disponibles pour répondre aux besoins d'autonomie ou non et de la volonté à suivre l'évolution de la valeur du contrat afin d'obtenir le maximum de rendement et de minimiser les risques. Il existe plusieurs modes de gestion en assurance vie :

- **La gestion libre** : pour les contrats d'assurance vie multisupport ou monosupport, ce mode de gestion est toujours disponible à la souscription et elle donne une liberté totale à l'assuré sur le choix de ses supports ainsi que la possibilité de les changer tout au long de la vie du contrat. Cependant, ce mode de gestion nécessite une bonne connaissance, une bonne capacité d'analyse et un suivi de l'évolution des marchés pour mettre en place une stratégie d'investissement qui maximise le rendement tout en acceptant des risques modérés. C'est à l'assuré de choisir les fonds à privilégier, la répartition de ces fonds dans son contrat et le bon moment pour faire les arbitrages. C'est le mode de gestion par défaut en assurance vie.
  
- **La gestion pilotée ou profilée** : elle est adaptée aux assurés qui ne veulent pas se soucier de suivre l'évolution du marché financier et qui ne disposent pas des connaissances dans le domaine de la finance. La seule décision que l'assuré a à prendre, c'est le niveau de risque qu'il est prêt à accepter lors de la souscription du contrat, c'est-à-dire la part et la nature des unités de compte qu'il veut détenir dans son contrat. La gestion est opérée par l'assureur ou la société de gestion en fonction du profil choisi par le client. En général, il existe 3 types de profil :
  - \* Le profil prudent ou sécurisé : la proportion de capital placée en unité de compte est investie sur les produits obligataires et monétaires est majoritaire. Ce type de profil présente un faible risque, mais un potentiel de gain faible aussi.
  - \* Le profil équilibré : il y a autant de fonds en euros que de fonds en UC dans le contrat. C'est un profil qui adopte une stratégie de maximisation de la rentabilité et une prise de risque mesurée. Ce type de profil peut profiter des surperformances des fonds en UC comme peut-être protégé des crises sur le marché des actions à des proportions moindres.
  - \* Le profil dynamique : les actions présentant une grande espérance de rendement et



un risque élevé sont majoritaires dans le contrat.

- **La gestion à l'horizon** : elle est destinée aux assurés qui ont une stratégie de placement de long terme, c'est-à-dire aux assurés qui veulent rester pendant une période assez longue dans le portefeuille de l'assureur. La répartition du capital évoluera en fonction de l'âge et du profil de l'assuré. Le capital est majoritairement investi sur des fonds risqués au début et s'oriente vers les fonds moins risqués au fil des années.

### 1.1.2 Les options et garanties

Les contrats d'assurance vie offrent très souvent des options et garanties en faveur des assurés afin de les attirer ou de les inciter à rester dans le portefeuille. Nous retiendrons principalement deux types d'options et garanties en assurance vie :

- Les options et garanties qui dépendent du comportement des assurés : leur exécution résulte principalement de la volonté de l'assuré de les utiliser ou non.
  - \* L'option d'avance : elle permet à l'assuré d'emprunter une partie de son capital pour répondre à un besoin de trésorerie de court ou moyen terme sans que l'opération ne soit considérée comme un retrait. Cette avance ne diminue pas la valeur du contrat et continue de produire des intérêts. L'avance ne peut pas dépasser la valeur du contrat.
  - \* L'option de réduction : elle concerne uniquement les contrats à primes périodiques, c'est-à-dire que le versement de primes se fait périodiquement. Lorsque l'assuré décide de ne plus verser ou de diminuer les primes périodiques prévues dans son contrat, l'assureur n'a pas le droit de l'y obliger. La valeur de l'engagement de l'assureur envers l'assuré doit être diminuée en fonction des primes versées.
  - \* L'option de rachat : elle donne le droit à l'assuré de retirer avant la date d'échéance du contrat tout ou une partie de son capital. Le rachat total (retrait de tout son capital) met un terme au contrat alors que le rachat partiel (retrait d'une partie du capital) diminue juste la valeur des encours du contrat.
  - \* L'option d'arbitrage : elle donne le droit à l'assuré de réorienter tout ou une partie de son épargne vers d'autres supports disponibles. Cette option est assimilée à une redistribution de l'épargne dans le contrat d'assurance vie tout en gardant le même montant de capital. Il n'y a donc ni entrée ni sortie d'argent.
- Les options et garanties indépendantes du comportement des assurés : elles dépendent plutôt de la structure de placement de l'assuré et de la performance des actifs financiers sur le marché.

\* Le taux minimum garanti (TMG) :

Le TMG est le taux de rendement minimum de la provision mathématique des fonds en euros que l'assureur doit verser chaque année. Il est fixé chaque année, contrat par contrat, sans dépasser 85% du rendement des actifs de l'assureur au cours des deux dernières années et doit être exprimé de façon annualisée. Le rendement annuel du contrat de l'assuré adossé sur un fonds en euro ne peut être donc inférieur à ce taux indépendamment de la performance réalisée par les actifs sur les marchés financiers. Le TMG joue donc un rôle important dans le choix des supports d'investissement et dans la répartition du capital de l'assuré.

\* Le taux de participation aux bénéfices (PB) :

La participation aux bénéfices correspond aux montants redistribués aux assurés des bénéfices réalisés par l'assuré dans l'année. La gestion des fonds d'assurance vie donne des profits techniques et financiers, dont la majeure partie doit être redistribuée par l'assureur. En effet, le taux minimal de participation aux bénéfices est de 85 % pour le résultat financier et de 90% quant au résultat technique. Chaque année, l'assureur a l'obligation de calculer un montant minimal de participation aux bénéfices techniques (différence entre frais prélevés et frais réels) et bénéfices financiers (les gains réalisés grâce au placement de l'épargne des assurés). Les participations ne sont pas obligatoirement distribuées à la même année durant laquelle l'assureur avait réalisé des plus-values, mais peuvent être rendues à la même année ou dans un délai maximum de 8 ans. C'est pour cela qu'on fait la distinction entre taux de participation aux bénéfices contractuel et le taux de participation aux bénéfices réellement distribué.

Le taux servi sur les fonds en euros dépend donc de son TMG et de son taux de participation aux bénéfices réellement distribué. En effet, la rémunération des fonds en euros est égale à la somme du TMG et du taux de participation aux bénéfices dont on soustrait le taux de chargement.

Les options et garanties apparaissent comme un avantage pour l'assuré, mais représentent plutôt un risque pour l'assureur, il est donc important de les appréhender et de les modéliser.

## 1.2 L'option et le risque d'arbitrage

### 1.2.1 L'option d'arbitrage

L'option d'arbitrage donne le droit à l'assuré de transférer son épargne d'un ou plusieurs supports vers un ou plusieurs autres supports disponibles dans le portefeuille de l'assureur. Il s'agit ainsi d'une modification de la structure de l'épargne de l'assuré au sein de son contrat

d'assurance vie selon les supports non seulement entre le fonds en euros et les unités de compte, mais également sur les UC proposées par l'assureur. En effectuant un arbitrage, la valeur de l'encours de l'épargne de l'assuré ne change presque quasiment pas, puisqu'il n'y a ni entrée ni sortie de capitaux.

Il existe 5 types d'options d'arbitrage :

- L'arbitrage libre : il s'agit des arbitrages effectués par l'assureur lui-même permettant de transférer tout ou partie de son capital d'un support d'investissement vers un autre. L'arbitrage libre est généralement réalisé par des assurés qui ont une mode de gestion libre ce qui implique qu'il doit suivre l'évolution des marchés financiers et doit disposer des connaissances en matière d'investissement.
- L'arbitrage automatique : il consiste à automatiser certains arbitrages pour respecter la répartition initiale de l'épargne de l'assuré en cas de déséquilibre. Ce type d'arbitrage ne nécessite pas l'intervention directe de l'assuré pour être réalisé.
- Dynamisation des intérêts annuels : cette option permet de transférer automatiquement les intérêts produits par les fonds en euros pendant l'année, vers un ou plusieurs supports en unités de compte pour l'année suivante.
- Sécurisation des gains : cette option permet de sécuriser automatiquement les plus-values vers un support sécurisé lorsque le rendement des supports risqués dépasse un seuil de référence.
- Stop loss ou limitation des moins-values : elle permet d'arbitrer automatiquement vers des fonds moins risqués lorsque le rendement des supports risqués est en dessous d'un seuil de référence. Pour limiter les pertes, les capitaux restants sur les fonds risqués seront transférés vers un fond sécurisé.

Par la suite nous ne faisons pas de distinction pour la modélisation des différentes options d'arbitrage, cependant, il est important de distinguer les arbitrages selon la nature de leurs facteurs déclenchants pour pouvoir bien les expliquer. On distingue généralement les arbitrages qui ne dépendent pas de l'environnement économique mais plutôt des facteurs dits structurels, qu'on appelle arbitrages structurels, de ceux qui en dépendent et qu'on appelle alors arbitrages conjoncturels ou dynamiques.

### 1.2.1.1 Les arbitrages structurels

Selon la directive de l'ACPR<sup>1</sup> dans l'ONC<sup>2</sup> 2013, les lois comportementales structurelles des assurés, et notamment les arbitrages structurels, doivent être modélisées sur la base des lois d'expérience pour vérifier la conformité avec les réalisations passées ou à défaut avec les

---

1. Autorité de Contrôle Prudentiel et de Résolution  
2. Orientations Nationales Complémentaires

données du marché. Les arbitrages structurels dépendent des paramètres dits structurels et correspondent parfois à un besoin de sécurité ou de gain (par exemple, la préparation d'une retraite auquel cas l'assuré préfère orienter ses placements vers des supports moins risqués) en dehors de l'influence de la conjoncture économique.

Parmi les facteurs structurels, on peut en citer deux grands types : ceux qui sont liés à au contrat (l'ancienneté du contrat, l'encours, la périodicité des primes, le type de gestion, le profil d'investissement, la part des unités de compte dans le contrat. . .) et ceux qui sont liés à l'assuré (âge, sexe, situation matrimoniale. . .).

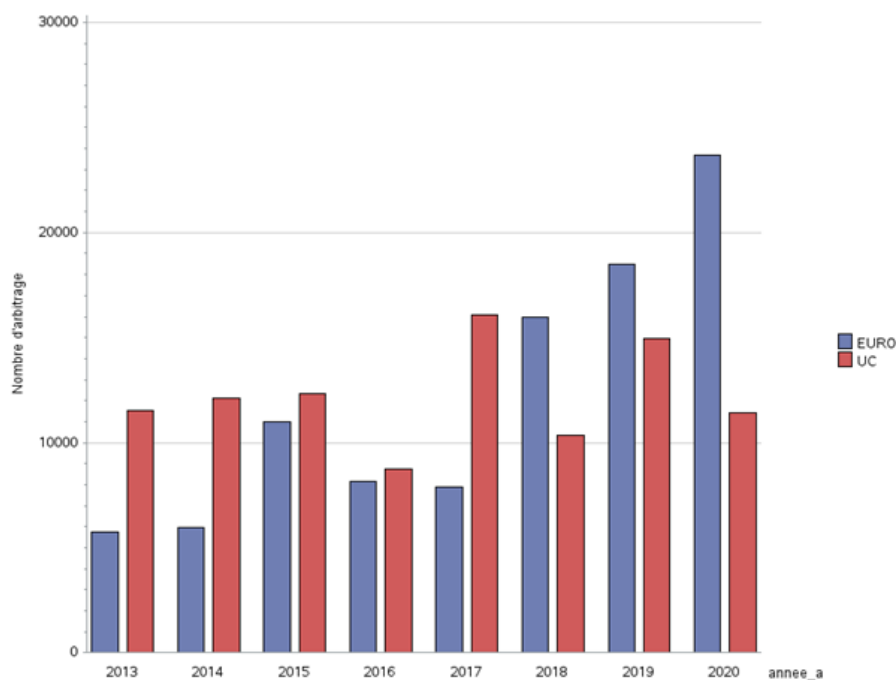
### 1.2.1.2 Les arbitrages conjoncturels

Les arbitrages dynamiques ou conjoncturels en revanche dépendent de l'environnement économique dans lequel se trouve l'assuré. Intuitivement, un assuré va chercher à comparer le rendement de son contrat et le rendement des supports en UC ou en euros. Ainsi, selon le principe de rationalité de l'assuré, si ce dernier constate que le rendement des supports en UC ou en euros est plus avantageux alors il réorientera son contrat afin de le placer sur les supports performants. Ces arbitrages viennent s'ajouter avec les arbitrages structurels.

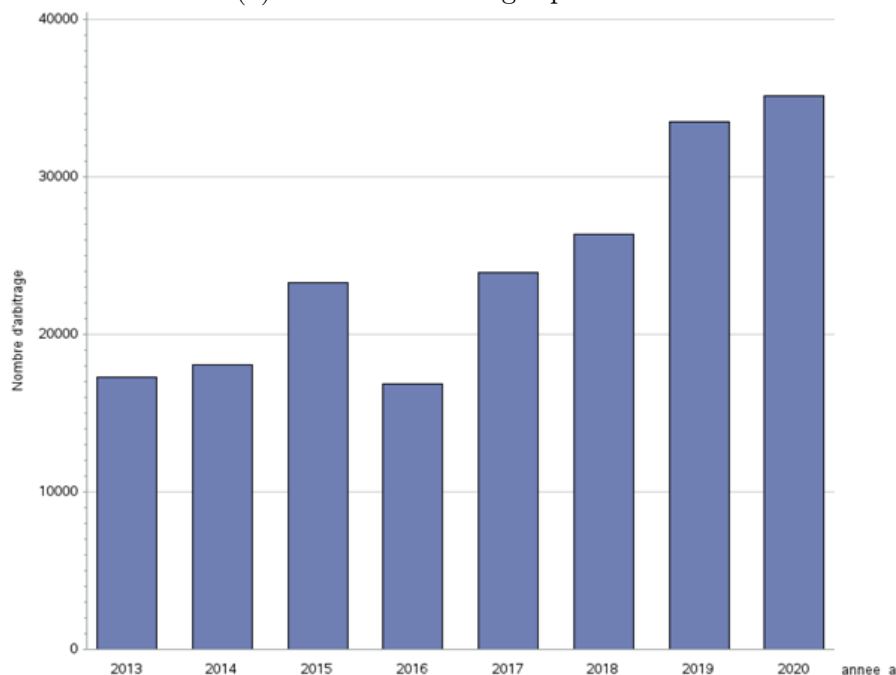
Toutefois, il y a d'autres types de facteurs conjoncturels en dehors de l'écart entre le rendement du contrat et le rendement des supports UC ou euros comme : le contexte économique et financier (l'évolution des taux d'intérêt, de chômage, de la croissance. . .), le changement de législation (avantage fiscale des arbitrages) . . . Pour la suite, nous ne retiendrons que les variables financières pour la modélisation des arbitrages conjoncturels.

## 1.2.2 État des lieux des arbitrages

Dans cette partie, les observations ont été agrégées par année, c'est-à-dire que l'ensemble des arbitrages effectués dans un contrat au cours de l'année est considérée comme un seul mouvement d'arbitrage dans l'analyse. Une analyse à un pas de temps mensuel des arbitrages aurait éventuellement apportée donné une meilleure compréhension et aurait fourni un meilleur modèle, mais la volumétrie des données et les observations qui ont été à notre disposition ne nous le permettent pas. Nous ne ferons pas de distinction entre le sens des arbitrages dans notre analyse, c'est-à-dire un arbitrage euro vers UC sera considéré de la même façon qu'un arbitrage UC vers euro du point de vue de la décision d'arbitrage. La distinction des deux arbitrages sera prise en compte en revanche lors de l'analyse des taux d'arbitrages.



(a) Nombre d'arbitrages par fonds



(b) Nombre d'arbitrages par année

FIGURE 1.1 – Nombre d'arbitrages entre 2013 et 2020

Entre 2013 à 2017, nous remarquons que le nombre d'arbitrages UC vers euro est toujours supérieur au nombre d'arbitrages euro vers UC et cette tendance s'est inversée à partir de l'année 2018. Les assurés ont alors commencé à s'orienter vers le fonds en UC en termes de nombre. Ce qui justifierait la nécessité de la prise en compte de sa modélisation pour connaître d'un côté la préférence des clients et d'un autre côté pour mieux appréhender les capitaux à immobiliser. Ces graphiques mettent aussi l'accent sur le fait que les clients sont très réactifs par rapport au rendement et à la nature de leur support.

Ce graphique illustre aussi l'effet du contexte du taux bas sur le choix des fonds d'investissement par les assurés dans leur contrat d'assurance vie. En effet, à partir de 2017, nous observons de plus en plus des pics d'arbitrage en faveur des fonds vers UC malgré le contexte actuel de pandémie qui a eu des impacts sur le marché financier. Cependant, un grand nombre d'arbitrages des fonds en UC vers des fonds en euros ou l'inverse ne signifie pas forcément qu'en net, le montant va suivre la même tendance d'où la nécessité de compléter l'analyse des nombres par l'analyse des montants ou du taux d'arbitrage.

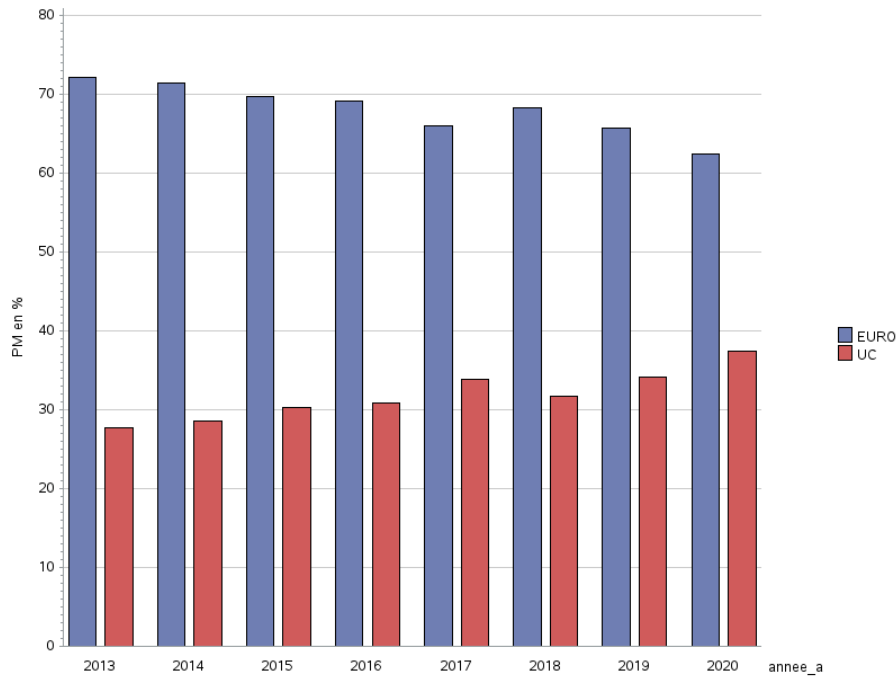
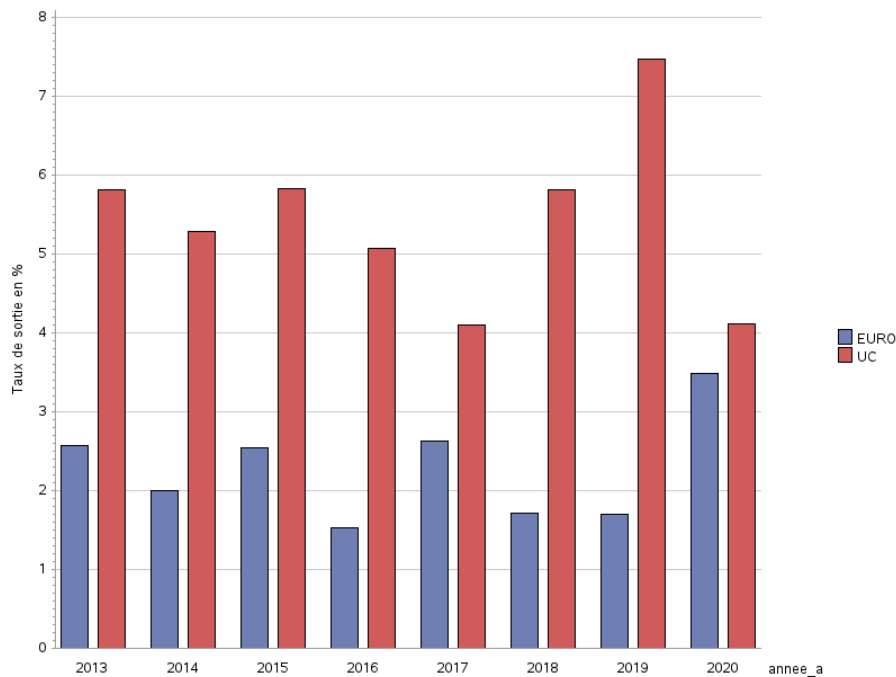
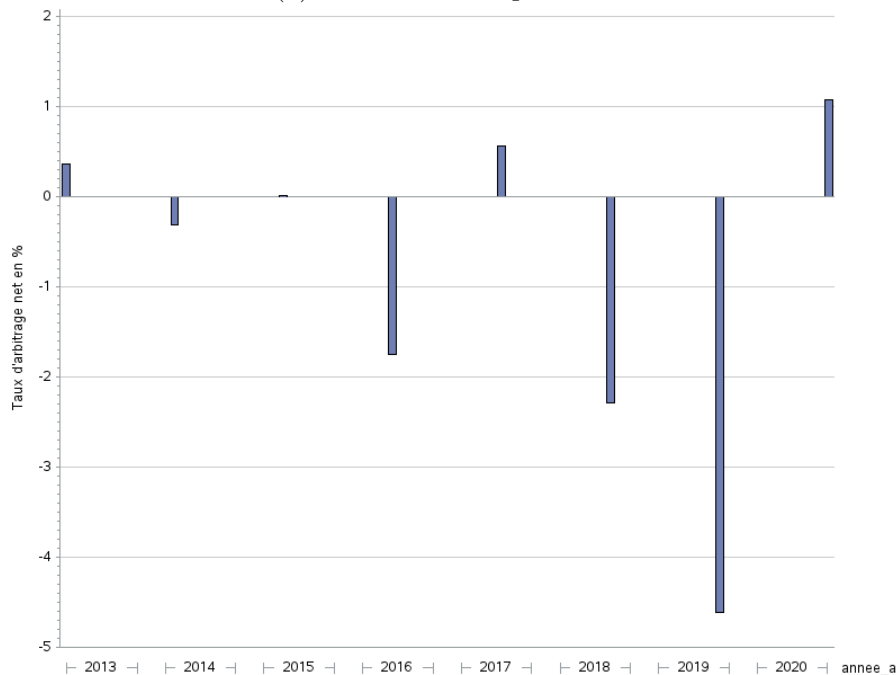


FIGURE 1.2 – Répartition de la PM



(a) Taux de sortie par fonds



(b) Taux d'arbitrage net par année

FIGURE 1.3 – Taux d'arbitrages entre 2013 et 2020

La figure 1.2 montre l'évolution de la structure de la provision mathématique (PM) dans le portefeuille de GENERALI. Depuis 2013, nous observons une tendance vers la hausse de la part en UC qui peut être expliquée par l'environnement de taux bas. Cependant, les assurés présentent toujours une aversion au risque, car entre 2013 et 2020, plus de 60 % de leur capital est placé sur le fonds en euros.

Entre 2013 et 2015, nous remarquons une stabilité du taux d'arbitrage net, c'est-à-dire un faible montant d'arbitrage en net par rapport au montant de la provision mathématique. À

partir de 2016, un pic de taux d'arbitrage net en faveur du fonds en euros apparaît. En effet, sous l'effet de la chute du cours du pétrole et de la crise du secteur bancaire italien en 2016, la bourse parisienne, et notamment les actions sur lesquelles les fonds en UC sont adossées ont enregistré des pertes au cours des deux premiers mois de l'année ce qui peut expliquer ce pic d'arbitrage. L'année 2019 a été marquée par une performance annuelle record de 26.37 % de l'indice CAC 40 et le feuilleton d'une possible dégradation des relations commerciales entre la Chine et les Etats-Unis ce qui peut traduire une partie du comportement de sécurisation des gains des assurés vers le fonds en euros.

### 1.2.3 Le risque d'arbitrage

Avantageuses pour l'assuré, les opérations d'arbitrage présentent des risques pour l'assureur. En effet, le fait de changer la nature des supports peut avoir des impacts ou changer les options et garanties offertes sur un contrat d'assurance vie. Cela peut aussi avoir des impacts sur l'évaluation de la provision (engagement) et sur la solvabilité de l'assureur. De plus, en tant qu'éléments du passif du bilan économique de la compagnie d'assurance, le Best Estimate et la solvabilité (et le comportement d'arbitrage) sont sensibles à la variation des taux d'intérêt et des indices financiers. Il est donc important de bien cerner et de bien distinguer les arbitrages qui découlent de la conjoncture économique de ceux qui n'en dépendent pas.

Le contexte récent de l'environnement de taux bas et la venue d'une crise sanitaire mettent en difficulté les compagnies d'assurance vie, d'une part en dégradant le rendement des obligations, ce qui a des conséquences négatives sur la solvabilité des assureurs-vie et d'autre part en créant une crise sur le marché financier, qui à son tour va augmenter le risque et la volatilité du rendement des fonds en unités de compte. D'un côté, l'incertitude sur les marchés financiers peut inciter les assurés à orienter leurs placements vers les fonds en euros qui obligeront les assureurs à investir sur des actifs à faible rendement, ce qui diluerait le rendement global du fonds en euros. L'assureur est contraint d'acquérir des nouvelles obligations sur le marché pour investir les nouveaux placements entrants sur le fonds en euros. Le rendement de ces nouvelles obligations est inférieur à ceux des anciennes obligations détenues par l'assureur pour les fonds en euros, et cette hausse d'arbitrage UC vers euros va donc diminuer le rendement global du fonds en euros. Or si le rendement du fonds en euros baisse, il serait difficile pour l'assureur de rémunérer les contrats d'assurance vie placés avec des taux garantis élevés qui constituent donc un coût en fonds propres importants et dégradent sa solvabilité. En plus du faible rendement, une hausse de l'arbitrage du fonds UC vers le fonds en euros expose le fonds en euros au risque de moins-values en cas de remontée de taux ou une éventuelle hausse des arbitrages euro vers UC car ces obligations sont acquises à des cours élevés (relation inverse entre taux de rendement et prix de l'obligation).

D'un autre côté, la baisse du rendement des fonds en euros peut inciter les assurés à se



tourner vers d'autres supports. Or en cas de retraits massifs ou de changement de support, la compagnie sera obligée de vendre ses actifs peut-être à des moments inopportuns (réalisation des moins-values) et peut se retrouver dans l'incapacité de faire face à ses engagements, ce qui entraînera ensuite un risque de faillite. Après une longue période de taux bas, l'assureur détiendra une part d'obligations à faible rendement et acquise à un prix élevé, donc si le taux sur le marché obligataire monte alors le cours des anciennes obligations qui ont un taux de rendement inférieur au taux de marché baisse. Ces obligations se retrouvent en situation de moins-value latente et comme elles sont aussi majoritaires dans le portefeuille de l'assureur alors le rendement global des fonds en euros sera moins intéressant que le rendement des fonds UC en obligation par exemple. Cela peut augmenter l'arbitrage euros vers UC. Une augmentation de l'arbitrage euros vers UC peut donc obliger l'assureur à vendre ses actifs (obligations en majorité) en situation de moins-value latente. Et si ces moins-values prennent de l'ampleur, l'assureur pourrait alors être en risque sur son bilan. Cependant, cette hausse de l'arbitrage euro vers UC permet en même temps de diminuer les fonds propres immobilisés, car les contrats ne seront plus garantis par l'assureur comme pour le cas des fonds en euros.

Par ailleurs, les arbitrages euros vers UC ou UC vers euros qui sont dus à des facteurs autres que la conjoncture économique peuvent avoir les mêmes impacts sur le Best Estimate et la solvabilité de l'assureur que ceux qui en dépendent.

Comprendre les facteurs qui influencent l'orientation des placements et le choix des supports des assurés et modéliser leurs comportements en matière d'arbitrage s'avèrent donc être importants pour l'assureur afin de faire face à ses engagements et améliorer sa solvabilité. Cela revient au grand dilemme de la statistique : un compromis entre la précision et l'interprétabilité du modèle.

L'assureur calcule le montant de son engagement envers l'assuré via un outil de gestion d'actif/passif. La projection des flux dans cet outil se fait par groupe de contrats. Deux approches de modélisation seront considérées pour modéliser le comportement d'arbitrage des assurés : une modélisation à la maille contrat qui permet d'avoir une meilleure compréhension des arbitrages et une meilleure précision à court terme (1 an), et une modélisation par groupe de contrats qui permet de réaliser une projection à long terme et une évaluation de l'impact du comportement d'arbitrage sur le Best Estimate.

Deux problématiques se dégagent alors dans cette étude à savoir :

- Est-il possible de construire un modèle qui permet de séparer les arbitrages dynamiques et les arbitrages structurels ?
- Les modèles de machine learning ou GLM peuvent-ils modéliser le comportement d'arbitrages des assurés ? Et sont-ils interprétables ?

# CADRE THÉORIQUE ET UNE NOUVELLE PERSPECTIVE DE MODÉLISATION DES ARBITRAGES

## 2.1 Revue de la littérature

Après la crise de 2008, la modélisation des arbitrages a suscité l'intérêt de beaucoup d'auteurs. En effet, la crise sur les marchés financiers et la chute du rendement en UC ont fait prendre conscience à l'assureur du coût et de l'impact des arbitrages sur sa solvabilité et sur la valeur de son engagement. Dans le domaine académique, il en ressort deux grandes approches pour modéliser le comportement d'arbitrage des assurés : l'approche microéconomique fondée sur la rationalité des assurés et l'approche statistique basée sur l'analyse des données.

### 2.1.1 Approche microéconomique

L'approche microéconomique est fondée sur la rationalité des assurés : elle tente de modéliser les préférences des assurés à travers la théorie de l'espérance d'utilité. En effet, dans le domaine de la microéconomie, la théorie de l'espérance d'utilité apparaît comme le fondement de la théorie de la décision individuelle en univers incertain et utilisée ici pour modéliser la décision d'arbitrage. Dans cette approche, l'assuré est perçu comme un agent économique rationnel qui cherche à maximiser sa fonction objectif qui n'est rien d'autre que son espérance d'utilité.

Karim Zennaf (2012) (11) dans son mémoire suppose que les assurés sont des agents rationnels qui disposent d'une richesse initiale et il cherche à déterminer la répartition optimale entre le fonds en euros et le fonds en UC que doit détenir les assurés jusqu'à l'échéance de leur contrat d'assurance vie. Il considère que l'espérance utilité est un système dynamique discret dont l'état à chaque instant  $t$  est fonction de  $\alpha$  qui est la proportion en unités de compte dans le contrat d'assurance vie avec un horizon d'investissement  $T$ . Il fait l'hypothèse que les agents sont averses au risque et ont une fonction d'utilité de type CRRA (Constant Relative Risk Aversion) et que les assurés peuvent arbitrer sans frais. L'objectif de l'assuré à un instant  $t$  quelconque sera donc de résoudre le problème suivant :

$$V(t, C_t) = \sup_{(\alpha_s)_{t \leq s \leq T}} E(U(C_T^\alpha) | X_t = C_t)$$

avec  $U$  la fonction d'utilité de l'assuré :

$$U(x) = x^{1-\gamma}$$

Le paramètre d'aversion au risque,  $\gamma$ , étant négatif pour traduire le fait que les assurés sont averses au risque,  $(\alpha_s)_{t \leq s \leq T}$  la part en UC dans le contrat, un processus à valeurs dans  $[0, 1]$  et  $C_t^\alpha$  le capital à l'instant  $t$  issu de la stratégie  $\alpha$ .

La fonction admet un maximum en :

$$\alpha^* = \begin{cases} 0 & \text{si } \frac{\mu-r}{\gamma\sigma^2} < 0 \\ \frac{\mu-r}{\gamma\sigma^2} & \text{si } \frac{\mu-r}{\gamma\sigma^2} \in [0, 1] \\ 1 & \text{si } \frac{\mu-r}{\gamma\sigma^2} > 1 \end{cases}$$

avec  $r$  le taux sans risque,  $\mu$  l'espérance de rendement de l'actif risqué et  $\sigma$  sa volatilité. En effet, l'auteur suppose que l'actif risqué suit une dynamique de Black-Scholes.

Il obtient une valeur de  $\alpha$  qui dépend du signe de l'écart entre le rendement du fonds en euros et du fonds en UC. Les principales limites de son travail sont les suivantes : il suppose que tous les assurés sont rationnels, ce qui n'est pas vrai dans la réalité ; il avance également que la solution du problème favorise le placement sur le fonds en euros, c'est-à-dire que les assurés arbitrent toute la totalité de leur capital vers le fonds en euros lorsqu'ils observent un écart de rendement en défaveur du fonds en UC, ce qui est loin d'être observé dans la pratique.

Khaoula Lyoubi (2020)(12) propose plutôt un modèle de choix multiple dans son mémoire pour modéliser les arbitrages. Elle adopte un modèle de choix probabiliste pour dépasser le cadre de la maximisation de l'espérance d'utilité. Elle résume le problème de l'assuré à un problème de choix d'une loterie dans un ensemble de loteries possibles finies. La loterie représente le poids du fonds en euros dans le portefeuille (la valeur est comprise entre 0 et 1) dans un ensemble fini. L'assuré va chercher à comparer une à une toutes les alternatives possibles de manière itérative pour construire la probabilité d'être la meilleure alternative de chaque loterie en utilisant la probabilité  $P(A_i, A_j)$  qui se définit comme la probabilité qu'un décideur choisisse l'alternative  $A_i$  au lieu de l'alternative  $A_j$  dans un contexte de choix binaire. Notons que la probabilité  $P(A_i, A_j)$  dépend d'une fonction discriminante. À l'initialisation du problème, chaque loterie a la même chance d'être choisie comme la meilleure alternative.

$$P(X, Y) = P(X \succ Y) = \frac{\phi(X')}{\phi(X') + \phi(Y')}$$

avec :

$$\begin{aligned} X' &= EU(X) - EU(X \wedge Y)(x)^1 \\ Y' &= EU(Y) - EU(X \wedge Y)(x) \end{aligned}$$

et  $\phi$  une fonction discriminante choisie par l'auteur.

### 2.1.2 Approche statistique

Il existe généralement deux types d'approche statistique pour modéliser le comportement d'arbitrage des assurés. La première consiste à utiliser une approche individuelle pour expliquer et modéliser la décision d'arbitrage en utilisant des modèles probabilistes et le taux d'arbitrage comme dans un modèle fréquence sévérité en assurance non-vie. La deuxième approche passe par la création des groupes homogènes d'assurés afin de modéliser le taux d'arbitrage de chaque groupe en utilisant soit la régression linéaire, soit le modèle linéaire généralisé (GLM) ou encore les modèles de machine learning.

Omar Hamaoui (2012) (16) utilise la régression linéaire pour modéliser le taux d'arbitrage euro vers UC et le taux d'arbitrage UC vers euro séparément en fonction des variables liées à des caractéristiques du contrat et des variables de la conjoncture économique. Une analyse du résidu de la régression lui a permis de montrer que les variables dépendantes devraient subir une transformation de Box-Cox<sup>2</sup> avant d'être utilisées dans la régression pour aplatir les variances des résidus. Ensuite, il fait l'hypothèse selon laquelle les assurés se basent sur le rendement du CAC40 et du TME pour effectuer leurs arbitrages et propose une modélisation par des séries temporelles en utilisant les fonctions de transfert de la chronique de taux d'arbitrage net UC vers euro en fonction d'une chronique d'entrée  $x_t$ . La chronique  $x_t$  à son tour dépend de l'écart entre le rendement du CAC40 et le TME.

$$y_t = \mu + v(B)x_{t-b} + \eta_t$$

où  $y_t$  le taux d'arbitrage,  $v(B)$  est un polynôme symbolique en  $B$  :  $v(B) = v_0 + v_1B + \dots + v_kB^b$ , et  $B$  correspond à l'opérateur retard défini par :  $BX_t = X_{t-1}$

$$x_t = 1 - \arctan(\alpha + \beta \times (\text{cac}_t - \text{tme}_t))$$

Guillaume Arquembourg (2014) (9) propose une approche GLM (régression logistique) pour modéliser le taux d'arbitrage euro vers UC, il traite les arbitrages structurels en ne soumettant que les caractéristiques des contrats comme variables explicatives au modèle. Pour capter les effets conjoncturels, il introduit la variable année avec les caractéristiques des contrats comme

---

1.  $L \wedge L'$  désigne la plus grande limite inférieure sur les loteries  $L$  et  $L'$  en termes de dominance stochastique. Il n'existe donc pas de loterie qui puisse dominer stochastiquement  $L \wedge L'$  et qui est dominée par  $L$  et  $L'$  (12)

2.  $\frac{x^\lambda - 1}{\lambda}$

variables explicatives de son modèle. Par la suite, il établit une relation entre le ratio du taux d'arbitrage euro/UC de l'année en intégrant la variable année sur le taux d'arbitrage euro/UC de l'année sans tenir compte de la variable année et la valeur d'un indice financier.

## 2.2 Une nouvelle perspective de modélisation des arbitrages

L'objectif de cette partie est de proposer un modèle qui permet de prédire séparément et d'identifier les facteurs explicatifs des arbitrages structurels et des arbitrages dynamiques. Dans le cadre de cette étude, la variable d'intérêt à modéliser sera le taux d'arbitrage net euro vers UC, c'est-à-dire que si le taux est positif pour un contrat donné alors il y a une sortie d'argent du fonds en euros vers le fonds en UC. À l'inverse, un taux d'arbitrage net négatif signifie qu'il y a un transfert de capitaux du fonds UC vers le fonds en euros. Par la suite, lorsqu'on parle de taux d'arbitrage, on fait allusion au taux d'arbitrage net euro vers UC.

Soit  $Y$  la variable aléatoire correspondant au taux d'arbitrage de l'assuré, alors  $Y$  peut s'écrire comme :  $Y = Y_s + Y_c$  où  $Y_s$  est le taux d'arbitrage structurel et  $Y_c$  le taux d'arbitrage conjoncturel. Soient  $X_s$  (âge, sexe, ancienneté...) l'ensemble des facteurs structurels et  $X_c$  (rendement du fonds en euros...) l'ensemble des facteurs liés à l'environnement économique qui expliquent les arbitrages. Nous supposons que les facteurs  $X_s$  sont indépendants des facteurs  $X_c$ . Le modèle qui sera mis en place s'inspire de l'approche fréquence/sévérité souvent utilisée en assurance non vie. Nous postulons une hypothèse forte d'indépendance temporelle du comportement d'arbitrage des assurés. Soient  $N$  la variable aléatoire représentant le nombre d'arbitrages dans le portefeuille de l'assureur,  $S$  le montant total d'arbitrage,  $B_i$  le montant d'arbitrage de l'assuré  $i$  et  $PM_i$  le montant de provision mathématique annuelle de l'assuré  $i$  précédant l'année d'arbitrage. On suppose que  $N$  est indépendant de  $B_i$  pour tout  $i$ . Le montant total d'arbitrage s'écrit comme suit :

$$S = \sum_{i=1}^N B_i$$

Avec  $B_i = Y_i * PM_i$ . Comme les  $PM_i$  sont déterministes alors modéliser les  $B_i$  revient donc à modéliser les  $Y_i$ .

On cherche donc un modèle de fréquence (GLM ou modèle de machine learning) pour modéliser le nombre d'arbitrages et un modèle de sévérité qui sera détaillé par la suite pour modéliser le taux ou le montant d'arbitrage à partir des variables explicatives.

### 2.2.1 Régression en deux étapes

L'hypothèse adoptée pour cette approche est que le taux d'arbitrage s'écrit comme une combinaison linéaire de ses facteurs explicatifs. La régression en deux étapes est souvent utilisée

pour résoudre le problème d'endogénéité en économétrie en utilisant des variables instrumentales. Mais à travers des justifications théoriques qui seront détaillés par la suite, il s'avère que cette méthode permet de séparer la part du modèle expliquée par deux variables explicatives non corrélées.

La méthode de régression en deux étapes s'opère de la manière suivante. À la première étape, on régresse la variable  $Y$  par toutes les variables explicatives structurelles pour obtenir la part du modèle expliquée par les facteurs structurels. Dans la seconde étape, on récupère les résidus de la première régression et on régresse ces résidus sur les variables conjoncturels pour prédire la part du modèle expliquée par les variables de la conjoncture économique. Si  $Y$  s'écrit comme une combinaison linéaire de ses facteurs explicatifs alors on a :

$$\begin{aligned} Y &= Y_s + Y_c \\ &= \phi(X_s) + \psi(X_c) \end{aligned}$$

avec  $Y_s = \phi(X_s) = \beta X_s$  et  $Y_c = \psi(X_c) = \delta X_c$

Première étape : Estimation des paramètres associés aux arbitrages structurels

$$Y = \alpha + \beta X_s + \epsilon \quad \text{avec} \quad \text{cov}(X_s, \epsilon) = 0$$

Deuxième étape : Estimation des paramètres associés aux arbitrages conjoncturels

$$\epsilon = \gamma + \delta X_c + u$$

Finalement on a :

$$\begin{aligned} Y &= \alpha + \beta X_s + \epsilon \\ &= \alpha + \beta X_s + (\gamma + \delta X_c + u) \\ &= (\alpha + \gamma) + \beta X_s + \delta X_c + u \end{aligned}$$

Dans le cas d'une régression du taux d'arbitrage sur toute les variables explicatives, comme  $\text{cov}(X_s, X_c) = 0$  (par hypothèse) alors on retrouve les mêmes coefficients pour les variables sauf qu'on ne saura pas distinguer la part de constante du modèle associée aux arbitrages structurels et conjoncturels.

$$Y = \zeta + \beta X_s + \delta X_c + u \quad \text{avec} \quad \zeta = \alpha + \gamma \tag{2.1}$$

## 2.2.2 Une approche par l'espérance conditionnelle

L'espérance conditionnelle<sup>3</sup> de  $Y$  sachant  $X$  est la fonction de  $X$  donnant la meilleure approximation de  $Y$  quand  $X$  est connu. En effet, en statistique, pour résoudre les problèmes de prédiction, il est important de pouvoir prédire une variable aléatoire sur laquelle nous ne disposons que d'une information partielle. Ce qui justifie l'importance de l'utilisation de l'espérance conditionnelle. La difficulté rencontrée dans l'utilisation de l'espérance conditionnelle pour résoudre un problème de prédiction est la détermination de la famille de fonction dans laquelle appartient l'espérance conditionnelle. En effet, dans la pratique, nous ne connaissons pas la loi jointe entre la variable cible et les variables explicatives d'où la nécessité d'approcher l'espérance conditionnelle par une famille de fonction (linéaire par exemple). Les méthodes les plus utilisées pour effectuer l'approximation sont les GLM et les modèles de machine learning.

Dans notre cas, étant donné que nous ne disposons que de l'information sur l'assuré, le contrat et la conjoncture économique précédant l'année d'arbitrage, on cherche alors à approcher le taux d'arbitrage à partir de ces variables. A priori, nous ne connaissons pas la loi de probabilité jointe entre le taux d'arbitrage et les variables explicatives. Il s'avère donc important d'approcher l'espérance conditionnelle du taux d'arbitrage sachant les informations précédant l'année d'arbitrage pour prédire sa valeur.

Par la suite, nous supposons que le taux d'arbitrage conjoncturel ne dépend pas des variables structurelles et que son espérance est nulle. En général, les assurés adoptent le même comportement suite à un changement de la conjoncture économique (rendement des fonds). De plus, ils font des arbitrages soit pour respecter leur profil de risque, par exemple garder 60% de support en UC dans le contrat tout au long de l'année soit pour suivre un plan d'épargne en fonction de l'âge c'est-à-dire que le capital est majoritairement investi sur des fonds risqués au début et s'oriente vers les fonds moins risqués au fil des années. Les assurés ne font alors des arbitrages dynamiques que dans des situations extrêmes, c'est-à-dire lorsque l'écart entre le rendement des deux fonds est très significatif (en cas de crise financière par exemple ou une surperformance des fonds en UC). De plus, nous supposons que le mouvement d'arbitrage dynamique est symétrique dans le sens où l'assuré va réagir de la même façon s'il constate un écart de rendement en faveur du fonds en euro ou en faveur du fonds en UC, par exemple s'il y a un écart de 10% du rendement en faveur du fonds en euro alors le client va arbitrer  $x$  % de la PM de son fonds en UC. De même s'il y a un écart de rendement de 10% en faveur du fonds en UC alors le même assuré va arbitrer  $x$  % de la PM de son fonds en euros, donc nous

---

3. Soit  $(X, Y)$  un couple aléatoire, avec  $Y$  intégrable. L'espérance conditionnelle de  $Y$  sachant  $X$  est l'unique variable aléatoire fonction de  $X$ , notée  $\mathbb{E}[Y | X]$ , telle que pour toute fonction bornée  $u : \mathbb{R} \rightarrow \mathbb{R}$ , on ait :

$$\mathbb{E}[u(X)Y] = \mathbb{E}[u(X)\mathbb{E}[Y | X]]$$

Ainsi il existe une fonction  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  mesurable telle que  $\mathbb{E}[Y | X] = \varphi(X)$ .

pouvons supposer que l'arbitrage conjoncturel ne dépend pas des facteurs structurels et son espérance est nulle. Notons que ces hypothèses sont très fortes mais s'avèrent être nécessaires pour la suite.

$$\begin{aligned}
 Y &= Y_s + Y_c \\
 \mathbb{E}(Y \mid X_s) &= \mathbb{E}(Y_s \mid X_s) + \mathbb{E}(Y_c \mid X_s) \\
 &= \mathbb{E}(Y_s \mid X_s) + \mathbb{E}(Y_c) \quad \text{car } Y_c \perp\!\!\!\perp X_s \\
 &= \mathbb{E}(Y_s \mid X_s) \quad \text{car } \mathbb{E}(Y_c) = 0
 \end{aligned}$$

Pour estimer les paramètres associés aux taux d'arbitrage structurel, cette espérance conditionnelle sera donc approchée par un modèle GLM ou un modèle de machine learning (XGboost par exemple). Les résidus de ce modèle correspondent donc aux taux d'arbitrage conjoncturel et seront approchés par un modèle GLM ou un modèle de machine learning à partir des variables conjoncturelles.

---

**Algorithm 1** Modélisation du taux d'arbitrage

---

- Soit  $x$  variable explicative,  $N$  nombre d'année d'entraînement et  $M$  nombre d'année à prédire
- Estimer la valeur du taux d'arbitrage structurel  $Y_s$  en approchant par un modèle GLM et d'autres modèles de machine learning sur les données de l'année  $1, \dots, N$
- Calculer les résidus du modèle  $\varepsilon = Y - \hat{\mathbb{E}}(Y \mid X_s)$ ,  $\hat{\mathbb{E}}(\cdot)$  correspond à la valeur estimée d'une variable par un modèle
- Approcher l'espérance conditionnelle de ce résidu  $\mathbb{E}(\varepsilon \mid X_c)$  par un GLM ou d'autres méthodes de machine learning (taux d'arbitrage conjoncturel)

**Pour**  $m = N + 1, \dots, N + M$  **faire**

$$\begin{aligned}
 \hat{f}(x) &= \hat{Y} = \hat{\mathbb{E}}(Y \mid X_s) + \hat{\mathbb{E}}(\varepsilon \mid X_c) \text{ pour chaque individu} \\
 \hat{B}_i &= \hat{f}_i(x) * PM_i ; 1, \dots, n \text{ avec } B_i : \text{Montant d'arbitrage} \\
 \hat{Y}_m &= \frac{\sum_{i=1}^{n_m} B_i^m}{\sum_{i=1}^{n_m} PM_i^m}
 \end{aligned}$$

**FinPour**

- Résultat : une série de taux d'arbitrage  $Y_{N+1}, \dots, Y_{N+M}$
-



# LES MODÈLES DE MACHINE LEARNING

Ce chapitre rappelle les principes du modèle GLM, de l'arbre de régression et du modèle XGBoost. Ensuite, il est question de présenter les outils nécessaires pour interpréter ces modèles et quantifier les incertitudes autour de leur prédiction.

En général, l'objectif de la modélisation statistique est d'expliquer une variable cible  $y$  par des variables explicatives  $x_1, x_2, \dots, x_k$ . Cette représentation permet à la fois d'ajuster un modèle pour expliquer et pour prédire la variable cible. Nous allons donc chercher une fonction des facteurs explicatifs qui approche la variable dépendante :

$$y \approx f(x_1, x_2, \dots, x_k)$$

## 3.1 Le modèle de régression linéaire

La modélisation la plus utilisée pour approcher une variable continue est la régression linéaire : elle suppose que la variable à expliquer s'écrit comme une combinaison linéaire des variables explicatives, c'est-à-dire nous cherchons la fonction  $f$  dans la famille des fonctions affines  $\mathcal{F}$  de  $\mathbb{R}$  dans  $\mathbb{R}$ . Le modèle de régression linéaire est défini par l'équation suivante :

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

où  $\varepsilon$  est le terme d'erreur, il regroupe les informations de  $y$  qu'on n'arrive pas à capter avec les régresseurs  $x_1, \dots, x_k$ . Soient  $n$  la taille de l'échantillon,  $Y = (y_i)_{i=1, \dots, n}$ ,  $X_j = (x_{j,i})_{i=1, \dots, n}$ ,  $X = (1, X_1, X_2, \dots, X_k)$ ,  $\beta = (\beta_0, \dots, \beta_k)$  et  $\varepsilon = (\varepsilon_i)_{i=1, \dots, n}$ . Pour que le modèle soit valide et cohérent, il faut imposer des hypothèses sur le terme d'erreur :

$$(\mathcal{H}) \left\{ \begin{array}{l} (\mathcal{H}_1) : \mathbb{E}[\varepsilon_i] = 0 \text{ pour tout indice } i \\ (\mathcal{H}_2) : \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ pour } i \neq j \\ (\mathcal{H}_3) : \text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \text{ pour } i = j \end{array} \right.$$

Il est aussi possible d'adopter une représentation matricielle du modèle pour simplifier les notations :  $Y = X\beta + \varepsilon$ . L'objectif de la régression est de trouver la valeur des paramètres qui permet d'approcher le mieux  $y$ . Pour déterminer  $\hat{\beta}$ , on cherche donc l'hyperplan qui passe le plus près possible du nuage de points défini par l'échantillon. Ce qui revient à chercher le paramètre  $\beta$  qui minimise la somme des carrés des résidus. On obtient alors l'estimateur des

moindres carrés ordinaires qui s'écrit :

$$\hat{\beta}_{MCO} = \arg \min_{\beta_0, \dots, \beta_k} \sum_{i=1}^n \left( y_i - \left( \beta_0 + \sum_{j=1}^k \beta_j x_{j,i} \right) \right)^2$$

Lorsque la matrice  $X'X$  est inversible, la solution explicite de ce problème s'écrit :

$$\hat{\beta}_{MCO} = (X'X)^{-1}(X'Y)$$

Dans le cas de la régression linéaire, l'importance d'une variable  $x_j$  est définie par la t-statistique :  $t_{\beta_j} = \left| \frac{\beta_j}{\sigma_{\beta_j}} \right|$ , rapport entre le poids estimé et son écart-type. Ainsi plus le poids d'une variable est grand, plus la variable est importante.

Parfois, il est souhaitable d'exclure certaines variables explicatives dans le modèle pour éviter le sur-apprentissage et la non-convergence du modèle. En effet, lorsque deux variables explicatives sont parfaitement corrélées alors la matrice  $X'X$  n'est pas inversible, alors il n'est pas possible d'estimer la valeur des paramètres. Parfois, il s'avère être nécessaire de pénaliser la régression en introduisant la norme des vecteurs des paramètres dans la fonction objectif pour s'assurer de la convergence du modèle.

La régularisation Ridge ou de Tykhonov est un cas particulier de régularisation, dans lequel on utilise pour régulariser la régression linéaire le carré de la norme du vecteur des paramètres  $\beta$ . Plus précisément, il s'agit de la norme  $\|\cdot\|_2$ , ou norme euclidienne, c'est-à-dire :

$$\|\beta\|_2^2 = \sum_{j=1}^k \beta_j^2$$

L'estimateur  $\beta_{Ridge}$  est donc la solution de ce programme de minimisation :

$$\arg \min_{\beta_1, \dots, \beta_k} \sum_{i=1}^n \left( y_i - \left( \sum_{j=1}^k \beta_j z_{ij} \right) \right)^2 + \lambda \sum_{j=1}^k \beta_j^2$$

$z_j$  : forme centrée réduite de  $x_j$  (il est nécessaire de standardiser les variables en amont puisque la solution de ce programme n'est pas invariante à un facteur d'échelle).

La régression Ridge permet de réduire l'amplitude des coefficients d'une régression linéaire et d'éviter le sur-apprentissage. Cependant, il peut être souhaitable d'annuler certains coefficients, pour davantage de régularisation. Les variables qui auront un coefficient égal à zéro ne feront plus partie du modèle, qui en sera d'autant simplifié. Un tel modèle, avec beaucoup de coefficients nuls, est appelé un modèle parcimonieux (ou « sparse » en anglais).

Les variables  $X_i$  n'étant pas nécessairement toutes pertinentes pour expliquer  $y$ , l'objectif est d'éliminer les variables les moins pertinentes et uniquement celles-ci, il s'agit là de l'idée directrice de la régression Lasso. Pour arriver à cela, il suffit de remplacer le terme de régularisation de la régression Ridge, autrement dit la norme  $\|\beta\|_2^2$  de  $\beta$ , par la norme  $\|\beta\|_1$  de ce

vecteur, c'est-à-dire :

$$\|\beta\|_1 = \sum_{j=1}^k |\beta_j|$$

Le programme de la régression Lasso s'écrit alors :

$$\arg \min_{\beta_1, \dots, \beta_k} \sum_{i=1}^n \left( y_i - \left( \sum_{j=1}^k \beta_j z_{ij} \right) \right)^2 + \lambda \sum_{j=1}^k |\beta_j|$$

La solution de ce programme s'appelle le Lasso, pour (*Least Absolute Shrinkage and Selection Operator*). « Absolute » pour l'utilisation de la norme 1, « shrinkage » (réduction) parce qu'on contraint les coordonnées de  $\beta$  à avoir des valeurs faibles, et « selection » parce qu'on va tellement les réduire que certaines seront nulles. Contrairement à la régression Ridge, ce problème d'optimisation n'a pas de solution explicite mais il peut être optimisé par l'algorithme de descente du gradient.

Le graphique suivant illustre pour le cas de 2 variables explicatives, la forme du sous espace d'optimisation pour la régression Lasso et la régression Ridge.

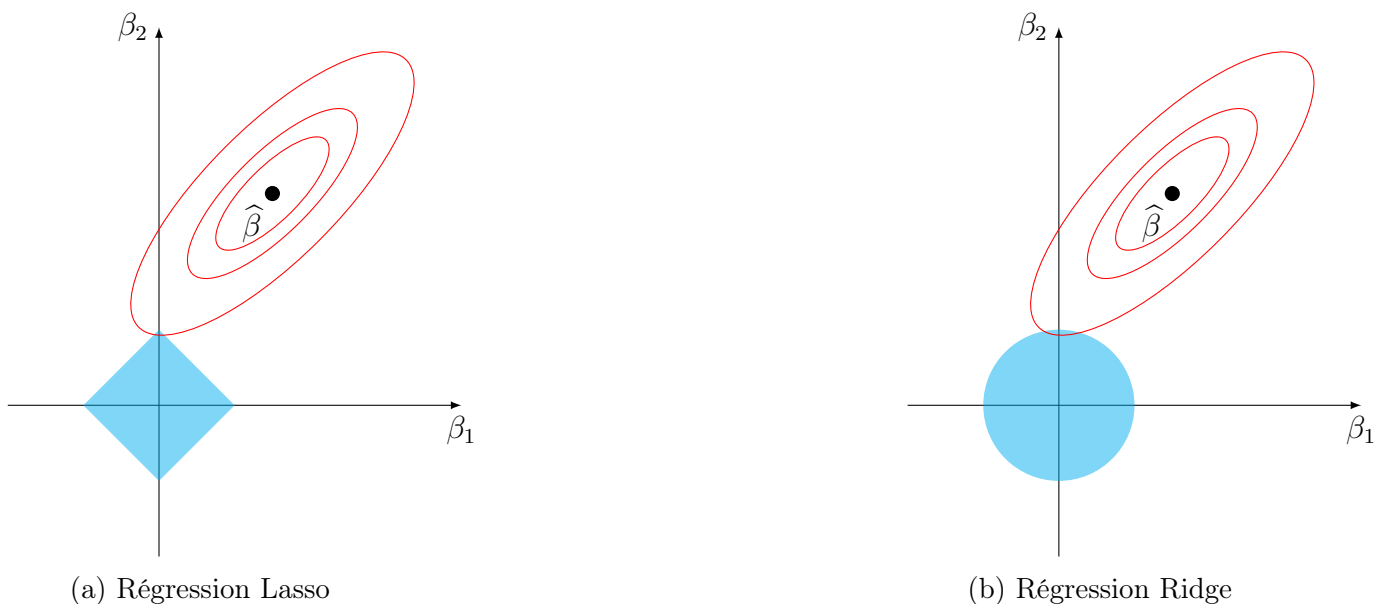


FIGURE 3.1 – Illustration de la restriction de l'espace d'optimisation par le Lasso et le Ridge

### 3.2 Le modèle GLM

Le modèle linéaire généralisé est une extension du modèle de régression linéaire multiple. Au lieu de modéliser directement la variable cible  $y$ , le GLM cherche plutôt à exprimer l'espérance conditionnelle de la variable réponse en fonction d'une combinaison linéaire des variables explicatives par le biais d'une fonction de lien. A l'inverse de la régression linéaire qui ne peut modéliser que les variables réelles continues dans  $\mathbb{R}$ , le GLM peut modéliser à la fois des va-

riables catégorielles et des variables continues. Nous avons donc la relation suivante :

$$g(\mathbb{E}[y | X]) = \beta_0 + \sum_{j=1}^p \beta_j x_j$$

où  $g$  est la fonction de lien, supposée monotone et différentiable.

Nous supposons aussi que la variable  $y$  admet une distribution issue de la famille exponentielle, c'est-à-dire sa fonction de densité s'écrit sous la forme :

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

où est  $\theta$  le paramètre naturel de la fonction exponentielle et  $\phi$  le paramètre de dispersion. En générale, le log vraisemblance d'un GLM s'écrit :

$$\ell(\theta) = \frac{y\theta - b(\theta)}{\phi} + c(Y; \phi)$$

La condition de premier ordre donne :

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} &= \frac{y - b'(\theta)}{\phi} = 0 \\ \mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right) &= \frac{\mathbb{E}(y) - b'(\theta)}{\phi} = 0 \end{aligned}$$

Donc on a :

$$\mathbb{E}(y) = b'(\theta)$$

Alors

$$\begin{aligned} \theta &= (b')^{-1}(y) \\ &= (b')^{-1}(g^{-1}(X'\beta)) \equiv h(X^{\top}\beta) \end{aligned}$$

où  $h$  s'écrit :

$$h = (b')^{-1} \circ g^{-1} = (g \circ b')^{-1}$$

Finalement, on a :

$$\ell_n(\beta; y, X) = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{\phi} \tag{3.1}$$

$$= \sum_{i=1}^n \frac{y_i h(X'_i \beta) - b(h(X'_i \beta))}{\phi} \tag{3.2}$$

### 3.2.1 La régression logistique

Dans le cas de la régression logistique,  $Y$  est une variable binaire  $\in \{0, 1\}$  et elle consiste à modéliser la probabilité conditionnelle  $\mathbb{P}(Y = 1 | X = x)$  comme fonction de la combinaison linéaire des variables explicatives. La valeur d'une probabilité devrait être comprise entre 0 et

1, d'où l'utilisation en générale de la fonction de lien logit.

$$g(\mathbb{P}(Y = 1 | X)) = \beta_0 + \sum_{i=1}^p \beta_i X_i$$

$$\text{logit}(\mathbb{P}(Y = 1 | X)) = \ln \left( \frac{\mathbb{P}(Y = 1 | X)}{1 - \mathbb{P}(Y = 1 | X)} \right) = \beta_0 + \sum_{i=1}^p \beta_i X_i$$

Étant donné qu' $Y$  est une variable binaire, le modèle prédit une probabilité et comme la régression logistique suppose que la loi de  $Y$  conditionnellement à  $X$  suit une loi de Bernoulli de paramètre  $p(x) = \mathbb{P}(Y = 1 | X = x)$ , alors la fonction de vraisemblance pour un échantillon de taille  $n$  s'écrit :

$$\mathcal{L}(\beta; y, X) = \prod_{i=1}^n \mathbb{P}(y_i = 1 | X_i = x^{(i)})^{y_i} (1 - \mathbb{P}[y_i = 1 | X_i = x^{(i)}])^{1-y_i}$$

$$\begin{aligned} \log(\mathcal{L}(\beta; y, X)) &= \ell_n(\beta; y, X) = \sum_{i=1}^n y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i)) \\ &= \sum_{i=1}^n \log(1 - p(x_i)) + \sum_{i=1}^n y_i \log \frac{p(x_i)}{1 - p(x_i)} \\ &= \sum_{i=1}^n \log(1 - p(x_i)) + \sum_{i=1}^n y_i \left( \beta_0 + \sum_{j=1}^p \beta_j x_j^{(i)} \right) \\ &= \sum_{i=1}^n -\log \left( 1 + e^{(\beta_0 + \sum_{j=1}^p \beta_j x_j^{(i)})} \right) + \sum_{i=1}^n y_i \left( \beta_0 + \sum_{j=1}^p \beta_j x_j^{(i)} \right) \end{aligned}$$

L'interprétation des résultats se fait généralement par odds ratio pour les variables catégorielles et à travers le paramètre directement pour les variables quantitatives. L'odds ratio se définit comme le rapport des côtes. La côte (au sens des jeux de paris) d'un évènement se définit comme le rapport de la probabilité de l'évènement à celle de son complémentaire. La côte pour une variable binaire  $x_1$  de modalité  $x_1 = 1$  vaut :

$$\frac{\mathbb{P}(y = 1 | x_1 = 1, \tilde{x}_{(p-1)})}{\mathbb{P}(y = 0 | x_1 = 1, \tilde{x}_{(p-1)})} = e^{(\beta_0 + \beta_1 + \beta_{p-1} \tilde{x}_{(p-1)})}$$

avec  $\tilde{x}_{(p-1)}$  une valeur quelconque fixée des autres variables explicatives. Donc, l'odds ratio de la variable  $x_1$  s'écrit :

$$\frac{\frac{\mathbb{P}(y=1|x_1=1,\tilde{x}_{(p-1)})}{\mathbb{P}(y=0|x_1=1,\tilde{x}_{(p-1)})}}{\frac{\mathbb{P}(y=1|x_1=0,\tilde{x}_{(p-1)})}{\mathbb{P}(y=0|x_1=0,\tilde{x}_{(p-1)})}} = e^{\beta_1}$$

Un individu qui à la modalité  $x_1 = 1$  a  $e^{\beta_1}$  plus de risque (côte) de subir ou réaliser l'évènement  $y = 1$  qu'un individu qui a la modalité  $x_1 = 0$ . La valeur de référence pour l'odds ratio est égale à 1, une valeur égale à 1 signifie que les deux modalités ont le même effet sur la variable réponse. Pour une variable continue, elle s'interprète comme l'impact sur la cote de la variable

expliquée d'une augmentation de 1 unité de la variable.

### 3.2.2 La régression tanh

Vu la distribution voir graphique 4.3 et la spécificité de la variable cible taux d'arbitrage, une adaptation du modèle GLM s'avère nécessaire pour bien modéliser le taux d'arbitrage. Le taux d'arbitrage a la particularité d'évoluer entre -1 et 1 et possède une distribution quasi symétrique avec des queues lourdes. Donc a priori, aucun modèle GLM classique ne peut modéliser notre variable cible. Cependant, comme la logit, la fonction tangente hyperbolique est souvent utilisée en machine learning comme fonction d'activation d'un réseau de neurones. Elle a l'avantage d'être inversible, monotone et bornée entre -1 et 1. Son inverse, l'arc tangente hyperbolique peut être donc utilisée comme une fonction de lien pour modéliser le taux d'arbitrage. La forme symétrique de la distribution nous laisse plutôt penser que sa distribution appartient à la famille de la loi normale. Un GLM avec une famille de distribution normale et une fonction de lien  $\operatorname{arctanh}(x)$  sera donc le modèle GLM utilisé pour modéliser le taux d'arbitrage que nous appelons « **la régression tanh** ».

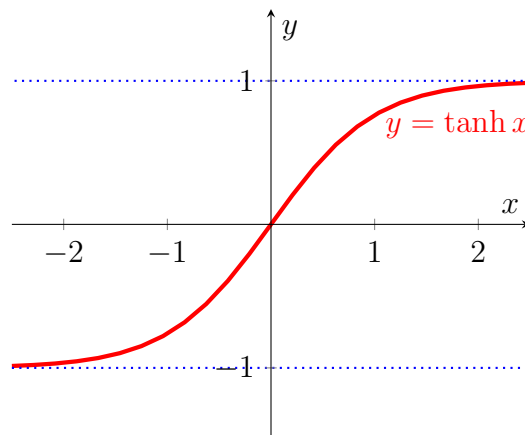


FIGURE 3.2 – La fonction tanh

Dans ce cas, nous avons la relation suivante :

$$\operatorname{arctanh}(\mathbb{E}[y | X]) = \beta_0 + \sum_{j=1}^p \beta_j x_j$$

Donc

$$\mathbb{E}[y | X] = \tanh \left( \beta_0 + \sum_{j=1}^p \beta_j x_j \right)$$

La densité d'une variable gaussienne de moyenne  $\mu$  et de variance  $\sigma^2$  s'écrit :

$$\begin{aligned} f_{\theta}(y) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\ &= \exp \left\{ \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2} \left( \frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right\} \end{aligned}$$

Donc on a :  $\theta = \mu, \phi = \sigma^2, b(\theta) = \frac{\theta^2}{2}$  et  $c(y, \phi) = -\frac{1}{2} \left( \frac{y^2}{\phi} + \log(2\pi\phi) \right)$

La fonction  $h$  de la régression tanh sera donc :

$$h = (g \circ b')^{-1} = \tanh(x)$$

avec  $g(x) = \operatorname{arctanh}(x)$  et  $b(\cdot)$  la fonction identité.

En utilisant l'équation 3.2, nous pouvons écrire la fonction de vraisemblance de la régression tanh :

$$\ell_n(\beta; y, X) = \sum_{i=1}^n \frac{y_i \tanh(X_i' \beta) - \tanh(X_i' \beta)}{\phi}$$

Comme ce type de régression n'existe pas encore sur  $\mathbb{R}$ , alors une adaptation de la fonction glm sur  $\mathbb{R}$  dont notamment sa fonction de lien a été faite pour l'estimation des paramètres dans le cadre de cette étude.

$\forall i \in \{1, \dots, p\}$ , on a :

$$\frac{\partial y}{\partial x_i} = \beta_i \left( 1 - \tanh^2 \left( \beta_0 + \sum_{j=1}^p \beta_j x_j \right) \right)$$

A priori, il n'existe pas une interprétation naturelle des paramètres de ce modèle. Nous pouvons juste dire qu'une augmentation d'une unité d'une variable quantitative  $x_i$  aura au plus comme impact une augmentation de  $\beta_i$  de  $y$ .

### 3.3 L'arbre de décision CART

Un autre type de modèle très utilisé en machine learning pour effectuer des classifications et des régressions est l'arbre de décision. En effet, en dehors des modèles linéaires, l'arbre de décision est l'un des modèles le moins complexes et faciles à interpréter. Nous allons voir par la suite la méthode de construction d'un arbre de décision et comment on peut le complexifier pour obtenir des modèles moins interprétables, mais plus performant et plus complexe comme le XGboost.

Il existe plusieurs types d'algorithme d'arbre de décision, mais dans le cadre de ce mémoire nous allons juste détailler le principe de l'algorithme CART : Classification and Regression Tree. La méthode CART a été conçue et formalisée par Leo Breinman et al.(1984) (13) et est qualifiée de méthode de partitionnement récursif ou de segmentation. En effet, le principe de base de l'algorithme consiste à partitionner successivement en deux sous-groupes les individus pour pouvoir les classifier ou prédire une variable quantitative.

#### Principe de l'arbre de décision

Soient  $n$  la taille de l'échantillon,  $Y$  la variable réponse qualitative à  $m$  modalités ou quantitative et  $p$  le nombre des variables explicatives pouvant être catégorielles ou continues  $(X_1, \dots, X_p)$ . Pour fixer les idées, nous nous plaçons dans le cadre d'un arbre de classification binaire qui est

basée sur la détermination d'une suite de nœuds.

- Un nœud est défini par une variable explicative et une division qui permet de faire une séparation en deux classes autrement dit, un nœud donné de l'arbre correspond un sous-ensemble d'échantillon auquel est appliquée une règle qui permet de séparer en deux classes les individus présents dans le nœud.
- Une division est définie par une séparation en deux sous-ensembles des modalités si la variable est qualitative ou le choix d'une valeur seuil si la variable est continue.
- Tous les individus de l'échantillon sont présents dans le nœud initial. On applique la division pour obtenir deux sous-échantillons et cette procédure sera itérée sur chacun des sous-ensembles jusqu'à ce qu'on obtienne des nœuds terminaux.

Dans ce cas, pour fonctionner, l'algorithme a besoin d'un critère qui permet de sélectionner la meilleure division, une règle qui permet de décider qu'un nœud est terminal : c'est ce qu'on appelle la feuille de l'arbre, c'est-à-dire qu'il n'y aura plus de partition possible de ce nœud et que chaque feuille est affectée à l'une des modalités de la variable réponse si elle catégorielle ou une valeur si elle est continue.

### **Critère de division**

Une restriction sur le critère de division est nécessaire pour limiter le nombre de division possible pour gagner en termes de temps. En effet, le critère d'admissibilité des divisions permet de limiter et de connaître le nombre de divisions candidates pour séparer un nœud. Nous disons qu'une division est admissible si les deux nœuds descendants sont non vides, autrement dit il y a au moins un individu dans chaque nœud créé par la division. Donc, si la variable explicative est catégorielle ordonnée avec  $m$  modalités alors il existe  $m - 1$  divisions admissibles dans le cas contraire il y a  $2^{m-1} - 1$  divisions admissibles. Le cas d'un facteur explicatif continu revient au même qu'une variable explicative ordinale.

### **Fonction d'hétérogénéité**

Pour le choix de la meilleure division d'un nœud, il est nécessaire de définir une fonction d'hétérogénéité qui permettra de déterminer laquelle des variables explicatives permet de partager en deux sous-ensembles les plus homogènes possibles les individus présents dans un nœud au sens de la variable cible. Cette fonction devrait être positive, égale à 0 si, et seulement si, les individus qui ont la même modalité ou la même valeur de la variable réponse se retrouvent tous dans le même sous-ensemble et maximale si les valeurs de la variable réponse sont très dispersées ou équiprobables. Notons  $\kappa_G$  et  $\kappa_D$  les deux nœuds fils créés à partir d'une division d'un nœud  $\kappa$ . L'algorithme retiendra donc la division qui minimise la somme de la fonction d'hétérogénéité des deux nœuds fils  $D_{\kappa_G} + D_{\kappa_D}$ . Le problème revient donc à chaque étape de



construction de l'arbre :

$$\max_{\{\text{divisions de } X^j; j=1, \dots, p\}} D_\kappa - (D_{\kappa_G} + D_{\kappa_D})$$

Pour une variable réponse  $Y$  quantitative, la fonction d'hétérogénéité d'un nœud  $\kappa$  est définie par :

$$D_\kappa = \frac{1}{|\kappa|} \sum_{i \in \kappa} (y_i - \bar{y}_\kappa)^2$$

où  $|\kappa|$  est le nombre d'observations dans le nœud  $\kappa$ .

L'objectif est de trouver pour chaque nœud de division la variable et la règle de division qui permettra de rendre le plus homogène possible les deux nœuds fils en minimisant la somme des variances intra classes :

$$\frac{|\kappa_G|}{n} \sum_{i \in \kappa_G} (y_i - y_{\kappa_G})^2 + \frac{|\kappa_D|}{n} \sum_{i \in \kappa_D} (y_i - \bar{y}_{\kappa_D})^2$$

Dans le cas d'une variable cible qualitative à  $m$  modalités, les fonctions d'hétérogénéité les plus utilisées sont :

— L'entropie :

$$D_\kappa = -2 \sum_{\ell=1}^m |\kappa| p_\kappa^\ell \log(p_\kappa^\ell)$$

$p_\kappa^\ell$  est la proportion de la classe  $\mathcal{T}_\ell$  de  $Y$  dans le nœud  $\kappa$ .

— L'indice de GINI :

$$D_\kappa = \sum_{\ell=1}^m p_\kappa^\ell (1 - p_\kappa^\ell)$$

De la même manière que pour une variable quantitative, l'objectif sera de trouver la variable et la division qui minimise l'une de ces deux fonctions.

### Règle d'arrêt

L'algorithme s'arrête à un nœud terminal donné ou feuille, lorsqu'il n'existe plus de partition admissible ou lorsqu'il est homogène ou lorsque le nombre des individus de la classe minoritaires ou le nombre d'observations présentes dans le nœud est inférieur à un seuil à choisir.

### Prédiction

Si la variable à prédire  $Y$  est quantitative, alors la valeur de la prédiction sera la moyenne de la variable  $Y$  des individus présents dans la feuille lors de l'entraînement. Si  $Y$  est catégorielle, pour chaque feuille de l'arbre est affectée une classe ou modalité de  $Y$  en prenant soit la classe majoritaire dans le nœud ou soit la classe a posteriori la plus probable au sens de Bayes si des probabilités a priori sont connues soit la classe la moins coûteuse si une fonction de coût de mauvais classement est définie.

### Elagage de l'arbre

La précédente démarche permet de construire l'arbre maximal, ce qui peut conduire parfois à un problème de sur-apprentissage. C'est donc un modèle à éviter au profit des modèles moins complexes (en termes de nombre de nœuds et de profondeur) pour avoir un modèle plus robuste. L'élagage de l'arbre maximal revient donc à trouver un arbre qui se situe entre l'arbre trivial (qui a une seule feuille) et l'arbre maximal  $A_{\max}$ . Cependant, il existe un nombre important de sous-arbres, donc il n'est pas possible de tous les considérer. Leo Breinman et al.(1984) (13) ont proposé une méthode qui consiste à construire une suite emboîtée de sous-arbre de l'arbre maximal et de choisir parmi cette suite.

### Construction de suite d'arbres

Soient  $A$  un arbre de décision et  $K_A$  le nombre de feuilles ou de nœuds terminaux de l'arbre  $A$  qui traduit sa complexité. La qualité d'ajustement de l'arbre  $A$  se définit comme :

$$D(A) = \sum_{\kappa=1}^{K_A} D_{\kappa}$$

où  $D_{\kappa}$  est la fonction d'hétérogénéité du feuille  $\kappa$  de l'arbre  $A$ .

Une pénalisation de la complexité de l'arbre sera introduite pour construire la séquence d'arbres emboîtés :

$$C(A) = D(A) + \gamma \times K_A$$

Pour  $\gamma = 0$  on a  $A_{\max} = A_{K_A}$ . Si on augmente la valeur de  $\gamma$ , l'une des divisions de  $A_{K_A}$ , celle dont l'amélioration de  $D$  est moindre (inférieur à  $\gamma$ ), apparaît comme non pertinent et ses deux nœuds fils seront donc regroupés (élagués) dans la feuille mère qui devient un nœud terminal, donc  $A_{K_A}$  devient  $A_{K_A-1}$  et on réitère le processus :

$$A_{\max} = A_{K_A} \supset A_{K_A-1} \supset \dots A_1$$

où  $A_1$  est le nœud initial de l'arbre.

On construit une séquence  $A_1, \dots, A_K$  pour des valeurs décroissantes du coefficient de pénalisation  $\gamma$ .

Une fois la construction de la séquence finie, il est question maintenant de trouver l'arbre optimal qui minimise une erreur de généralisation. On utilise souvent la méthode de validation croisée pour choisir la valeur de  $\gamma$  qui minimise l'erreur de prévision et l'arbre sera considéré comme optimal dans la séquence estimée sur l'échantillon d'entraînement.

L'algorithme d'élagage d'arbre dans (20) se résume comme suit :

**Algorithm 2** Sélection d'arbre ou élagage par validation croisée

---

Construction de l'arbre maximal  $A_{\max}$

Construction de la séquence  $A_K \dots A_1$  d'arbres emboîtés associée à une séquence de valeurs de pénalisation  $\gamma_\kappa$  sur tout l'échantillon

**Pour**  $v = 1, \dots, V$  **faire**

    Pour chaque échantillon, estimation de la séquence d'arbres associée à la séquence des pénalisations  $\gamma_\kappa$

    Estimation de l'erreur sur la partie restante de validation de l'échantillon

**FinPour**

Calcul de la séquence des moyennes de ces erreurs

L'erreur minimale désigne la pénalisation  $\gamma_{\text{opt}}$  optimale

Retenir l'arbre associé à  $\gamma_{\text{opt}}$  dans la séquence  $A_K \dots A_1$

---

où  $V$  désigne le nombre d'échantillons obtenus en divisant en  $V$  sous-groupes l'échantillon initial.

## 3.4 Le XGBoost

Les modèles d'arbres de décision produisent souvent des modèles moins performants et instables. D'autres types d'algorithmes comme le bagging<sup>1</sup> et le boosting sont appliqués sur les arbres de décision pour fournir un modèle plus robuste et plus performant. Ces méthodes consistent à construire plusieurs modèles (des arbres par exemple) et de les utiliser pour faire la prédiction.

### Le principe du Boosting

Le boosting est un algorithme basé sur l'amélioration d'un classifieur faible (faible taux de bonne prédiction) et l'agrégation par moyenne pondérée ou vote d'une famille de modèles. En effet, le boosting consiste à construire une séquence de modèle dont chaque modèle est une version adaptative du précédent en modifiant le poids de chaque individu en fonction de l'erreur d'estimation, plus l'observation est mal ajustée plus son poids dans le prochain modèle sera important.

### L'Adaboost

La première version du modèle de boosting est l'Adaboost ou adaptative boosting, c'est un problème de classification binaire dont la fonction de discrimination sera notée  $\delta$  et qui est facile à adapter pour des modèles de régression. Dans cet algorithme, le poids de chaque observation est initialisé à  $1/n$  pour le premier modèle puis évolue pour chaque modèle. Il augmente proportionnellement à l'erreur de prédiction que le précédent modèle commet et reste inchangé si l'observation est bien classée. La prédiction finale du modèle sera une combinaison linéaire pondérée par les qualités d'ajustement des différents modèles :  $\sum_{m=1}^M c_m \delta_m(\mathbf{x}_0)$ , où

---

1. Bagging : prévision par agrégation, il s'agit de la moyenne de la prédiction de plusieurs arbres si la variable réponse est quantitative et d'un vote si elle est qualitative

$$c_m = \log((1 - \hat{\varepsilon}_p)/\hat{\varepsilon}_p) \text{ et } \hat{\varepsilon}_p = \frac{\sum_{i=1}^n w_i 1\{\delta_m(x_i) \neq y_i\}}{\sum_{i=1}^n w_i}.$$

$$\text{Donc } \hat{f}_m(x) = \hat{f}_{m-1}(x) + c_m \delta(x; \gamma_m).$$

### Le Tree Gradient boosting

L'algorithme de Tree Gradient Boosting a la même base que l'algorithme d'Adaboost, la différence se situe au niveau de l'amélioration de chaque nouveau modèle en utilisant la direction de gradient de la fonction de perte. Soient  $M$  le nombre de modèles (arbres) à construire et  $l$  la fonction de perte à définir en fonction du problème, alors pour tout  $m$  appartenant à  $2, \dots, M$  :

$$\hat{f}_m = \hat{f}_{m-1} - \gamma_m \sum_{i=1}^n \nabla_{f_{m-1}} l(y_i, f_{m-1}(x_i))$$

L'algorithme de Gradient Tree Boosting dans (21) se résume comme suit :

---

#### Algorithm 3 Gradient Tree Boosting pour la régression

---

Soit  $x_0$  à prévoir

Initialiser  $\hat{f}_0 = \arg \min_{\gamma} \sum_{i=1}^n l(y_i, \gamma)$

**Pour**  $m = 1, \dots, M$  **faire**

Calculer  $r_{mi} = - \left[ \frac{\partial l(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}} ; i = 1, \dots, m$

Ajuster un arbre de régression  $\delta_m$  aux couples  $(x_i, r_{mi})_{i=1, \dots, n}$

Calculer  $\gamma_m$  en résolvant :  $\min_{\gamma} \sum_{i=1}^n l(y_i, f_{m-1}(x_i) + \gamma \delta_m(x_i))$

Mise à jour :  $\hat{f}_m(x) = \hat{f}_{m-1}(x) + \gamma_m \delta_m(x)$

**FinPour**

Résultat  $\hat{f}_M(x_0)$

---

### XGBoost

Le XGBoost est une version améliorée et optimisée de l'algorithme de Gradient Boosting. Sa prédiction est basée sur la moyenne pondérée de plusieurs algorithmes qualifiés de faibles. Son avantage par rapport au modèle de Gradient Boosting est que le XGBoost limite le nombre de séparations possibles d'un nœud alors que le Gradient Boosting considère la perte potentielle pour toute possibilité de séparation d'un nœud, ce qui est très coûteux en termes de temps de calcul lorsqu'on dispose de plusieurs variables explicatives.

La différence entre le XGBoost et le Gradient Boosting réside sur l'introduction des paramètres de régularisation pour éviter les problèmes de sur-apprentissage et pour avoir une meilleure performance. La fonction objectif devient donc :

$$\mathcal{L}(f) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{m=1}^M \Omega(\delta_m)$$

avec  $M$  le nombre des arbres et  $\delta_m \in \mathcal{F}$  avec  $\mathcal{F}$  la famille des arbres CART.

$$\Omega(\delta) = \alpha|\delta| + \frac{1}{2}\beta\|w\|^2$$

où  $|\delta|$  est le nombre de feuilles de l'arbre de régression  $\delta$  et  $w$  le vecteur des valeurs attribuées à chacune de ses feuilles.

Cette pénalisation ou régularisation limite l'ajustement de l'arbre ajouté à chaque étape et contribue à éviter un surapprentissage des données. L'augmentation du nombre d'itérations peut dégrader la qualité d'ajustement du modèle lorsque les observations mal prédites sont présentes. Le terme  $\Omega$  peut être vu donc comme une combinaison de régularisation Ridge de coefficient  $\beta$  et de pénalisation Lasso de coefficient  $\alpha$ .

### 3.5 Interprétation SHAP (Shapley Additive exPlanations)

En dehors des méthodes d'interprétation et de mesure d'importance de variable classiques du modèle XGBoost que nous allons voir dans le chapitre 5, nous allons utiliser la méthode SHAP. Cette méthode a été proposée par Lundberg dans le but de mesurer la contribution de chaque variable à la valeur de prédiction de la variable cible pour chaque individu. L'idée est de calculer la valeur Shapley comme en théorie de jeux pour toutes les variables explicatives pour tous les individus de l'échantillon. En effet, elle permet de calculer la contribution marginale d'une variable explicative dans la prédiction. De plus, cette approche a l'avantage de vérifier une propriété d'additivité, autrement dit la somme de la valeur SHAP de toutes les variables est égale à la prédiction du modèle pour tous les individus et que les variables manquantes n'ont pas d'impact sur la prédiction.

La valeur de Shapley  $\phi_i$  d'une variable  $i$  s'écrit :

$$\phi_i = \sum_{S \subseteq M \setminus i} \frac{|S|!(M - |S| - 1)!}{M!} (f_x(S \cup i) - f_x(S))$$

où  $M$  le nombre de variables explicatives,  $S$  un ensemble de variable et  $f_x$  la fonction de prédiction avec  $f_x(S) = \mathbb{E}[f(x) \mid x_S]$ .

Par additivité de l'approche SHAP, la prédiction du modèle peut être vue comme :

$$\hat{f}(x) = \phi_0 + \sum_{i=1}^M \phi_i z'_i$$

avec  $\phi_0$  la valeur moyenne de la prédiction sur l'échantillon,  $\hat{f}(x)$  la prédiction du modèle,  $z'_i \in \{0, 1\}$  quand la variable est observée  $z'_i = 1$  ou manquante  $z'_i = 0$ .

## 3.6 Intervalle de confiance

Au-delà de la précision et de l'interprétabilité, l'un des plus grands défis actuels en statistique est de trouver une méthode qui permet de quantifier le niveau de confiance que nous avons vis-à-vis de la prédiction fournies par un modèle de machine learning. La connaissance de l'incertitude autour de la valeur prédite permet de renforcer et d'améliorer la prise de décision. En effet, la valeur prédite par un modèle de machine learning n'est pas absolue, il est donc important de connaître les potentielles variations de la prédiction avec un niveau de certitude donné afin de mesurer sa robustesse. La notion d'intervalle de confiance renvoie généralement à la notion de la loi de probabilité et de la distribution d'une variable aléatoire, cependant, en machine learning la loi de probabilité de la variable cible n'est pas souvent connue. Il est donc nécessaire d'utiliser d'autres méthodes pour construire l'intervalle de confiance de la prédiction. En générale, il existe deux types de méthodes pour estimer les bornes de l'intervalle de confiance de la prédiction avec un niveau de confiance donné pour les modèles de machine learning : la "régression quantile" et le "percentile bootstrap".

### 3.6.1 Le percentile bootstrap

L'une des méthodes la plus utilisée en statistique pour faire de l'inférence sur la variabilité d'une variable est la méthode bootstrap. Le principe consiste à générer des échantillons factices par des tirages avec remise à partir de l'échantillon d'entraînement initial et entraîner un modèle sur chaque échantillon et prédire la variable cible sur l'échantillon test afin d'obtenir une distribution empirique de la prédiction. L'inconvénient de cette méthode est qu'elle très coûteuse en termes de temps, donc elle est difficile à mettre en œuvre pour des bases des données volumineuses.

---

**Algorithm 4** Bootstrap

---

Échantillon initial  $(Y_i, X_i)_{i=1\dots n}$ Échantillon test  $(Y_{test}, X_{test})$ **Pour**  $b = 1, \dots, B$  **faire**    Tirer avec remise un échantillon de taille  $n$  à partir de l'échantillon initial  $(Y_i, X_i)_{i=1\dots n}$     Entraîner un modèle  $f_b$  sur le nouvel échantillon    Prédire la variable cible  $\hat{y}_{test_b} = \hat{f}_b(x_{test})$ **FinPour**Résultat  $\{\hat{y}_{test_1}, \dots, \hat{y}_{test_B}\}$ 

---

La méthode de percentile bootstrap a pour principe de prendre directement le quantile empirique pour construire l'intervalle de confiance de la prédiction. Soit  $q_\alpha$  le quantile empirique d'ordre  $\alpha$  de la prédiction obtenu par la méthode de percentile bootstrap, nous obtenons

l'intervalle de confiance de la prédiction par :

$$IC_{1-\alpha} = [q_{\alpha/2}, q_{1-\alpha/2}] \quad (3.3)$$

### 3.6.2 La régression quantile

Soient  $Y$  la variable réponse,  $F_Y$  ( $F_Y(y) = P(Y \leq y)$ ) sa fonction de répartition et  $X$  l'ensemble des facteurs explicatifs. Le quantile d'ordre  $\alpha$  d'une variable aléatoire  $Y$  est définie par  $q_\alpha(Y) = \inf \{y : F_Y(y) \geq \alpha\}$ . La régression quantile est une méthode qui tente d'évaluer comment les quantiles conditionnels  $q_\alpha(Y | X) = \inf \{y : F_{Y|X}(y) \geq \alpha\}$  de la variable réponse évoluent en fonction des variables explicatives. L'idée derrière la régression quantile est de modifier la fonction objectif d'un modèle statistique ou d'un modèle de machine learning pour que ce dernier puisse prédire directement le quantile conditionnel d'ordre  $\alpha$  de la variable cible.

La fonction objectif de la régression quantile d'ordre  $\alpha$  se définit par :

$$L_\alpha(y, \hat{y}) = \sum_{i=y_i < \hat{y}_i} \alpha |y_i - \hat{y}_i| + \sum_{i=y_i \geq \hat{y}_i} (1 - \alpha) |y_i - \hat{y}_i|$$

Pour expliquer l'intuition derrière cette fonction de perte, nous allons considérer le quantile le plus utilisé qui n'est rien d'autre que la médiane, c'est-à-dire  $\alpha=0.5$ . Si nous remplaçons  $\alpha=0.5$ , la fonction de perte quantile devient la MAE. L'utilisation de la fonction de perte MAE permet donc de prédire la médiane d'une variable réponse. Considérons maintenant trois points sur une droite verticale avec des différentes distances par rapport à chacun : le point supérieur, le point moyen et le point inférieur. Dans ce problème à une dimension, la MAE coïncide avec la distance. L'hypothèse à tester ici est que la médiane (le point moyen) est le point qui minimise la MAE. Pour vérifier notre hypothèse, si nous rapprochons le point moyen du point supérieur d'une distance donnée et nous éloignons le point inférieur du point moyen de la même distance, ce qui va faire augmenter la MAE (la distance moyenne entre les points). Le point moyen reste le point qui minimise la MAE. De même, si nous réalisons la même chose dans le sens inverse, notre hypothèse sera toujours vérifiée c'est-à-dire que le point moyen est à la fois la médiane et le point qui minimise la fonction de perte MAE. Ce résultat est toujours vérifié même si nous augmentons le nombre d'observations.

Dans la fonction de perte de la régression quantile, pour  $\alpha = 0.5$ , les sous-prédiction et les sur-prédiction seront pénalisées avec le même poids et nous obtenons la médiane. Pour une valeur élevée de  $\alpha$ , les sur-prédictions seront plus pénalisées que les sous-prédictions. Pour  $\alpha=0.75$ , le poids des sur-prédictions sera égal à 0.75 et le poids des sous-prédictions sera égal à 0.25, c'est-à-dire que le modèle essaiera alors d'éviter les sur-prédictions environ trois fois plus que les sous-prédictions, et le quantile à 0,75 sera obtenu. La fonction de perte de la régression quantile présente un inconvénient d'avoir une dérivée seconde égale à 0, donc elle ne peut pas être utilisée pour des modèles de machine learning qui utilise le gradient et la matrice hessienne

de la fonction. Pour remédier à ce problème, il s'avère nécessaire d'approximer cette fonction de perte par une fonction qui possède une dérivée continue et différentiable.

Yogesh Bagul (2017) (23) dans son travail propose d'approximer la fonction  $f(x) = |x|$  par  $x \tanh(x/\mu)$  avec  $\mu > 0 \in \mathbb{R}$  et nous allons utiliser ce résultat pour proposer une nouvelle fonction de perte pour approximer la fonction de perte de quantile.

En effet, on a :

$$\begin{aligned} \left| \tanh\left(\frac{x}{\mu}\right) \right| &\leq 1 \\ |x| \left| \tanh\left(\frac{x}{\mu}\right) \right| &\leq |x| \\ |x| \left( 1 - \left| \tanh\left(\frac{x}{\mu}\right) \right| \right) &\leq 0 < |\mu| \end{aligned}$$

Donc :

$$1 - \left| \tanh\left(\frac{x}{\mu}\right) \right| < \frac{|\mu|}{|x|}$$

Alors

$$\begin{aligned} \left| |x| - x \tanh\left(\frac{x}{\mu}\right) \right| &= \left| |x| - |x| \tanh\left(\frac{x}{\mu}\right) \right| \\ &= \left| |x| - |x| \left| \tanh\left(\frac{x}{\mu}\right) \right| \right| \\ &= |x| \left\{ 1 - \left| \tanh\left(\frac{x}{\mu}\right) \right| \right\} \\ &< |x| \frac{|\mu|}{|x|} = |\mu| = \mu \end{aligned}$$

Finalement on a :

$$\lim_{\mu \rightarrow 0} \left| |x| - x \tanh\left(\frac{x}{\mu}\right) \right| = 0$$

Pour  $\mu > 0 \in \mathbb{R}$ , la fonction de perte de quantile d'ordre  $\alpha$  peut donc être approximée par :

$$L_\alpha(y, \hat{y}) = \sum_{i=y_i < \hat{y}_i} \alpha(y_i - \hat{y}_i) \tanh\left(\frac{y_i - \hat{y}_i}{\mu}\right) + \sum_{i=y_i \geq \hat{y}_i} (1 - \alpha)(y_i - \hat{y}_i) \tanh\left(\frac{y_i - \hat{y}_i}{\mu}\right)$$



## CONSTRUCTION ET DESCRIPTION DU PORTEFEUILLE

Ce chapitre est consacré à la description du portefeuille de GENERALI sur le produit épargne. Après une présentation des différentes variables, un test d'indépendance entre les variables explicatives et la variable réponse sera réalisé. La dernière partie de ce chapitre est consacrée aux choix des métriques d'évaluation des modèles.

### 4.1 Construction de la base

#### 4.1.1 Définition du périmètre d'étude

La modélisation des lois comportementales en assurance-vie nécessite de bien définir en premier le périmètre d'étude et le phénomène à analyser. La base de données qui sera utilisée pour la modélisation des arbitrages provient de la jointure de plusieurs tables SAS et de la création de quelques variables qui sont jugées pertinentes et utiles pour la modélisation. Les principales tables utilisées pour extraire les données sont les suivantes :

- Une table répertoriant les mouvements d'arbitrage par contrat 01/2004 à 06/2021 (76 116 726 lignes) ;
- Une table qui contient le montant de PM mensuelle par support de chaque contrat 01/2007 à 06/2021 (323 252 447 lignes) ;
- Une table répertoriant les caractéristiques des contrats et les caractéristiques des assurés ;
- L'historique du rendement des différents fonds en euro par année de 2012 à 2019, la reconstitution du rendement du fonds en UC s'est avérée difficile car on ne dispose de l'historique de leur performance donc on a utilisé la moyenne du rendement du CAC 40 et de l'Eurostoxx comme rendement des fonds en UC ;
- Une table répertoriant le taux de PB de chaque contrat.

Les opérations d'arbitrages dans la table initiale sont enregistrées à la maille journalière et par support (Fonds Euro 1, Fonds Euro 2, Fonds UC 1, Fonds UC2, ...). La première étape réalisée pour la construction de la base des données consiste à agréger les opérations par contrat, par année et par type de support, c'est-à-dire qu'on fait juste la distinction entre fonds en euros et fonds en UC. Le choix du pas de temps annuel a été justifié par le fait que certaines variables qui sont jugées très importantes comme le rendement des fonds en euros par exemple ne

sont disponibles qu'au pas annuel et que le pas de temps utilisé pour le calcul du BE est l'année.

Seules les observations dont l'écart d'arbitrage est inférieur à 5% sont retenues par la suite. L'écart d'arbitrage se calcule comme suit :

$$\text{Ecart} = \frac{\text{Montant sorti} - \text{Montant entrant} - \text{frais d'arbitrage}}{\text{Montant sorti}}$$

En théorie, la valeur de l'écart devrait être nulle, cependant, on a observé parfois quelques anomalies dans les tables initiales ce qui nous a amenés à retenir ce critère d'exclusion. Toutes les observations qui ont des valeurs manquantes sur la provision mathématique ou sur le montant d'arbitrage ont été supprimées du périmètre d'observation. En effet, l'absence de la provision mathématique ou du montant d'arbitrage ne nous permet pas de calculer le taux d'arbitrage de l'année d'un contrat donné. Après traitement, nous disposons de deux bases de données, lesquelles seront utilisées pour la modélisation de la fréquence d'arbitrage et du taux d'arbitrage. La base des données pour la fréquence comporte 2 920 000 observations et la base des données pour la modélisation du taux d'arbitrage comporte 450 000 observations. Les deux bases se portent sur la période 2013-2020. Le pas de temps utilisé dans la modélisation sera l'année.

## 4.1.2 Choix et Justifications des variables explicatives

### 4.1.2.1 Variable à modéliser

#### — Le nombre d'arbitrages

Cette variable traduira la décision à arbitrer ou non de l'assuré au cours de l'année donc elle prendra une valeur dans l'ensemble  $\{0, 1\}$ . Cette variable prendrait la valeur 1 si l'assuré avait effectué une ou plusieurs opérations d'arbitrage du fonds en euro vers le fonds en UC ou du fonds en UC vers le fonds en euro pendant l'année. A priori, le choix d'arbitrage d'un assuré va dépendre à la fois des caractéristiques de son contrat et de la conjoncture économique, donc on ne fera pas de distinction entre l'effet structurel et l'effet conjoncturel pour la modélisation de cette variable.

#### — Le taux d'arbitrage

Initialement et dans la réalité, nous ne pouvons pas observer la composante structurelle et la composante conjoncturelle des arbitrages. Nous ne disposons que du taux d'arbitrage total net de l'assuré pendant l'année qui prend ses valeurs dans l'intervalle  $[-1, 1]$ . Un taux d'arbitrage à  $-0.1$  signifie que l'assuré a effectué un arbitrage de 10% de son PM en UC vers le fonds en euros, inversement un taux d'arbitrage égal à  $0.1$  voudra dire que l'assuré a arbitré 10% de son PM en euros vers le fonds en UC.

#### 4.1.2.2 Les variables explicatives

— **Le montant de la PM**

Le montant de la PM peut être assimilé à la richesse initiale de l'assuré comme dans l'analyse microéconomique qui est perçue comme un indicateur important pour sa fonction d'utilité donc peut a priori influencer ses opérations d'arbitrage. Un résultat classique qu'on retrouve dans la microéconomie est que plus un agent dispose d'une richesse initiale importante plus il a de l'aversion pour le risque, ceci peut s'expliquer par le fait que les assurés qui détiennent un capital important dans leur contrat d'assurance-vie sont plus sensibles au rendement de leur placement. Cependant, nous resterons prudents dans l'interprétation de cette variable, car on ne peut pas assimiler directement le contrat d'assurance-vie de l'assuré à son patrimoine, il se peut qu'il possède d'autres formes de placement par exemple.

— **La Part UC dans le contrat**

Le poids du fonds UC dans l'épargne de l'assuré correspond au ratio du montant de PM des fonds en UC sur le montant de PM totale du contrat. Elle permet de qualifier à première vue le degré d'aversion au risque d'un assuré. En effet, plus la part UC du contrat est proche de 1, plus l'assuré a du goût pour le risque.

— **Nombre de fonds UC dans le contrat**

Cette variable a été créée pour surmonter la limite de la variable Part UC. A priori, on peut penser que deux assurés dont l'épargne est composée de 70% d'unités de compte chacun n'ont pas forcément la même appétence pour risque. L'assuré qui diversifie ses placements en unités de compte est plus averse au risque que celui qui ne le fait pas. La diversification du placement permet d'éliminer ou de réduire les risques idiosyncratiques. Un autre indicateur qui n'est rien d'autre que le rating des fonds en UC à partir de leur volatilité aurait été plus intéressant pour mieux décrire l'aversion au risque des assurés, cependant cette information n'était pas disponible dans les tables initiales.

— **Rendement du contrat**

C'est l'indicateur par excellence que les assurés utilisent pour prendre la décision d'arbitrer ou non et pour décider du montant à arbitrer. L'analyse du rendement de son contrat et du rendement du fonds en euro et du fonds en UC donne un aperçu au client du potentiel gain ou potentielle perte qu'il aura à l'avenir s'il décide de réorienter son placement dans l'un des deux fonds. Pour les contrats multisupports, le rendement du contrat est obtenu en faisant la moyenne des rendements pondérées par la provision mathématique.

— **L'âge de l'assuré**

Souvent utilisé comme variable explicative dans un modèle de tarification ou dans un modèle de rachat, l'âge apparaît généralement comme une variable discriminante pour les modèles individuels dans le domaine de l'assurance. En effet, l'objectif et l'horizon de

placement ont a priori une influence sur la décision d'arbitrage d'un assuré et ces deux variables peuvent être assimilées à son âge. Le raisonnement classique utilisé en assurance-vie est que plus l'âge avance plus les assurés ont tendance à sécuriser leur épargne pour préparer leur retraite par exemple.

— **L'ancienneté du contrat**

Elle est très pertinente dans l'analyse des rachats mais ce n'est pas le cas a priori pour les arbitrages. En effet, comme les capitaux ne sortent pas du portefeuille de l'assureur donc aucune fiscalité spécifique n'est appliquée lors de la réalisation d'une opération d'arbitrage. Cependant, étant donné le dynamisme de l'activité de l'assurance, l'offre des nouveaux produits d'assurance-vie peut inciter les clients qui sont restés longtemps dans le portefeuille à changer de contrat par exemple et leurs années d'adhésions peuvent leur donner aussi une meilleure connaissance de leur support d'investissement donc facilite le choix de continuer à rester sur le même fonds ou choisir un autre fonds qui peut procurer un meilleur rendement ou plus de sécurité.

— **Le sexe** Souvent utilisé comme l'âge comme variable explicative en assurance, nous n'avons pas cependant aucun a priori sur son influence sur les arbitrages. Une analyse effectuée par la suite nous dira si cette variable est discriminante ou non.

— **Le mode de gestion**

Il est recommandé de séparer la modélisation des arbitrages selon le mode de gestion selon l'initiateur de la décision d'arbitrage. En effet, pour le mode de gestion libre, les arbitrages dépendent directement du comportement des assurés alors que pour le mode de gestion automatique (profilée et à horizon), les arbitrages découlent juste de l'exécution des options à laquelle les assurés ont souscrites. La prise en compte de cette recommandation introduira une nouvelle complexité en plus de la séparation des arbitrages conjoncturels et structurels. Elle sera considérée en revanche comme une variable explicative dans notre modèle.

— **Indice de périodicité des primes**

Les contrats d'assurance-vie à prime unique sont adaptés pour le placement d'un capital plus ou moins important donc les assurés qui optent pour ce mode de versement de prime sont a priori plus sensibles aux variations du rendement de leurs contrats que ceux qui optent pour le versement périodique. De plus, les épargnants qui choisissent les contrats à primes périodiques ont généralement pour objectif de constituer une épargne progressive au fil du temps donc ils vont avoir tendance à diminuer au fur et à mesure le poids du fonds en UC quand la valeur de son contrat augmente.

— **TMG**

Pour les fonds en euros.

— **Taux de PB**

Pour les fonds en euros.

— **Taux CAC 40**

La prise en compte du rendement de l'indice CAC 40 permet de prendre en compte la conjoncture économique dans notre modèle. La situation du CAC 40 peut traduire la santé du marché boursier en France comme c'est le regroupement des 40 plus grandes entreprises françaises donc on peut supposer que l'évolution du rendement la plupart des fonds unités de compte va suivre la même tendance. De plus, c'est l'un des indices boursiers les plus suivis en France, donc a priori, on a une bonne raison de supposer que les assurés vont l'utiliser comme référence pour appréhender le rendement des unités de compte.

— **Taux Eurostoxx**

Tout comme le rendement du CAC 40, l'indice Eurostoxx représente la santé financière des grandes entreprises européennes. Par conséquent, elle peut être utilisée par les assurés qui prônent la diversification géographique de leur placement. Elle peut être utilisée aussi comme référence pour anticiper l'évolution de la performance des fonds en unité de compte.

— **Taux Moyen des emprunts de l'Etat**

Le taux moyen des emprunts de l'Etat est généralement utilisé par les institutions financières pour fixer le niveau du taux des obligataires. Donc, ce taux peut être utilisé par les assurés comme référence pour trouver le rendement espéré de leur placement s'ils choisissent les supports sécurisés. C'est aussi l'indice de référence en assurance-vie pour déterminer les taux techniques maxima ou le TMG.

— **Taux 10 ans** Utilisé dans la même logique que le TME, le taux OAT 10 ans représente en général le rendement à long terme des obligations émises par l'Etat donc il peut être représentatif du rendement des supports sans risque à long terme. De plus, ce type d'obligation est moins risqué que les obligations des entreprises (ou corporate) à long terme, mais offre moins de rendement que ces dernières en revanche. Donc, ce taux peut refléter le rendement à long terme des placements sans risque attendu par les assurés.

### 4.1.2.3 Transformation des variables financières

Les variables financières en soient ne semble pas a priori être pertinentes pour appréhender le comportement d'arbitrage des assurés. En effet, lorsque l'assuré reçoit le relevé de la situation de son épargne, l'information qui l'intéresse en premier c'est le taux qu'on lui a servi pour ses placements et la première question qu'il se pose est combien il aura gagné et combien il va gagner s'il aurait décidé et s'il décide de réorienter le placement de son capital dans son contrat d'assurance vie. Ensuite, la deuxième question qui vient dans son esprit est : quel risque va-t-il courir s'il décide de réaliser cette opération ? Donc, on peut dire que les assurés ne vont pas regarder un par un le rendement des différents supports de placement, mais vont plutôt

faire une comparaison entre le rendement de leur contrat et les performances des autres fonds pour décider s'ils vont réorienter ou non leur capital. Ainsi, nous avons décidé de considérer les variables énoncées ci-dessus à l'instar des variables financières dans notre base de données :

- \* Écart entre le rendement du contrat et le taux du CAC 40
- \* Écart entre le rendement du contrat et le taux Eurostoxx
- \* Écart entre le rendement du contrat et le TME
- \* Écart entre le rendement du contrat et le Taux 10 ans

TABLEAU 4.1 – Liste des variables

<b>Variables</b>	<b>Signification</b>
PM_1	Montant de la PM annuelle
PART_UC_1	Part en UC dans le contrat
NB_FONDS_UC	Nombre de fonds en UC dans le contrat
AGE	Age de l'assuré
ANCIENNETE	Ancienneté du contrat
IND_PERIODICITE	Indice de périodicité des primes
CD_SEXE	Sexe de l'assuré
ID_MOD_GESTION	Mode de gestion du contrat
TMG	TMG du contrat
TX_PB	Taux de PB du contrat
Ecart_cac	Écart entre le rendement du contrat et du CAC 40
Ecart_erstx	Écart entre le rendement du contrat et de l'Eurostoxx
Ecart_tme	Écart entre le rendement du contrat et le Tme
Ecart_tx10ans	Écart entre le rendement du contrat et le taux 10 ans
RDT_EUR	Rendement du fonds en euros dans le contrat
RDT_UC	Rendement du fonds en UC dans le contrat

Notons que toutes ces variables sont calculées ou prises l'année précédant l'année d'arbitrage.

### 4.1.3 Traitement des valeurs manquantes et des valeurs aberrantes

Avant de commencer l'analyse descriptive de la base et la modélisation des variables d'intérêt, nous traitons d'abord le problème de valeurs manquantes et de valeurs aberrantes dans les variables explicatives qui résultent parfois d'une anomalie ou d'une erreur de saisie lors de la construction de la base comme la présence d'un âge négatif ou d'un âge qui dépasse 200 ans.

Le choix de la méthode d'imputation des valeurs manquantes dépend généralement de la cause et de la source de leur présence dans la base. Il est important de connaître si elles sont juste le simple fruit du hasard ou si ce sont des données manquantes de façon non-aléatoire, c'est-à-dire que la probabilité que la variable soit manquante dépend de sa nature. Dans notre cas, nous pouvons dire que les données sont manquantes de façon aléatoire, car elles sont dues généralement à des oublis lors de la souscription ou de l'enregistrement des opérations et la probabilité que ces données soient absentes est identique pour tous les assurés.

Une méthode classique pour imputer les valeurs manquantes de manière aléatoire est la complétion par une combinaison linéaire des observations non-manquantes. Elle consiste à substituer les observations manquantes par une combinaison linéaire des valeurs présentes dans la base. La méthode la plus répandue est la méthode d'imputation par la moyenne et c'est ce que l'on a utilisé pour remplacer les valeurs manquantes des variables explicatives quantitatives dans notre base. En ce qui concerne les variables qualitatives, nous avons décidé d'imputer les données non observées par le mode ou la modalité la plus fréquente.

Pour le cas des valeurs aberrantes, étant donné qu'elles sont dues la plupart à une erreur d'enregistrement lors de la souscription ou de la réalisation d'une opération, alors nous avons décidé d'appliquer le même traitement que pour les valeurs manquantes.

## 4.2 Analyse exploratoire

### 4.2.1 Analyse univariée

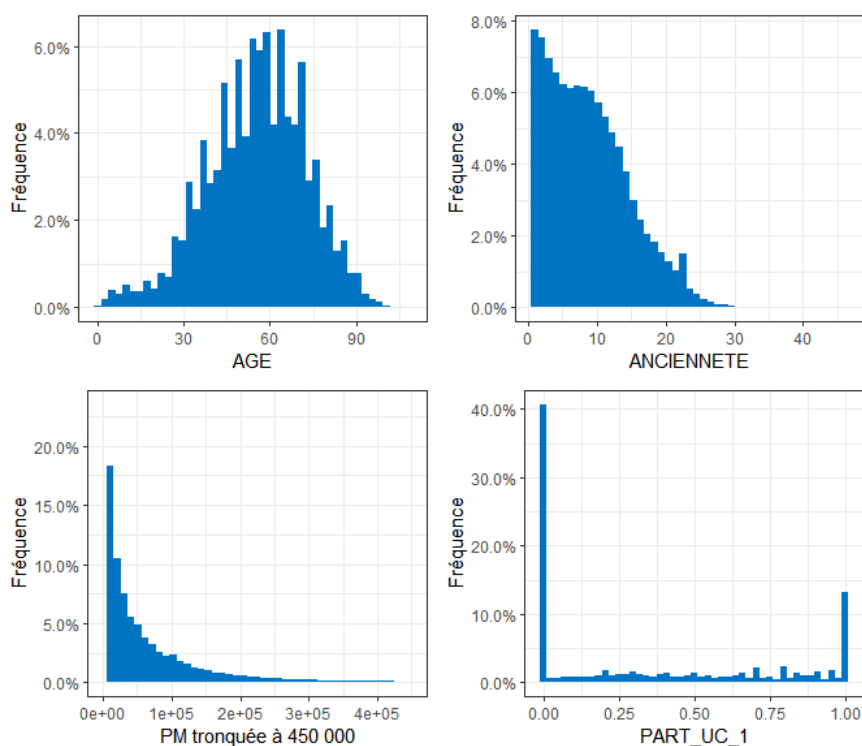


FIGURE 4.1 – Histogramme des variables quantitatives

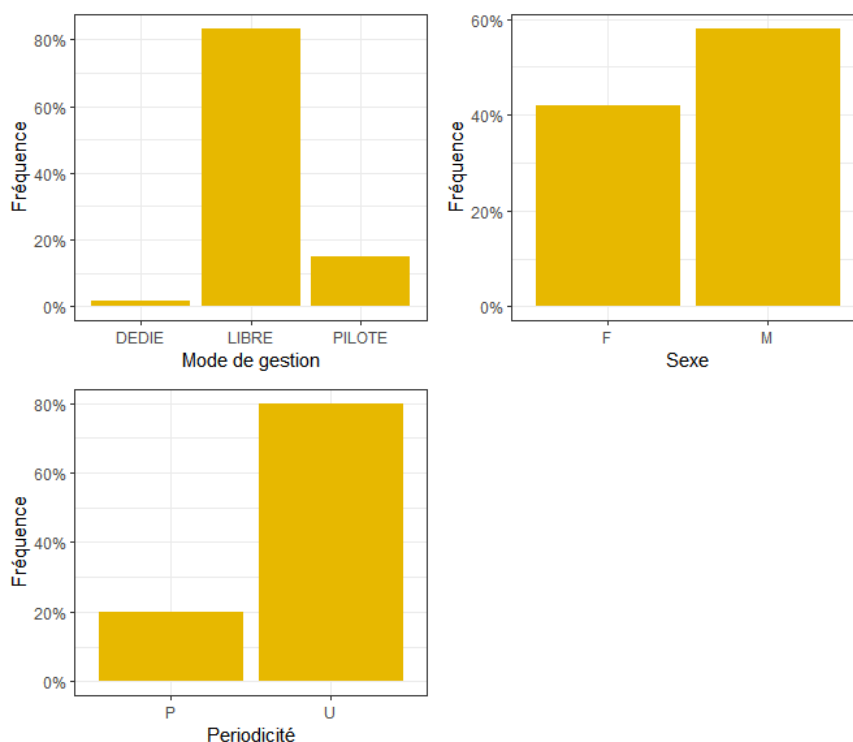


FIGURE 4.2 – Diagramme en bâtons des variables qualitatives

L'analyse univariée de chaque variable explicative montre que pour toute année confondue, un assuré moyen chez GENERALI possède un contrat d'assurance vie de 8 ans d'ancienneté et l'âge moyen des assurés est de l'ordre de 55 ans avec une distribution symétrique au tour de cette valeur centrale. En moyenne, un contrat d'assurance vie dans le portefeuille est constitué de 36 % de fonds en UC et opte en majorité (76 %) pour un mode de versement unique des primes. La majorité des clients présents dans le portefeuille (plus de 83 %) préfèrent gérer leurs contrats par eux-mêmes et choisissent dans ce cas le mode de gestion libre. 56,66 % des individus présents dans la base sont de sexe masculin.

### 4.2.2 Test d'indépendance

Après une description univariée des données, nous nous intéressons à la pertinence des variables explicatives dans notre base, c'est-à-dire pour savoir si elles ont une influence significative sur la variable d'intérêt qui n'est rien d'autre que la décision de faire un arbitrage ou non. Un test de khi-carré de Pearson ou test de khi 2 a été fait pour sélectionner les variables en amont du modèle. Ce test permet d'apprécier l'existence ou non d'une relation entre deux variables dans notre base. Pour que le test fonctionne, il faut qu'au moins, l'une des deux variables à tester soit qualitative. Notons que ce test ne permet pas de décrire le sens de la dépendance entre les deux variables, mais seulement l'existence de cette dépendance.

Le test de khi 2 permet de dire dans notre cas si la décision d'arbitrer ou non qui est attribuée à notre variable explicative (le mode de gestion par exemple) existe vraiment ou est



seulement le fruit du hasard. L'hypothèse nulle du test  $H_0$  est que les deux variables à tester sont indépendantes. Le niveau de significativité d'un test statistique est souvent exprimé par une p-value entre 0 et 1. Plus la valeur de la p-value est petite, plus on rejette l'hypothèse nulle.

Pour juger la significativité d'un test, il faut fixer a priori un seuil à partir duquel on peut dire si le test est significatif ou non. Le seuil standard utilisé en statistique est 0.05 donc une valeur de p-value inférieure à 5 % signifie qu'on a une évidence à 95 % de rejeter l'hypothèse nulle, c'est-à-dire que la probabilité que l'hypothèse soit vraie est inférieure à 5 %. Donc, nous rejetons l'hypothèse nulle au seuil de 5 %.

TABLEAU 4.2 – Tableau des tests de Khi 2

<b>Variables</b>	<b>p-value</b>
PM_1	< 2e-16
PART_UC_1	< 2e-16
NB_FONDS_UC	< 2e-16
AGE	< 2e-16
ANCIENNETE	< 2e-16
IND_PERIODICITE	< 2e-16
CD_SEXE	< 2e-16
ID_MOD_GESTION	< 2e-16
TMG	< 2e-16
TX_PB	< 2e-16
Ecart_cac	< 2e-16
Ecart_erstx	< 2e-16
Ecart_tme	< 2e-16
Ecart_tx10ans	< 2e-16
RDT_EUR	< 2e-16
RDT_UC	< 2e-16

Le tableau 4.2 montre que tous les tests de khi deux entre la variable réponse fréquence d'arbitrage et les variables explicatives sont significatifs au seuil de 5 %, donc nous avons une évidence à 95 % d'affirmer qu'il existe une relation de dépendance entre chaque variable explicative et la décision de réaliser ou non un arbitrage.

### 4.2.3 Analyse de la distribution des variables cibles

Nous allons maintenant analyser la distribution des variables à modéliser le taux d'arbitrage et la fréquence d'arbitrage des assurés pour toute année confondue et pour chaque année prise

séparément. La distribution du taux d'arbitrage présente en général une forme W-shaped et symétrique entre -1 et 1. Cette analyse permet de constater que la distribution du taux d'arbitrage change d'une année à l'autre, ce qui met en évidence l'importance de la conjoncture économique sur le montant à arbitrer des assurés.

Il est donc important de prendre en compte cet aspect comme nous l'avons mentionné dans la partie méthodologique. Nous observons aussi la présence d'une queue lourde des deux côtés pour chaque distribution, autrement dit, le nombre d'assurés qui change complètement la nature de leurs fonds est important chaque année, on ne saura pas dire a priori si cela est dû à la conjoncture économique ou bien plutôt un comportement structurel des assurés.

Le graphique des fréquences d'arbitrage montre plutôt une tendance stable, près de 14% des clients dans le portefeuille font des arbitrages chaque année. Et cette proportion ne change pas trop d'une année à une autre. Notons aussi que le nombre des clients qui réalisent des opérations d'arbitrage est très faible devant le nombre des clients qui ne le font pas, donc nous sommes en présence d'un problème de données déséquilibrées. Un traitement spécifique que nous allons détailler dans la partie modélisation des fréquences sera adopté pour surmonter cette difficulté.

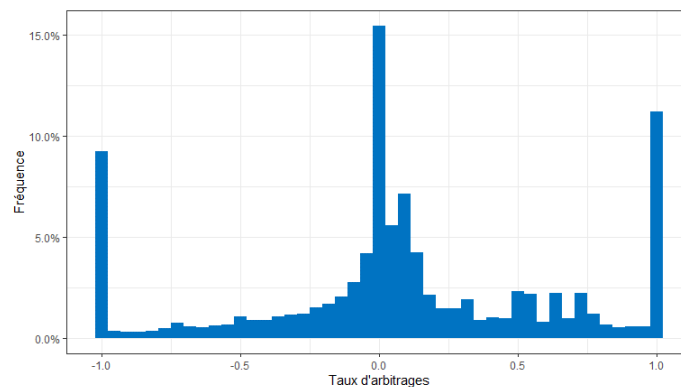


FIGURE 4.3 – Histogramme du taux d'arbitrage

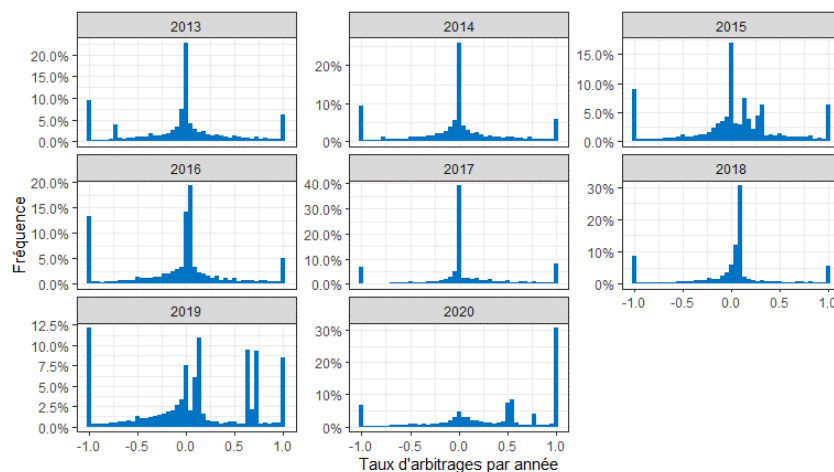


FIGURE 4.4 – Histogramme du taux d'arbitrage par année

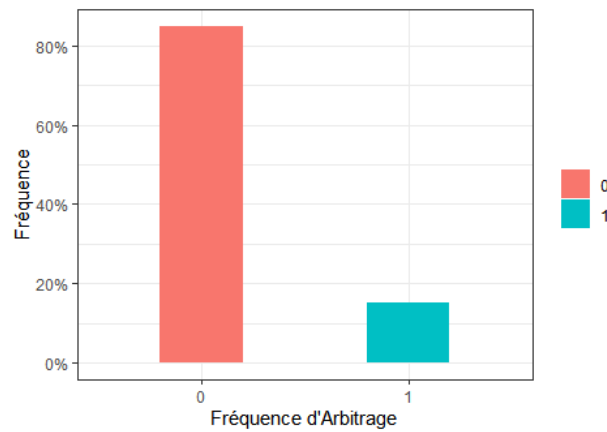


FIGURE 4.5 – Fréquence d’arbitrage

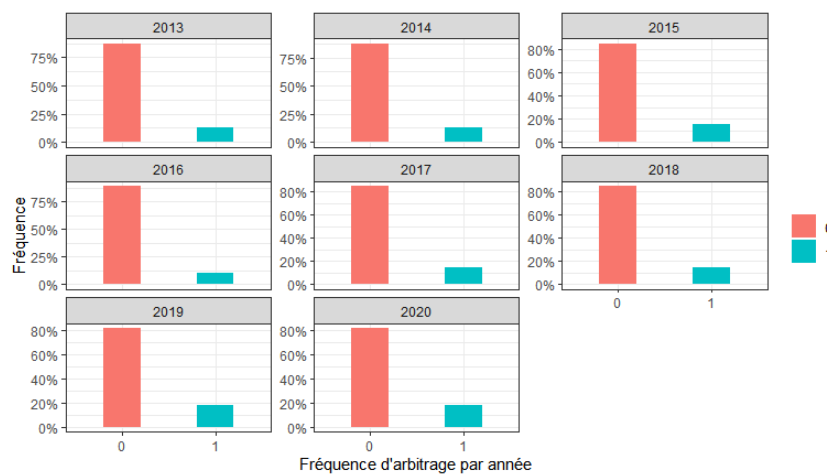


FIGURE 4.6 – Fréquence d’arbitrage par année

## 4.2.4 Analyse bivariée

### 4.2.4.1 Analyse des corrélations entre les variables explicatives

L’analyse des corrélations des variables permet de quantifier la relation de dépendance entre deux variables aléatoires et le sens de la dépendance si les deux variables sont continues. Lors de la mise en place d’un modèle statistique ou de machine learning, l’idéale serait d’avoir des variables explicatives décorréelées pour l’interprétabilité du modèle et pour sa convergence, ce qui n’est pas souvent le cas dans la réalité. La présence d’une forte corrélation entre les variables explicatives peut s’interpréter comme une redondance d’information pour le modèle ce qui peut donc induire un biais ou une non-convergence du modèle.

Le coefficient de corrélation de Pearson sera utilisé pour mesurer la dépendance entre les variables explicatives continues. Cependant, il n’est pas possible de quantifier le lien entre deux variables catégorielles par cette statistique. Le V de Cramer sera donc utilisé pour évaluer la corrélation entre ces variables, il est souvent utilisé dans le même cadre que la statistique de khi-deux pour mesurer la corrélation entre deux variables qualitatives et présente un avantage

de rester stable même si on augmente la taille de l'échantillon dans la même proportion d'intermodalités. Plus sa valeur est proche de 0, plus la relation de dépendance entre les deux variables est faible et vaut 1 en cas d'une parfaite dépendance.

Comme les entreprises qui composent l'indice CAC 40 font partie en majorité des entreprises composantes de l'indice Eurostoxx alors nous pouvons dire a priori que les deux variables peuvent présenter une relation de dépendance. De même pour le TME et le taux 10 ans, car le taux de 10 ans est utilisé pour déterminer le TME.

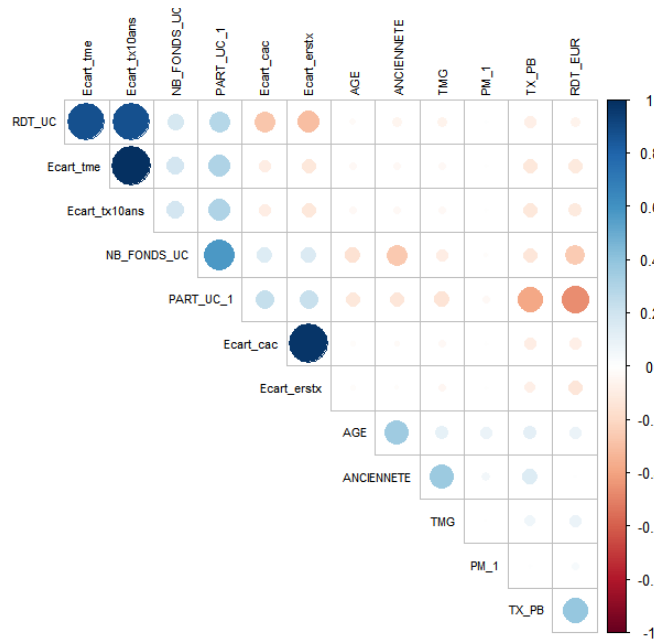


FIGURE 4.7 – Matrice de corrélation des variables continues

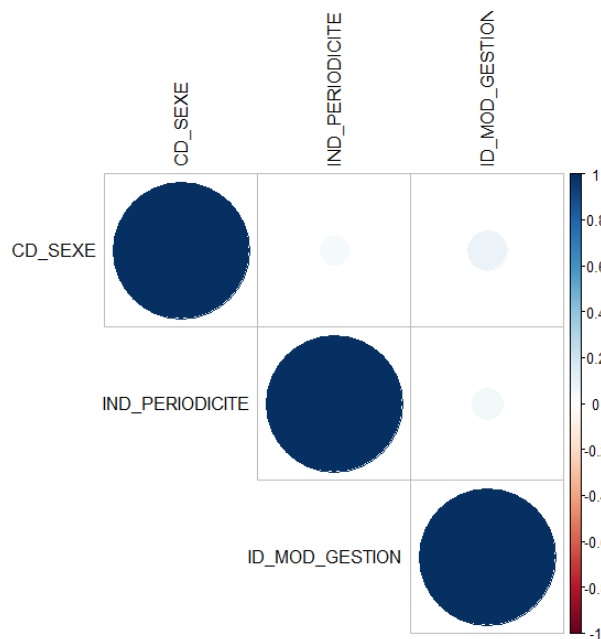


FIGURE 4.8 – Matrice de V de Cramer des variables qualitatives

Le graphique de la matrice de corrélation des variables continues et la matrice de V de

Cramer des variables qualitatives montrent qu'en globale les variables ne sont pas très corrélées entre elles. Nous remarquons aussi la présence d'une corrélation non négligeable entre la part en UC et le nombre de fonds en UC détenus dans le contrat. En effet, plus le nombre de fonds en UC détenu dans le contrat est élevé plus nous pensons que le montant investi dans le fonds en UC est important. La corrélation des variables sera prise en compte par le modèle lors de la sélection des variables explicatives à retenir pour prédire les variables cibles.

#### 4.2.4.2 Discrétisation des variables continues

Lors de la mise en place d'un modèle GLM, il est généralement conseillé de discrétiser les variables continues pour éviter la linéarité de leur effet sur la variable cible dans le modèle GLM. En effet, par construction, l'effet d'une variable explicative continue sur la variable dépendante est monotone, donc le GLM ne peut pas traduire un effet de changement de comportement à partir d'un certain seuil et un effet non-linéaire d'une variable continue. Par exemple : l'effet de l'âge sur le taux d'arbitrage sera croissant ou décroissant si nous gardons l'âge comme une variable continue dans notre modèle. Or la réalité montre en général que ce n'est pas le cas. Les jeunes assurés (les moins de 25 ans) et les personnes âgées (plus de 70 ans) arbitrent moins en termes de fréquence et en termes de taux d'arbitrage par rapport aux assurés entre 25 et 69 ans. Si nous gardons la variable continue alors le modèle ne saura pas identifier et prédire ce type de comportement.

Cependant, une discrétisation des variables continues au hasard aussi peut introduire des biais dans le modèle, car il se peut que certaines modalités ou classes créées disposent de moins d'individus ce qui peut rendre instable l'estimation des paramètres dans un modèle GLM. Un regroupement équiprobable, c'est-à-dire même nombre d'individus par classe peut créer un problème d'interprétation et de convergence des paramètres, car cette méthode peut diluer l'effet d'une variable par rapport à la variable cible.

Adrien Ehrhardt (2019) (1) propose une méthode pour trouver la meilleure discrétisation d'une variable en se basant sur la maximisation de l'AIC (Akaike Information Criterion). Sander Devriendt et al. (2021) (18) ont développé plutôt une méthode de discrétisation basée sur la pénalisation Lasso et ses variants. C'est cette dernière qui a été utilisée dans notre cas pour discrétiser les variables continues.

La discrétisation basée sur la pénalisation Lasso consiste à minimiser la fonction objectif suivante :

$$\mathcal{O}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) = f(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) + \lambda \cdot \sum_{j=0}^J g_j(\boldsymbol{\beta}_j)$$

où  $f()$  est le moins du log de vraisemblance divisé par la taille de l'échantillon,  $\mathbf{y}$  la variable cible,  $\mathbf{X}$  la matrice des variables explicatives,  $g_j(\cdot)$  une fonction convexe pour  $j$  de  $0, \dots, J$  et  $\boldsymbol{\beta}_j$  la valeur du coefficient associé à la variable  $x_j$  avec  $\boldsymbol{\beta}_0$  la constante. Comme la constante n'est

pas régularisée dans une régression Lasso alors  $g_j(0) = 0$ . La fonction de pénalité  $g_j(\cdot)$  permet d'éviter le sur-apprentissage des données d'entraînement alors que le paramètre  $\lambda$  permet de contrôler l'importance de la pénalité.

Il existe trois types de pénalisation en fonction de la nature et de la caractéristique de la variable explicative. En effet, pour une variable continue par exemple, il est important de savoir le nombre de classe optimal et la valeur prise par les bornes de l'intervalle de chaque classe.

— **Le Lasso :**

La pénalisation Lasso peut-être utilisée pour une variable catégorielle ou pour une variable continue et est donnée par :

$$g_{\text{Lasso}}(\boldsymbol{\beta}_j) = \sum_{i=1}^{p_j} w_{j,i} |\beta_{j,i}| = \|\mathbf{w}_j * \boldsymbol{\beta}_j\|_1$$

où  $p_j$  est le nombre de coefficient  $\beta_{j,i}$  qui se trouve dans le vecteur  $\boldsymbol{\beta}_j$  et  $\mathbf{w}_j$  un vecteur du poids des pénalités. En fonction de la valeur de  $\lambda$  et du vecteur de poids  $\mathbf{w}_j$ , la pénalisation va rendre la valeur de certains coefficients égale à 0 pour les enlever du modèle, ce qui réduit la variance du modèle, mais présente un risque d'augmentation du biais. Pour une variable continue, la pénalisation Lasso permet de sélectionner ou non la variable en lui attribuant un coefficient égal ou différent de 0. Pour les variables catégorielles, elle permet de sélectionner les modalités les plus déterminants.

— **Le groupe Lasso :**

La pénalisation Group Lasso utilise la norme  $L_2$  pour annuler certains coefficients  $\boldsymbol{\beta}_j$  du modèle :

$$g_{\text{grpLasso}}(\boldsymbol{\beta}_j) = w_j \sqrt{\sum_{i=1}^{p_j} \beta_{j,i}^2} = \|w_j \boldsymbol{\beta}_j\|_2$$

où  $w_j$  représente le vecteur de poids de pénalité de la variable  $x_j$ . Contrairement à la norme  $L_1$ , la norme  $L_2$  ne permet pas de séparer chaque coefficient dans  $\boldsymbol{\beta}_j$ . Si nous disposons d'une variable continue, c'est-à-dire un seul coefficient  $\boldsymbol{\beta}_j$ , alors le Group Lasso revient à la pénalisation Lasso. La pénalisation Group Lasso est adaptée pour tester si une variable explicative catégorielle est utile ou non pour la prédiction de la variable cible, car la valeur des estimateurs des coefficients  $\beta_{j,i}$  pour le vecteur  $\boldsymbol{\beta}_j$ , sont soit tous nuls ou tous non nuls. Elle est donc utile pour la sélection des variables catégorielles.

— **Le Fused Lasso :**

Pour regrouper les modalités consécutives d'une variable catégorielle ordinaire, c'est-à-dire que l'ordre des modalités a un sens, le Fused Lasso met une pénalité  $L_1$  sur la différence

de deux coefficients consécutifs de deux modalités consécutives d'une variable :

$$g_{\text{Lasso}}(\boldsymbol{\beta}_j) = \sum_{i=2}^{p_j} w_{j,i-1} |\beta_{j,i} - \beta_{j,i-1}| = \|D(\mathbf{w}_j) \boldsymbol{\beta}_j\|_1$$

Où  $D(\mathbf{w}_j)$  est la matrice de différence première des poids  $w_{j,i}$  :

$$D(\mathbf{w}_j) = \begin{bmatrix} -w_{j,1} & w_{j,1} & 0 & 0 & 0 \\ 0 & -w_{j,2} & w_{j,2} & \cdots & 0 & 0 \\ 0 & 0 & -w_{j,3} & 0 & 0 \\ & \vdots & & \ddots & w_{j,p_j-2} & 0 \\ 0 & 0 & 0 & -w_{j,p_j-1} & w_{j,p_j-1} \end{bmatrix}$$

Le Fused Lasso est adapté pour les variables catégorielles ordonnées et les variables continues transformées en variables catégorielles pour capter les effets non-linéaires de la variable. Comme c'est la différence entre les coefficients qui sera régularisée, alors nous avons besoin d'une modalité de référence pour avoir une unique valeur de  $\boldsymbol{\beta}$  qui minimise la fonction objectif. Donc le coefficient associé à la modalité de référence sera donc égal à 0 comme dans le cas d'une régression logistique ou d'une régression linéaire avec une variable catégorielle. Si deux modalités consécutives d'une variable catégorielle ordonnée ont le même effet sur la variable cible selon le modèle alors le Fused Lasso va rendre leur coefficient identique ce qui signifie que les deux modalités devraient être regroupées en une seule modalité. C'est la pénalisation Fused Lasso qui a été utilisée dans notre cas pour discrétiser les variables continues. L'idée est de discrétiser le plus fin possible les variables continues tout en restant raisonnable par rapport au temps de calcul et de la complexité que cela peut induire et laisser la pénalisation Fused Lasso regrouper les modalités qui ont la même caractéristique.

Les deux méthodes de discrétisation ont été appliquées sur notre jeu de données et nous avons retenu la méthode qui minimise l'AIC comme la meilleure approche pour discrétiser les variables continues présentes dans la base. Il en ressort que c'est l'approche qui utilise la pénalisation de type LASSO qui donne la meilleure discrétisation.

### 4.3 Découpage en échantillon d'entraînement et en échantillon test

Le choix de découpage de la base des données reste primordial en modélisation statistique, il doit dépendre de la nature, de la caractéristique de la variable à modéliser et de la manière

dont nous allons utiliser le modèle pour prédire les variables cibles futures. Dans notre cas, nous cherchons à prédire le taux d'arbitrages et la fréquence d'arbitrages annuels du portefeuille de l'assureur ce qui revient à prédire le taux d'arbitrages et la fréquence d'arbitrages des contrats présents dans le portefeuille l'année précédant l'année d'arbitrage. Pour cela, l'idéal sera donc d'avoir une observation des comportements d'arbitrages des assurés sur une année ou une période (série d'années consécutives) pour l'échantillon d'entraînement ainsi que des données tests qui correspondent à des opérations d'arbitrage au cours d'une année donnée ou une période donnée.

Les arbitrages observés entre 2013 et 2018 seront donc utilisés comme données d'entraînement et les arbitrages réalisés par les clients entre 2019 et 2020 comme données tests. Le choix de ces années n'a pas été fait aléatoirement, en effet, sur le marché financier, nous observons que le rendement du CAC en fin 2019 a atteint une valeur typiquement élevée et le comportement d'arbitrages en 2020 en dépend. De plus, l'année 2020 a été marquée par une crise sanitaire qui a créé une crise financière, donc prendre juste l'année 2020 revient à tester le modèle juste sur des situations atypiques qu'il n'a jamais vu sur l'échantillon d'apprentissage, ce qui peut fausser l'évaluation et l'appréciation de la qualité du modèle. D'où le choix d'inclure l'année 2019 dans les données tests.

## 4.4 Choix des métriques d'évaluation

Pour le choix du modèle, il est nécessaire de définir des indicateurs de performance pour juger lequel des modèles candidats est le plus adapté pour la variable cible et pour déterminer si les prédictions du modèle sont en adéquation avec les valeurs observées. Ces indicateurs seront utilisés pour choisir le meilleur modèle.

### 4.4.1 Métrique pour la fréquence

Étant donné que la modélisation de la fréquence des arbitrages est un problème de classification, alors nous allons utiliser des indicateurs qui utilisent la probabilité et l'appartenance à des classes de la variable à prédire.

- Le taux de bien classés : cette mesure permet d'apprécier la capacité globale du modèle à classifier la variable cible. Elle s'obtient en calculant le rapport entre le nombre de bonnes prédictions et le nombre total d'observations à prédire.

$$\text{Taux de bien classés} = \frac{\text{Vrai positif} + \text{Vrai négatif}}{\text{Total}}$$

Cet indicateur s'avère être insuffisant pour apprécier un modèle classification binaire si les incidences d'une mauvaise prédiction des deux classes ne sont pas identiques et si la répartition des deux classes est déséquilibrée.



- Le rappel ou la sensibilité : cet indicateur mesure la capacité du modèle à détecter les vrais positifs qui ne sont rien d'autre que les assurés qui n'ont pas réalisé des opérations d'arbitrage dans cette étude.

$$\text{sensibilité} = \frac{\text{Vrai Positif}}{\text{Vrai Positif} + \text{Faux Négatif}}$$

- La spécificité : appelée aussi le taux des vrais négatifs, cette mesure évalue la capacité à détecter tous les assurés qui ont effectué un arbitrage. C'est une mesure complémentaire de la sensibilité.

$$\text{spécificité} = \frac{\text{Vrai Négatif}}{\text{Vrai négatif} + \text{Faux Positif}}$$

- Le log loss : cette métrique utilise plutôt les probabilités pour évaluer les modèles. Elle augmente à mesure que la probabilité prédite par le modèle diffère de la vraie classe. Plus sa valeur est proche de 0, meilleure est la qualité du modèle.

$$\text{Logloss} = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$$

TABLEAU 4.3 – Matrice de confusion

		Valeur prédite		total
		0	1	
Vraie valeur	0	Vrai Positif	Faux Négatif	P'
	1	Faux Positif	Vrai Négatif	N'
total		P	N	

#### 4.4.2 Métrique pour le taux d'arbitrage

Une fois estimé, il est important de s'intéresser à la qualité du modèle. Ceci se fait à partir des indicateurs qui mesurent l'adéquation des prédictions du modèle  $\hat{y}_i$  aux vraies valeurs  $y_i$ .

On définit le résidu du modèle comme l'écart entre la valeur observée  $y_i$  et la valeur prédite  $\hat{y}_i$  :  $\varepsilon_i = y_i - \hat{y}_i$ . où  $\hat{y}_i$  est la valeur du taux d'arbitrage prédite par le modèle. Naturellement, si le résidu est proche de zéro, alors le modèle s'ajuste bien aux données.

##### Le coefficient d'ajustement

En utilisant le théorème de Pythagore, on a :  $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2$  où :

- $SCT = \sum_{i=1}^n (y_i - \bar{y})^2$  : représente la somme des carrés totale, il s'agit de la variabilité totale de  $y$  autour de sa moyenne  $\bar{y}$  ;
- $SCE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  est la somme des carrés expliqués, c'est-à-dire la variabilité de  $y$  expliqué par le modèle ;
- $SR = \sum_{i=1}^n \hat{\varepsilon}_i^2$  est la somme des carrées résiduelles, c'est-à-dire la variabilité de  $y$  que le modèle ne parvient pas à capter.

On définit le coefficient d'ajustement du modèle comme le rapport entre la variabilité expliquée par le modèle et la variabilité totale :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SR}{SCT}$$

Si  $R^2 = 1$ , le modèle explique tout, c'est à dire que  $y_i = f(x_{1,i}, \dots, x_{k,i})$  pour tout  $i$ . Plus généralement, si le  $R^2$  est proche de 1 alors le modèle s'ajuste bien aux données.

Si  $R^2 = 0$ , cela veut dire que  $SCE = 0$ , donc  $\hat{y}_i = \bar{y}$ , le modèle de régression est inadapté puisqu'on ne modélise rien de mieux que la moyenne ; plus généralement, si le  $R^2$  est proche de 0 alors les insuffisances du modèle sont grandes. Soit les variables explicatives sont inadéquates, soit la spécification du modèle n'est pas adaptée, c'est-à-dire que les variables  $X$  n'expliquent pas bien la variable réponse  $y$ .

### L'erreur quadratique moyenne

En anglais MSE ( Mean Square Error), c'est la moyenne arithmétique des carrés des résidus du modèle (entre les prévisions du modèle et les observations de la variable réponse). Si l'on compare deux estimateurs sans biais, le meilleur est celui qui présente la MSE la plus faible.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

D'autres mesures de précision du modèle peuvent être considérées, il s'agit notamment de :

- Root Mean Square Error (RMSE) :

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

- Mean Absolute Error (MAE) :

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

## IMPLÉMENTATION DES MODÈLES D'ARBITRAGE STRUCTUREL ET CONJONCTUREL

En statistique, notamment lors d'un problème de classification, il est toujours important de prendre en compte la distribution de la variable cible à modéliser. Dans notre étude, le nombre des contrats qui ont réalisé un arbitrage pour toutes années confondues et pour chaque année prise séparément est très faible devant le nombre des contrats qui n'ont pas effectué un arbitrage, il est de l'ordre 85 % contre 15 %. Nous sommes donc en présence d'un problème de données déséquilibrées ou « imbalanced data », la non-prise en compte de ce déséquilibre peut donc biaiser le modèle et détériorer ses performances. En effet, lors de la phase d'entraînement, le modèle n'apprend pas assez sur la classe minoritaire et aura donc du mal à classer cette dernière. L'idéal sera d'avoir des classes équilibrées lors de l'apprentissage ce qui n'est pas le cas de notre jeu de données. Nous étions contraints de ne travailler que sur 70 % de la base fréquence d'entraînement à cause de sa volumétrie. Pour surmonter le problème, plusieurs méthodes d'échantillonnage seront donc mises en œuvre pour modifier la répartition des classes dans la base d'apprentissage.

### — Undersampling :

La méthode d'Undersampling consiste à réduire le nombre d'observations de la classe majoritaire pour le ramener au nombre d'observations de la classe minoritaire, c'est-à-dire des assurés qui n'ont pas réalisé un arbitrage au cours de l'année dans la base d'entraînement pour que la répartition de la variable nombre d'arbitrages soit équilibrée. Elle permet aussi en même temps de réduire la taille de notre jeu de données et donc de réduire le temps pour la compilation du modèle.

Cependant, une suppression d'un nombre important de la classe majoritaire peut faire perdre des informations importantes à nos données d'apprentissage pour bien classer les contrats qui n'ont jamais réalisé d'arbitrage.

### — L'Oversampling :

Cette méthode s'applique sur la classe minoritaire, elle permet d'équilibrer le nombre des classes en répliquant ou bootstrapant les observations dans la classe minoritaire jusqu'à ce que les deux classes aient le même nombre d'observations. Elle augmente dans ce cas le nombre d'observations et le temps de calcul en même temps.

Aucune information sur la classe minoritaire ne sera donc perdue, mais elle risque de conduire au sur-apprentissage si des observations de la classe minoritaire se répètent plusieurs fois.

— **Un mélange des deux :**

Cette méthode a l'avantage de garder la taille initiale de la base d'apprentissage identique tout en diminuant le nombre de la classe majoritaire et en augmentant le nombre de la classe minoritaire. Elle a l'avantage de faire le compromis entre pertes d'information via la réduction de la classe majoritaire et sur-apprentissage à travers l'augmentation de la classe minoritaire.

## 5.1 Modèle GLM fréquence

Le premier modèle qui sera implémenté dans le cadre de cette étude est le modèle GLM fréquence d'arbitrage. Étant donné la nature de la variable cible nombre d'arbitrages ou décision à arbitrer ou non qui appartient dans l'ensemble  $\{0, 1\}$ , donc la seule famille candidate et plausible pour notre modèle GLM fréquence est la famille binomiale avec une fonction de lien logit qui n'est rien d'autre que la régression logistique. C'est l'approche de classification par excellence du modèle GLM et elle a l'avantage d'être facile à interpréter.

Le choix du meilleur modèle GLM fréquence se porte donc sur la méthode d'échantillonnage et non plus sur la famille de loi. En plus de la base d'apprentissage initiale, une base après undersampling et une base après mélange des deux méthodes seront testées pour déterminer la bonne base d'apprentissage qui permet de bien classifier la décision d'arbitrage. La méthode oversampling n'a pas pu être réalisée vu la taille de notre base de données et le temps de calcul que ça aurait pu générer. Il en ressort que c'est la régression logistique avec une base après mélange des deux méthodes qui a été retenue car elle présente des meilleures performances par rapport aux deux autres jeux de données.

Une sélection de variables a été aussi réalisée en utilisant l'approche backward du modèle GLM, il s'agit d'enlever à chaque étape la variable la moins significative et de regarder le critère d'information AIC. Il s'avère que c'est le modèle sans les variables rendement du fonds en euros et rendement du fonds en unités de compte qui présente la meilleure performance. Les résultats de ce modèle sont présentés dans le tableau ci-dessus :

TABLEAU 5.1 – Résultat GLM fréquence

	Estimate	Std. Error	t value	Pr(>  t )
(Intercept)	-3.835377	0.029044	-132.055	< 2e-16
IND_PERIODICITEU	0.001328	0.005246	0.253	0.800155
CD_SEXEM	0.192164	0.004138	46.438	< 2e-16
AGE(25,30]	0.333218	0.015459	21.555	< 2e-16
AGE(30,40]	0.228839	0.012081	18.942	< 2e-16
AGE(40,70]	0.145398	0.010983	13.239	< 2e-16
AGE(70,109]	-0.040067	0.012025	-3.332	0.000862
ANCIENNETE(4,8]	-0.112558	0.005362	-20.991	< 2e-16
ANCIENNETE(8,10]	-0.199815	0.006752	-29.594	< 2e-16
ANCIENNETE(10,13]	-0.272152	0.007193	-37.833	< 2e-16
ANCIENNETE(13,18.5]	-0.401382	0.008702	-46.125	< 2e-16
ANCIENNETE(18.5,46.1]	-0.913470	0.017629	-51.818	< 2e-16
NB_FONDS_UC1	1.434311	0.010265	139.731	< 2e-16
NB_FONDS_UC]1, 5]	1.821220	0.010500	173.455	< 2e-16
NB_FONDS_UC]5, 10]	2.062872	0.011962	172.459	< 2e-16
NB_FONDS_UC]10, 15]	2.267249	0.015094	150.204	< 2e-16
NB_FONDS_UC>15	2.774869	0.013140	211.183	< 2e-16
PM_1]4200 ; 10000]	0.327735	0.008669	37.806	< 2e-16
PM_1]10000 ; 20000]	0.709791	0.008349	85.015	< 2e-16
PM_1]20000 ; 50000]	1.026580	0.007787	131.834	< 2e-16
PM_1]50000 ; 85000]	1.184006	0.008626	137.262	< 2e-16
PM_1]85000 ; 150000]	1.229721	0.008927	137.748	< 2e-16
PM_1]150000 ; 200000]	1.297714	0.011653	111.362	< 2e-16
PM_1]200000 ; 500000]	1.425023	0.010212	139.542	< 2e-16
PM_1>500000	1.741993	0.013782	126.394	< 2e-16
PART_UC_1(0.1,0.2]	0.106956	0.011431	9.357	< 2e-16
PART_UC_1(0.2,0.35]	0.024237	0.011137	2.176	0.029538
PART_UC_1(0.35,0.5]	0.024390	0.013545	1.801	0.071746
PART_UC_1(0.5,0.6]	0.279568	0.014340	19.496	< 2e-16
PART_UC_1(0.6,0.75]	0.278976	0.013835	20.164	< 2e-16
PART_UC_1(0.75,0.85]	0.042425	0.015533	2.731	0.006309
PART_UC_1(0.85,0.95]	0.328555	0.015616	21.039	< 2e-16
PART_UC_1(0.95,1]	-0.557908	0.014294	-39.031	< 2e-16
TX_PB(0,0.85]	0.060687	0.006740	9.004	< 2e-16
TX_PB(0.85,0.95]	0.064995	0.007608	8.543	< 2e-16
TX_PB(0.95,1]	-0.047681	0.005330	-8.946	< 2e-16
TMG(0,0.01]	0.247167	0.010249	24.116	< 2e-16
TMG(0.01,0.02]	-0.547343	0.266317	-2.055	0.039857
TMG(0.02,0.03]	-1.777427	0.052187	-34.059	< 2e-16
TMG(0.03,0.1]	-0.665682	0.036698	-18.139	< 2e-16
ID_MOD_GESTIONLIBRE	-0.055409	0.014527	-3.814	0.000137
ID_MOD_GESTIONPILOTE	0.767018	0.015833	48.443	< 2e-16
Ecart_cac(-0.13,-0.05]	0.154475	0.011908	12.972	< 2e-16
Ecart_cac(-0.05,-0.02]	0.366693	0.017621	20.809	< 2e-16
Ecart_cac(-0.02,0]	0.598696	0.024515	24.422	< 2e-16
Ecart_cac(0,0.025]	0.636026	0.018932	33.595	< 2e-16
Ecart_cac(0.025,0.04]	0.623783	0.018135	34.396	< 2e-16
Ecart_cac(0.04,0.155]	-0.581059	0.111608	-5.206	1.93e-07
Ecart_tme(-0.01,0]	1.039249	0.031739	32.744	< 2e-16
Ecart_tme(0,0.03]	1.013443	0.040979	24.731	< 2e-16
Ecart_tme(0.03,0.06]	0.794582	0.041475	19.158	< 2e-16
Ecart_tme(0.06,0.1]	0.933603	0.045028	20.734	< 2e-16
Ecart_tme(0.1,0.248]	0.751634	0.071444	10.521	< 2e-16
Ecart_tx10ans(-0.01,0]	-0.522337	0.031958	-16.345	< 2e-16
Ecart_tx10ans(0,0.05]	-0.707534	0.041059	-17.232	< 2e-16
Ecart_tx10ans(0.05,0.08]	-0.424697	0.044022	-9.647	< 2e-16
Ecart_tx10ans(0.08,0.1]	-0.565618	0.045494	-12.433	< 2e-16
Ecart_tx10ans(0.1,0.12]	-0.197057	0.069971	-2.816	0.004858
Ecart_tx10ans(0.12,0.248]	-0.409932	0.071305	-5.749	8.98e-09
Ecart_erstx(-0.05,-0.02]	-0.089820	0.012852	-6.989	2.77e-12
Ecart_erstx(-0.02,0]	-0.293030	0.009934	-29.498	< 2e-16
Ecart_erstx(0,0.015]	-0.316584	0.013244	-23.904	< 2e-16
Ecart_erstx(0.015,0.196]	-0.314098	0.014300	-21.965	< 2e-16

L'un des avantages des modèles GLM est qu'ils sont faciles à interpréter contrairement aux autres modèles de machine learning. La fonction de lien s'applique à la combinaison linéaire de nos variables explicatives pour obtenir la variable cible qui n'est rien d'autre que la décision à arbitrer ou non à la fin de l'année. La fonction de lien utilisée pour la fréquence GLM est le logit, ce qui fait que plus le logarithme des probabilités d'arbitrer est élevé, plus nous pouvons dire que l'assuré va arbitrer au cours de l'année. Donc, les coefficients positifs signifient que la probabilité d'arbitrer augmente, tandis que des coefficients négatifs indiquent qu'elle diminue. Et comme nous avons discrétisé les variables continues, alors les coefficients estimés dans le tableau 5.1 s'interprètent de la manière suivante : si l'assuré a les caractéristiques de référence alors il y a une probabilité égale à 0.0211 de procéder à une opération d'arbitrage à la fin de l'année. Étant donné que tous nos prédicteurs sont des variables catégorielles alors interpréter une variable catégorielle à deux modalités et une variable qualitative à plusieurs modalités s'avère être suffisante, car nous pouvons reprendre la même interprétation pour les autres variables en changeant juste la valeur des coefficients et les modalités. Pour la variable indice de périodicité de versement des primes, lorsque l'indice de périodicité d'un assuré passe d'unique à périodique toutes choses égales par ailleurs, alors le logarithme népérien de la probabilité d'arbitrer au cours de l'année augmente de 0.029. Pour la variable montant de provision mathématique, une augmentation de la provision mathématique de la tranche de ]10000 , 20000] à ]20000 , 50000] augmente la probabilité d'arbitrer de 37.27% si toutes les autres variables explicatives restent inchangées ( $\exp(1.0266 - 0.7098) = 1.3727$ ).

L'analyse des effets marginaux de chaque variable explicative permet aussi d'évaluer la significativité et d'interpréter les paramètres d'un modèle GLM. Elle a l'avantage d'être facile à interpréter, mais a l'inconvénient d'avoir plusieurs méthodes d'estimation qui peuvent parfois conduire à des résultats différents. Il existe deux méthodes pour calculer l'effet marginal d'une variable, la première consiste à calculer la moyenne de l'effet marginal de tous les individus et la deuxième consiste à calculer l'effet marginal de l'individu moyen. Nous retenons la première définition dans cette étude. L'effet marginal d'une variable catégorielle binaire  $x_1$  sur  $P(y = 1 | x)$  s'obtient en faisant la différence des deux probabilités suivantes :

$$\frac{1}{n} \sum_{i=1}^n [P_i(y = 1 | \{x_{i,1} = 1, x_{i,2}, \dots, x_{i,k}\}) - P_i(y = 1 | \{x_{i,1} = 0, x_{i,2}, \dots, x_{i,k}\})]$$

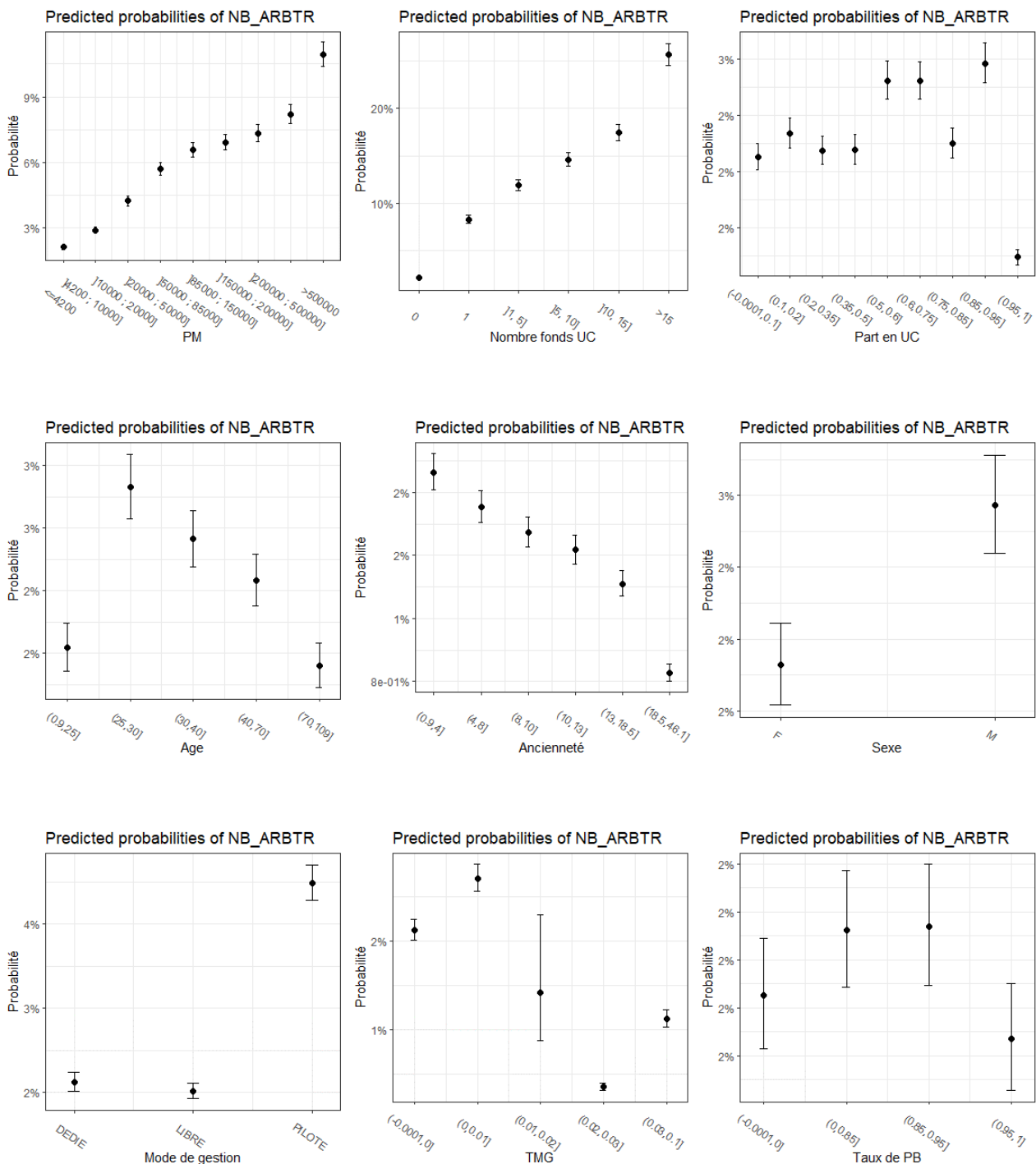
$$\frac{1}{n} \sum_{i=1}^n \left[ \frac{1}{1 + e^{-(\beta_0 + \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_K x_{Ki})}} - \frac{1}{1 + e^{-(\beta_0 + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_K x_{Ki})}} \right]$$

Et il s'étend de la même manière pour une variable catégorielle à plus de deux modalités :

$$\Delta_2 = \frac{1}{n} \sum_{i=1}^n [P_i(y = 1 | \{x_{i,1} = 2, x_{i,2}, \dots, x_{i,k}\}) - P_i(y = 1 | \{x_{i,1} = 0, x_{i,2}, \dots, x_{i,k}\})]$$

$$\Delta_3 = \frac{1}{n} \sum_{i=1}^n [P_i(y = 1 | \{x_{i,1} = 3, x_{i,2}, \dots, x_{i,k}\}) - P_i(y = 1 | \{x_{i,1} = 0, x_{i,2}, \dots, x_{i,k}\})]$$

La figure 5.1 montre les effets marginaux des variables explicatives dans la meilleure GLM fréquence. Elle montre que la probabilité d'arbitrer augmente lorsque le montant de la PM de l'assuré augmente. Par exemple, la probabilité qu'un assuré qui possède une PM entre ]10000 , 20000] est de l'ordre de 3 % et cette valeur atteint l'ordre de 6 % lorsque le montant de la PM de l'assuré appartient à la tranche ]50000 , 85000], donc si le montant de la PM de l'assuré passe de la classe ]10000 , 20000] à la classe ]50000 , 85000] alors la probabilité que l'assuré réalise un arbitrage augmente d'environ 3 %, toutes choses égales par ailleurs. Il est important aussi de prendre en compte les écarts-types des effets marginaux pour juger la précision de ces derniers en calculant leur intervalle de confiance.



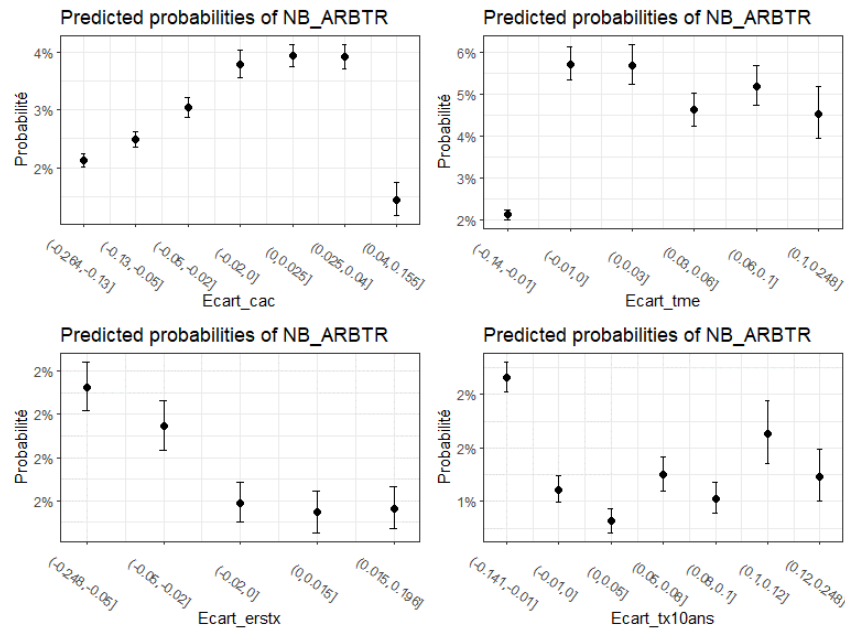


FIGURE 5.1 – Effets marginaux pour la fréquence GLM

Notons aussi que le choix du seuil de classification permet aussi de choisir le degré de prudence de notre modèle. En effet, étant donné la forte dominance de la classe d'assurés qui n'ont pas fait d'arbitrage dans notre jeu de données d'entraînement alors le modèle aura tendance à bien classifier cette classe si on choisit un seuil de classification à 0.5. Une diminution raisonnable de ce seuil permettrait donc de rester prudent pour ne pas sous-estimer le nombre des clients qui vont réaliser des arbitrages à l'avenir. En revanche, si on a tendance à trop diminuer ce seuil pour bien prédire la classe des assurés qui vont arbitrer alors la classe majoritaire sera sous-estimée. Néanmoins, l'avantage de l'approche fréquence/sévérité est qu'elle utilise la probabilité d'arbitrer au lieu de l'appartenance à la classe lors de la prédiction.

L'analyse de l'importance des variables permet d'identifier les variables qui contribuent les plus à la construction de notre modèle. La méthode la plus courante pour mesurer l'importance des variables pour le modèle GLM est la t-statistique. En effet, elle permet d'apprécier et de statuer sur la significativité des coefficients pour évaluer l'importance de la variable dans le modèle. Comme certaines variables dans la base fréquence ont plusieurs modalités, alors nous avons décidé de prendre la moyenne de la t-statistique comme mesure de l'importance de la variable.



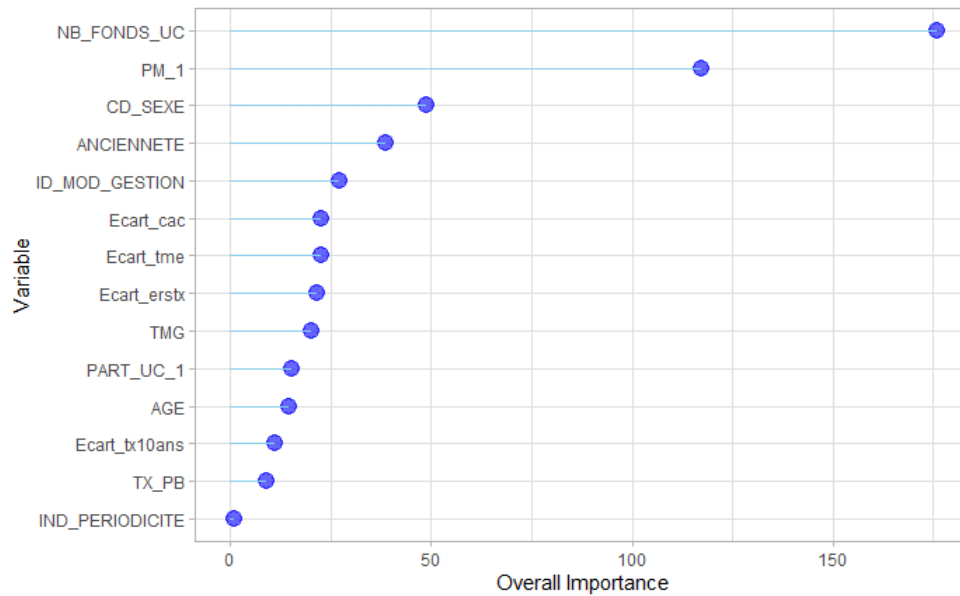


FIGURE 5.2 – Importance des variables dans le modèle GLM Fréquence

La figure 6.2 montre que c'est le montant de la provision mathématique et le nombre de fonds en UC détenu par l'assuré qui sont les variables qui contribuent le plus à la construction de notre modèle. Le taux de PB et l'indice de périodicité ont une contribution faible à la décision d'arbitrer ou non des assurés.

## 5.2 Modèle XGBoost Fréquence

Le deuxième modèle à implémenter pour modéliser la fréquence d'arbitrage dans cette étude est le modèle XGBoost. C'est un modèle très connu pour ses performances du point de vue de la métrique d'évaluation des modèles pour les problèmes de classification et du point de vue de l'implémentation du modèle sur le logiciel R dans notre cas. De plus, les divers paramètres du modèle permettent d'avoir un contrôle sur le modèle pour apprendre sur les données afin d'éviter les problèmes de sur-apprentissage et les problèmes de compilation. Le Xgboost a aussi l'avantage d'être capable de prendre en compte les effets non linéaires d'une variable continue, donc ne nécessite pas une discrétisation à l'entrée des variables continues.

De la même façon que pour le modèle GLM fréquence, nous avons testé le XGBoost sur les 3 méthodes d'échantillonnage et il s'avère que c'est le XGBoost avec la base après mélange des deux méthodes avec une proportion de 30% de la classe des assurés qui ont réalisé des arbitrages qui présente les meilleures performances. La présence des variables rendement du fonds en euro et du rendement du fonds en UC ont perturbé l'estimation de notre modèle donc nous avons décidé de les exclure pour la modélisation de la fréquence. Notons aussi que nous avons utilisé les paramètres par défaut de la fonction XGBoost de R pour le choix de la base des données.

Pour trouver le meilleur modèle XGBoost pour prédire la décision d'arbitrer ou non, il s'avère nécessaire de trouver les paramètres optimaux, car une variation de la valeur de ces derniers

aura des impacts différents sur le résultat du modèle. L'idéal sera donc de donner un ensemble de valeur pour tous les paramètres et de comparer les résultats pour toutes les combinaisons possibles de ces paramètres par validation croisée. Mais vu la taille de notre base de fréquence, il est impossible de réaliser une telle opération. Nous avons donc décidé de chercher deux par deux les paramètres optimaux en fixant les autres paramètres et faisant varier la valeur de deux paramètres donnés pour trouver la bonne valeur qui optimise notre modèle. Nous étions aussi limités au niveau du nombre des valeurs candidates vu la taille de la base des données. Nous cherchons donc à comparer les résultats obtenus par validation croisée pour choisir les paramètres optimaux. L'objectif est d'entraîner le modèle afin qu'il apprenne seulement sur les variables explicatives pour prédire la variable cible et non sur les bruits.

La fonction "xgboost" de R contient plus de 30 paramètres à ajuster, mais nous avons décidé d'ajuster juste 6 paramètres pour des raisons de complexité et de temps :

- `eta` : le learning rate ou le taux d'apprentissage. Il doit être compris entre 0 et 1. Il permet de contrôler comment les informations qui proviennent des nouveaux arbres peuvent être utilisées dans le processus de boosting. Plus il est proche de 1, plus les informations sur les nouveaux arbres sont importantes pour la construction du modèle. Une grande valeur d'`eta` implique dans ce cas une vitesse de convergence rapide du modèle et peut aussi conduire à un sur-apprentissage des données d'entraînement.
- `colsample_bytree` : le ratio de sous-échantillonnage des colonnes lors de la construction d'un arbre. Les colonnes seront sous-échantillonnées pour la construction d'un nouvel arbre.
- `max_depth` : le nombre maximum de profondeurs des arbres. Les arbres très profonds peuvent conduire à un sur-apprentissage alors que les arbres peu profonds peuvent conduire aussi à un problème de sous-apprentissage des données.
- `n_rounds` : Le nombre d'itération du boosting.
- `sub_sample` : il permet de déterminer si le modèle fait du pur boosting ou du boosting stochastique. En effet, si sa valeur est égale à 1 alors nous sommes dans le cas d'un pur boosting, c'est-à-dire que l'algorithme utilisera tout notre jeu de données pour faire grandir l'arbre alors que si sa valeur est entre 0 et 1, l'algorithme ne va sélectionner qu'une partie de la base. Il permet en général de résoudre le problème des valeurs aberrantes, car elles seront généralement supprimées sur la plupart des échantillons sélectionnés lors de la phase d'apprentissage.
- `min_child_weight` : le nombre minimal d'observation dans un nœud terminal. Une limitation de ce nombre peut conduire à un problème de sous-apprentissage.

Nous allons commencer par analyser les paramètres taux d'apprentissage `eta` et le nombre d'itérations. A priori, une valeur élevée du taux d'apprentissage et du nombre d'itérations peut conduire à un sur-apprentissage donc nous avons décidé de nous limiter à des petites valeurs d'`eta` et de `n_rounds` comme valeurs candidates. Le paramètre par défaut d'`eta` dans

la fonction "xgboost" de R est égale à 0.3 et 100 pour `n_rounds`. Les valeurs candidates d'`eta` à tester sont : {0.01, 0.05, 0.1, 0.3, 0.5} et celles de `n_rounds` sont {80, 90, 100, 110, 120} et les autres paramètres sont fixés par défaut. Nous utilisons le log loss, le taux de bien classés et la spécificité comme critère de choix des paramètres optimaux. Il en ressort que `n_rounds` = 100 et `eta` = 0.5 qui optimisent notre critère.

Pour la suite, nous allons déterminer la valeur optimale de `max_depth` et de `colsample_bytree`. Pour s'y faire, la valeur d'`eta` et la valeur de `n_rounds` seront remplacées par leur valeur optimale, nous faisons varier la valeur de `max_depth` ({4, 5, 6, 7, 8}) et de `colsample_bytree` ({0.25, 0.5, 0.75, 1}) et les autres paramètres gardent leur valeur par défaut. Après validation croisée, nous obtenons `max_depth` = 6 et `colsample_bytree` = 1. De la même manière, nous obtenons `sub_sample` = 1 et `min_child_weight` = 1.

Pour mieux comprendre le modèle XGBoost fréquence, le calcul des différentes mesures d'importance du XGBoost a été réalisé. En effet, cette analyse permet de déterminer pourquoi un assuré a réalisé ou non un arbitrage et de mesurer la contribution d'une variable à cette décision d'un point de vue individuelle et d'un point de vue global.

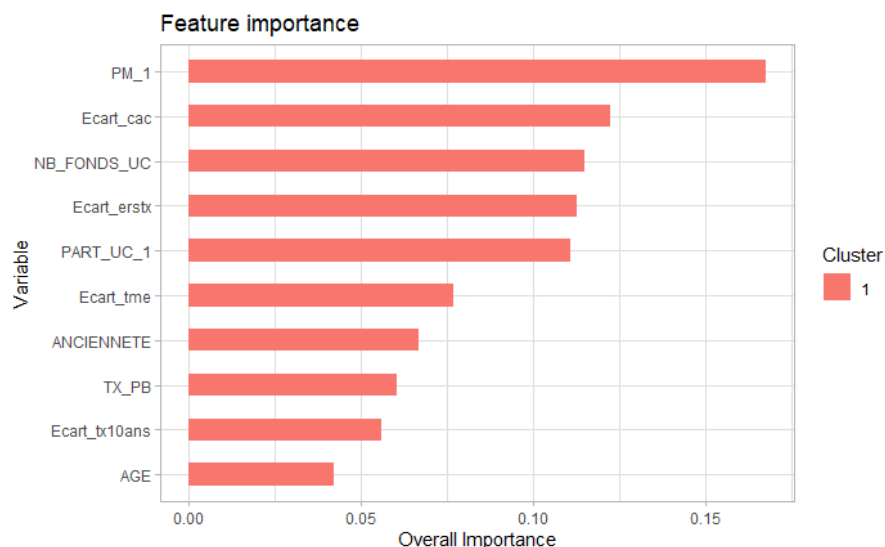
Nous commençons par analyser l'importance globale des variables à travers l'analyse des contributions globale de chaque variable "**gain**"<sup>1</sup>, "**weight**"<sup>2</sup> et "**cover**"<sup>3</sup>. Le graphique ?? montre que c'est la provision mathématique et le nombre de fonds en UC qui sont les variables les plus utilisées pour séparer les données à travers l'ensemble des arbres construits par le modèle. Et ces deux mêmes variables sont les variables qui réduisent les plus notre fonction de coût (logloss ici) quand elles sont utilisées pour une division au niveau d'un nœud.

---

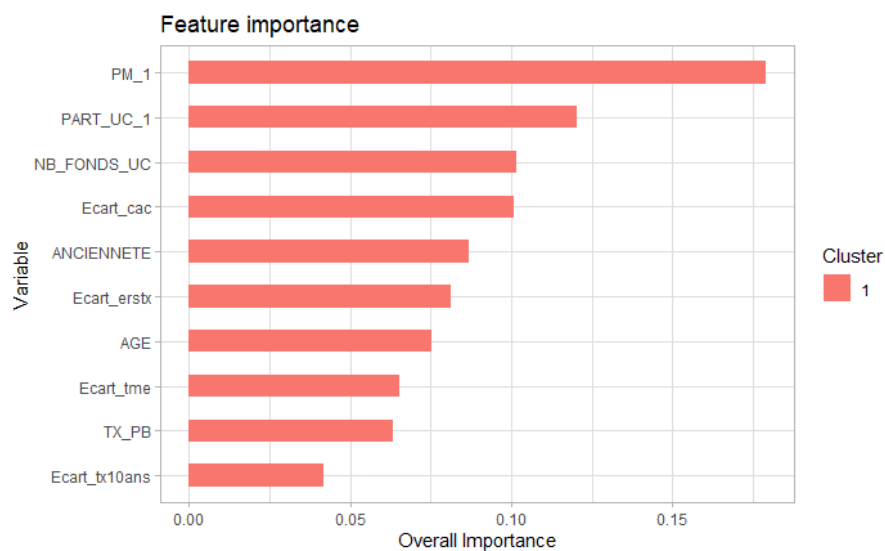
1. gain : Il s'agit de la moyenne de la réduction de la fonction coût (pour les données d'entraînement) quand une variable est utilisée pour une division au niveau d'un nœud.

2. weight : Il s'agit du nombre de fois où une variable est utilisée pour diviser les données à travers l'ensemble des arbres.

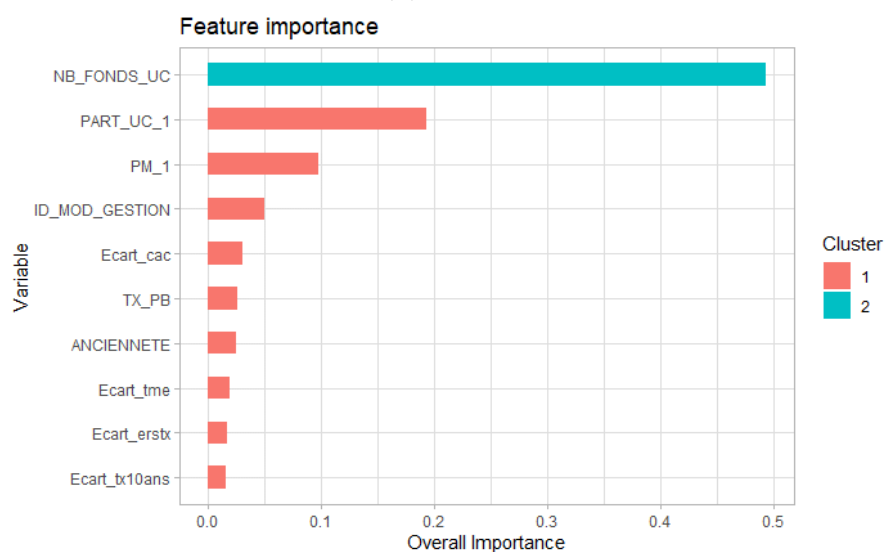
3. cover : Il s'agit du nombre de fois où une variable est utilisée pour diviser les données à travers l'ensemble des arbres, pondéré par le nombre de données (d'entraînement) passées par ce nœud. On peut en effet avoir des nœuds créés lors de l'entraînement, mais où presque aucune donnée ne circule : on considère donc que l'influence de ce nœud est moindre.



(a) Cover



(b) Weigh



(c) Gain

FIGURE 5.3 – Importance des variables dans le modèle XGBoost Fréquence

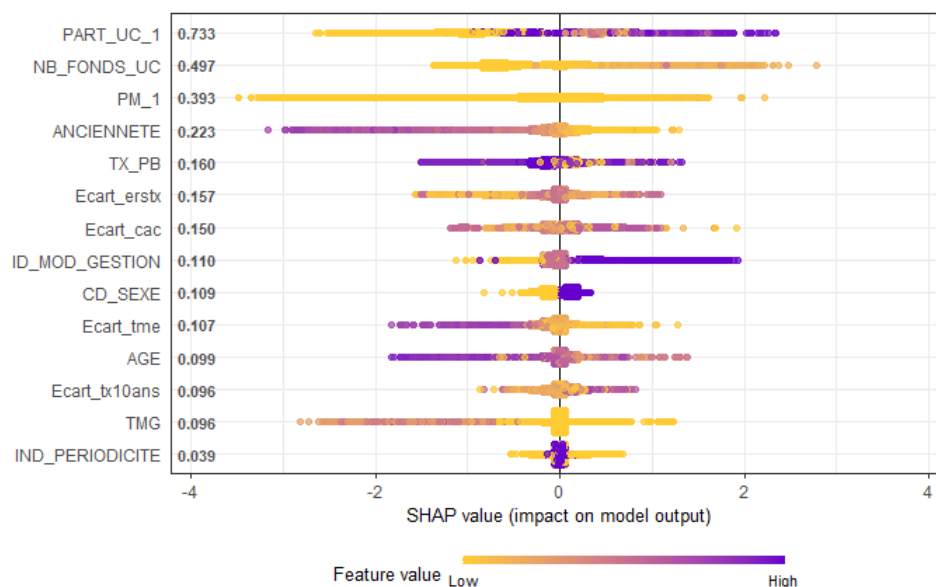


FIGURE 5.4 – Importance des variables par la méthode SHAP

L'objectif de l'explication des modèles par la méthode SHAP est d'expliquer la valeur de la prédiction du taux d'arbitrage par la somme des contributions des variables explicatives à la prédiction. Pour avoir une vision globale de l'importance des variables explicatives par la méthode shape pour le meilleur modèle XGBoost (XGBoost mélange, nous justifierons le choix de ce modèle plus tard), une présentation de la valeur des shapes de chaque individu sur chaque variable permet de constater la distribution des shapes et sa valeur en moyenne. Chaque point de la figure 5.4 représente la valeur SHAP de chaque individu sur chaque variable explicative. Ici pour des raisons de volumétrie, nous avons décidé de ne représenter que 100 000 individus de la base, mais le résultat reste le même pour tous les assurés de la base. La couleur représente la valeur du shap en allant du plus important au moins important. L'importance d'une variable est obtenue en calculant la moyenne de la valeur absolue des valeurs SHAP des individus sur la variable. La figure 5.4 montre que c'est la part en UC du contrat de l'assuré qui est la variable la plus importante suivie du nombre de fonds en UC dans le contrat et le montant de la PM. Le modèle XGboost mélange stipule que c'est l'aversion au risque de l'assuré et le montant de sa richesse qui sont les facteurs les plus déterminants dans sa décision à arbitrer ou non.

TABLEAU 5.2 – Performances des modèles

Modèles	GLM mélange	GLM normal	XGBoost normal	XGBoost mélange
Taux de bien classés	82.24 %	86.35 %	85.89 %	83.35 %
Sensitivité	87.36 %	98.55%	96.53 %	91.57 %
Spécificité	57.98 %	27.28%	34.80 %	43.91 %
LogLoss	0.3792	0.3467	0.3489	0.3727

Étant donné que la base et la nature des variables utilisées dans les modèles GLM et les modèles XGboost sont différentes, alors la comparaison de ces modèles se feraient donc sur les données tests. Ainsi, tous les résultats de performance du tableau sont calculés sur les données tests, à savoir le nombre d'arbitrages ou la fréquence d'arbitrage entre 2019 et 2020. Sur la base du taux de bien classés et du log loss, le modèle GLM avec une base d'apprentissage normal (sans undersampling et oversampling) qui présente la meilleure performance par rapport aux autres modèles. Cependant, c'est le modèle qui sous-estime le plus la classe des assurés qui arbitrent au cours de l'année, donc ce n'est pas un modèle très prudent pour prédire les arbitrages. Le modèle GLM avec une base mélangée (oversampling et undersampling) présente la meilleure performance en termes de spécificité, mais son taux de bon classement (82.24 %) est inférieur au taux de la classe positive (des assurés qui n'ont pas réalisés des arbitrages) des données tests qui est égal à 82.76 %, donc ce modèle ne fera pas mieux qu'un modèle qui ne prédit que de classe positive en termes de bon classement. Comme le XGboost avec une base d'apprentissage mélangée présente une meilleure spécificité et les autres indicateurs sont comparables à ceux du XGboost avec une base normale, le modèle XGboost mélange sera le modèle jugé le plus en adéquation avec notre base des données et notre objectif.

### 5.3 Régression en deux étapes du taux d'arbitrage

La méthode de régression en deux étapes, certes, présente une propriété mathématique intéressante qui permet de séparer les arbitrages structurels et conjoncturels, mais ne permet pas de garantir que le taux d'arbitrage sera compris entre -1 et 1. Les résultats de ce modèle ne sont pas très satisfaisants donc ne sera présentés dans cette étude.

### 5.4 Modèle GLM taux d'arbitrage

Pour l'implémentation du GLM taux d'arbitrage, la méthodologie diffère un peu du modèle GLM fréquence. En premier lieu, nous avons appliqué la régression tanh sur la variable cible taux d'arbitrage en fonction des variables structurelles pour obtenir le taux d'arbitrage structurel. Ensuite, le résidu de ce modèle sera considéré comme le taux d'arbitrage conjoncturel et sera appréhendé par une régression linéaire (car rien ne garantit plus que le résidu devrait être compris entre -1 et 1) sur les variables conjoncturelles de notre base.

#### 5.4.1 Modèle GLM taux d'arbitrage structurel

De même que pour la régression fréquence, nous avons discrétisé les variables explicatives continues par la méthode de discrétisation Lasso en fonction du taux d'arbitrage. Nous mettons en place un modèle GLM tanh (famille normale et fonction de lien  $\operatorname{arctanh}(x)$ ) et nous avons

appliqué une méthode backward pour la sélection des variables du modèle. C'est le critère AIC qui était utilisé pour choisir de retenir ou d'exclure une variable. Les variables utilisées pour le modèle GLM structurel sont : PM\_1, Part\_UC\_1,...Finalement, le modèle retenu est le modèle qui utilise toutes les variables structurelles. Le résultat de ce modèle est donné dans le tableau suivant :

TABLEAU 5.3 – Résultat GLM structurel

	Estimate	Std. Error	t value	Pr(>  t )
(Intercept)	0.3038	0.0100	30.46	< 2e-16
IND_PERIODICITEU	-0.0177	0.0019	-9.41	< 2e-16
CD_SEXEM	0.0111	0.0015	7.29	3.19e-13
AGE(25,40]	-0.0570	0.0046	-12.42	< 2e-16
AGE(40,50]	-0.0380	0.0045	-8.36	< 2e-16
AGE(50,101]	-0.0441	0.0044	-10.13	< 2e-16
ANCIENNETE(8,10]	-0.0146	0.0022	-6.61	3.94e-11
ANCIENNETE(10,18.5]	-0.0468	0.0021	-22.83	< 2e-16
ANCIENNETE(18.5,38.1]	-0.0380	0.0069	-5.52	3.35e-08
NB_FONDS_UC1	-0.5086	0.0048	-106.54	< 2e-16
NB_FONDS_UC]1, 5]	-0.3692	0.0048	-77.41	< 2e-16
NB_FONDS_UC]5, 10]	-0.3469	0.0052	-66.97	< 2e-16
NB_FONDS_UC]10, 15]	-0.5294	0.0057	-93.10	< 2e-16
NB_FONDS_UC>15	-0.6857	0.0054	-126.96	< 2e-16
PM_1]4200 ; 85000]	-0.0502	0.0029	-17.31	< 2e-16
PM_1]85000 ; 200000]	-0.0233	0.0034	-6.92	4.65e-12
PM_1]200000 ; 500000]	0.0098	0.0039	2.49	0.0128
PM_1>500000	0.0386	0.0049	7.85	4.25e-15
PART_UC_1(0.1,0.2]	0.0299	0.0047	6.39	1.64e-10
PART_UC_1(0.2,0.35]	0.0558	0.0043	13.09	< 2e-16
PART_UC_1(0.35,0.5]	0.0757	0.0044	17.19	< 2e-16
PART_UC_1(0.5,0.6]	0.1291	0.0046	28.18	< 2e-16
PART_UC_1(0.6,0.75]	0.0648	0.0043	15.12	< 2e-16
PART_UC_1(0.75,0.85]	0.1040	0.0045	23.06	< 2e-16
PART_UC_1(0.85,0.9]	0.0850	0.0047	18.17	< 2e-16
PART_UC_1(0.9,0.95]	0.1312	0.0052	25.14	< 2e-16
PART_UC_1(0.95,1]	0.0698	0.0043	16.31	< 2e-16
TX_PB(0,0.85]	0.0969	0.0022	44.82	< 2e-16
TX_PB(0.85,0.95]	0.0618	0.0029	21.65	< 2e-16
TX_PB(0.95,1]	0.0122	0.0019	6.36	2.08e-10
TMG(0,0.01]	0.1135	0.0042	26.75	< 2e-16
TMG(0.01,0.0451]	-0.0491	0.0148	-3.32	0.0009
ID_MOD_GESTIONLIBRE	-0.0401	0.0082	-4.91	9.02e-07
ID_MOD_GESTIONPILOTE	0.4411	0.0085	51.88	< 2e-16

L'importance des variables est mesurée par la t-statistique des coefficients du modèle dans le



tableau 5.3. Notons aussi que les coefficients n'ont pas de signification particulière comme dans le cas des autres GLM classiques. Étant donné que toutes les variables sont catégorielles, pour avoir une idée sur l'importance globale de chaque variable explicative, il faudra calculer plutôt la moyenne de la valeur absolue des t-statistiques des différentes modalités d'une variable. Il en ressort que c'est le nombre de fonds en UC qui est la variable la plus déterminante pour la détermination du taux d'arbitrage structurel.

### 5.4.2 Modèle GLM taux d'arbitrage conjoncturel

Comme la valeur taux d'arbitrage conjoncturel du modèle est égale à la différence entre le taux d'arbitrage observé et le taux d'arbitrage structurel du modèle, a priori nous ne pouvons pas dire que sa valeur doit être comprise entre -1 et 1. Par exemple, si la valeur observée est égale à 0.5 et la valeur de taux d'arbitrage structurel prédit par le modèle est égale à -0.55 alors la valeur du taux d'arbitrage conjoncturel est  $0.5 - (-0.55) = 1.05$ . Donc le modèle GLM qui s'adapte à notre variable réponse est un modèle GLM de famille normale avec une fonction de lien identité qui n'est rien d'autres que la régression linéaire. De même que pour le GLM taux d'arbitrage structurel, nous discrétisons les variables explicatives pour que notre modèle puisse tenir compte des effets non-linéaires de chaque variable. Nous utilisons uniquement les variables explicatives conjoncturelles dans cette partie (Ecart\_cac, Ecart\_tme, RDT\_EUR, RDT\_UC, Ecart\_erstx et Ecart\_tx10ans). Une approche backward (critère AIC) a été réalisée sur le modèle pour sélectionner les variables pertinentes et il s'avère que le rendement des fonds en euros du contrat n'a pas été retenu par le modèle. Le tableau 5.4 montre le résultat du modèle après sélection des variables :

TABLEAU 5.4 – Résultat GLM conjoncturel

	Estimate	Std. Error	t value	Pr(>  t )
(Intercept)	-0.0209	0.0168	-1.25	0.2117
Ecart_erstx(-0.15,-0.12]	0.1924	0.0910	2.12	0.0344
Ecart_erstx(-0.12,-0.05]	0.1991	0.0911	2.18	0.0289
Ecart_erstx(-0.05,0]	0.1823	0.0912	2.00	0.0456
Ecart_erstx(0,0.015]	0.1806	0.0912	1.98	0.0477
Ecart_erstx(0.015,0.02]	0.2090	0.0912	2.29	0.0220
Ecart_erstx(0.02,0.17]	0.2237	0.0922	2.43	0.0152
Ecart_tme(-0.01,0]	0.0099	0.0102	0.97	0.3321
Ecart_tme(0,0.03]	-0.0862	0.0121	-7.10	1.28e-12
Ecart_tme(0.03,0.06]	-0.0899	0.0125	-7.17	7.60e-13
Ecart_tme(0.06,0.1]	-0.0896	0.0132	-6.79	1.09e-11
Ecart_tme(0.1,0.14]	-0.1973	0.0242	-8.16	3.32e-16
Ecart_tme(0.14,0.248]	0.1707	0.0946	1.80	0.0713
Ecart_tx10ans(-0.01,0]	0.0217	0.0088	2.47	0.0135
Ecart_tx10ans(0,0.05]	0.1327	0.0113	11.76	< 2e-16
Ecart_tx10ans(0.05,0.08]	0.1602	0.0116	13.78	< 2e-16
Ecart_tx10ans(0.08,0.1]	0.1421	0.0127	11.22	< 2e-16
Ecart_tx10ans(0.1,0.14]	0.2420	0.0237	10.20	< 2e-16
Ecart_tx10ans(0.14,0.248]	-0.0534	0.0946	-0.56	0.5723
Ecart_cac(-0.15,-0.1]	-0.2530	0.0897	-2.82	0.0048
Ecart_cac(-0.1,-0.05]	-0.2557	0.0899	-2.85	0.0044
Ecart_cac(-0.05,-0.02]	-0.2369	0.0900	-2.63	0.0085
Ecart_cac(-0.02,0]	-0.3162	0.0901	-3.51	0.0005
Ecart_cac(0,0.02]	-0.1270	0.0900	-1.41	0.1582
Ecart_cac(0.02,0.04]	-0.2252	0.0900	-2.50	0.0123
Ecart_cac(0.04,0.136]	0.1700	0.1072	1.59	0.1127
RDT_UC(0,0.03]	0.0771	0.0043	18.05	< 2e-16
RDT_UC(0.03,0.1]	-0.0130	0.0039	-3.34	0.0008
RDT_UC(0.1,0.249]	0.0238	0.0046	5.18	2.27e-07

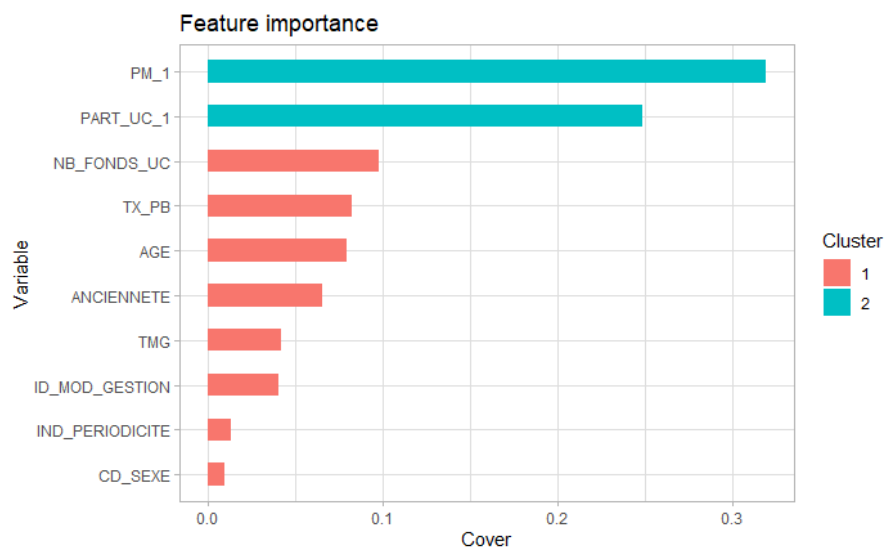
Les t-statistiques montrent que c'est le rendement du fonds en unités de compte qui influe le plus sur le montant à arbitrer ou le taux d'arbitrage de l'assuré selon le modèle GLM.

## 5.5 Modèle XGBoost taux d'arbitrage

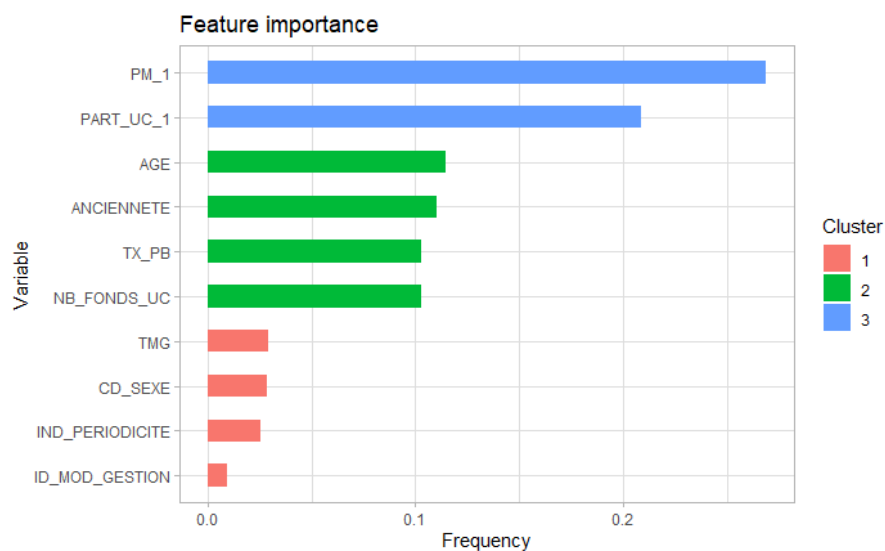
Comme pour le modèle GLM, le modèle XGBoost a été estimé en deux étapes. La première consiste à approcher le taux d'arbitrage par un modèle XGboost qui n'utilise que les variables structurelles et la deuxième étape consiste à modéliser le résidu du premier XGBoost par les variables conjoncturelles.

### 5.5.1 Modèle XGBoost taux d'arbitrage structurel

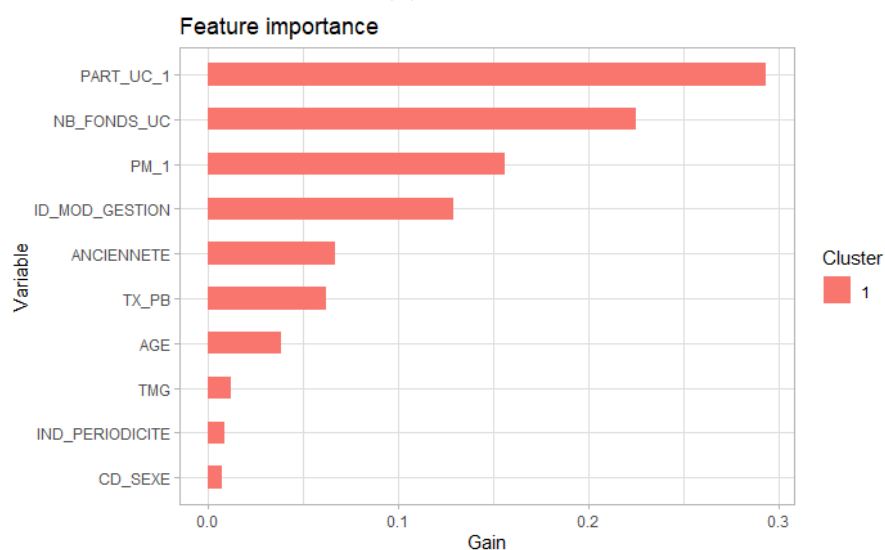
Comme le modèle XGBoost tient compte des effets non-linéaires des variables explicatives continues, donc nous avons gardé les variables comme telle pour le modèle XGBoost. Comme pour le modèle XGBoost fréquence, nous avons cherché les valeurs optimales des paramètres du modèle XGBoost taux d'arbitrage structurel par la méthode de validation croisée en utilisant la fonction RMSE comme critère de sélection. Nous avons donné des valeurs candidates pour chaque paramètre cité dans le modèle XGBoost fréquence, comme la taille de la base taux d'arbitrage est modérée donc il était possible d'estimer la RMSE par validation croisée pour toutes les combinaisons possibles des paramètres candidats. Il en ressort que les paramètres optimaux sont : `n_rounds = 200`, `max_depth=10`, `col_sample_by_tree = 0.5`, `eta=0.1`, `gamma=0`, `min_child_weight = 5`, `subsample = 1`.



(a) Cover



(b) Weigth



(c) Gain

FIGURE 5.5 – Importance des variables dans le modèle XGboost structurel

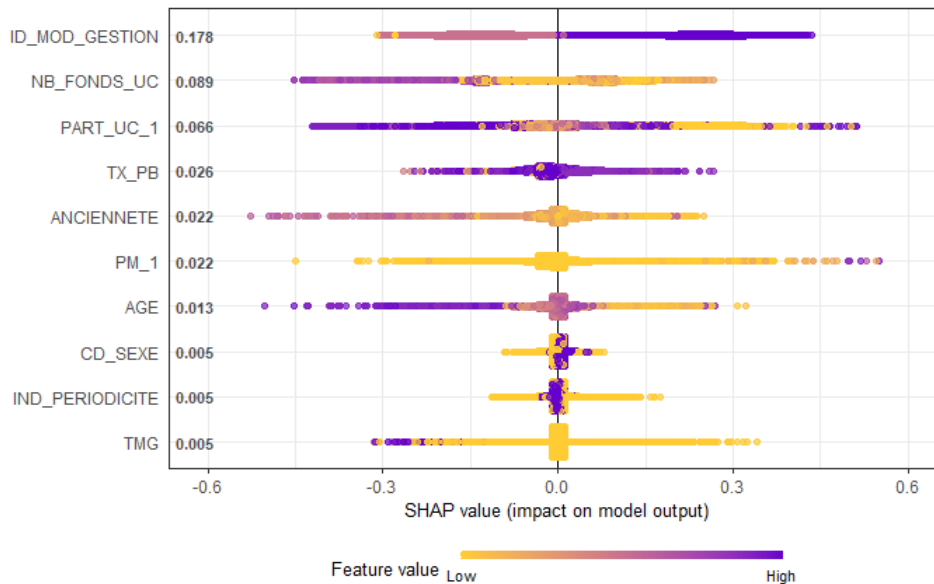
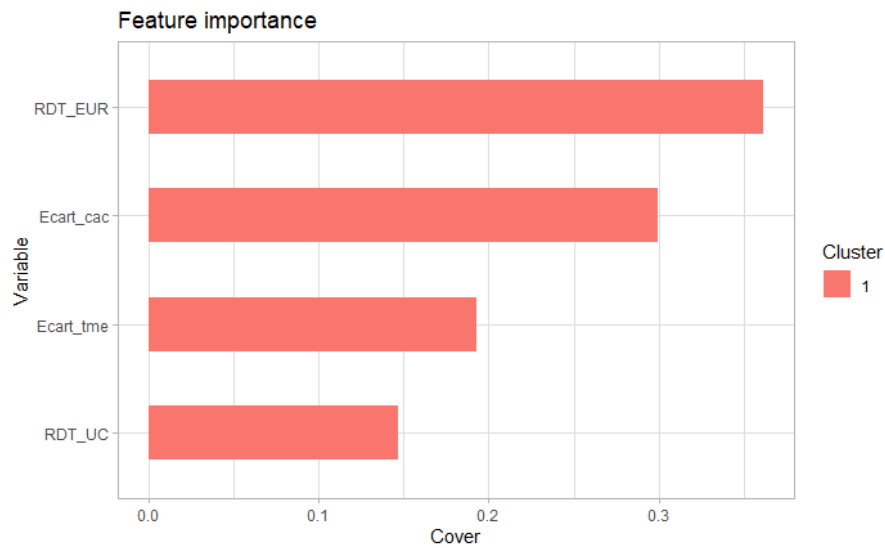


FIGURE 5.6 – Importance des variables dans le modèle XGboost structurel par la méthode SHAP

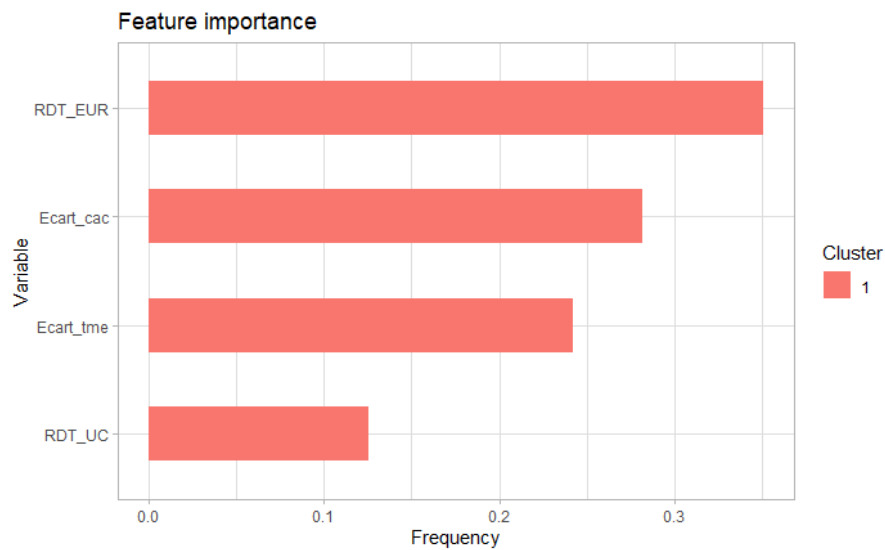
En termes d'importance des variables, le montant de la provision mathématique et la part en UC dans le contrat sont les variables les plus utilisées et qui contribuent le plus à la construction du modèle pour les mesures weight, gain et cover. Le modèle XGBoost arrive à traduire le comportement d'aversion au risque des assurés. Ensuite, viennent les variables âges et le nombre de fonds en UC, donc le modèle XGBoost est capable de bien capter et d'affirmer notre hypothèse de départ qui stipule que l'âge de l'assuré a une incidence sur son comportement d'arbitrage. Du point de vue de la contribution à la prédiction, il s'avère que le mode de gestion est la variable qui contribue le plus en moyenne à la prédiction du taux d'arbitrage structurel. Que ce soit du point de vue de l'utilisation à la construction du modèle ou de l'interprétation SHAP, les variables part en UC et nombre de fonds en UC restent toujours très déterminantes pour la prédiction du taux d'arbitrage structurel.

### 5.5.2 Modèle XGboost taux d'arbitrage conjoncturel

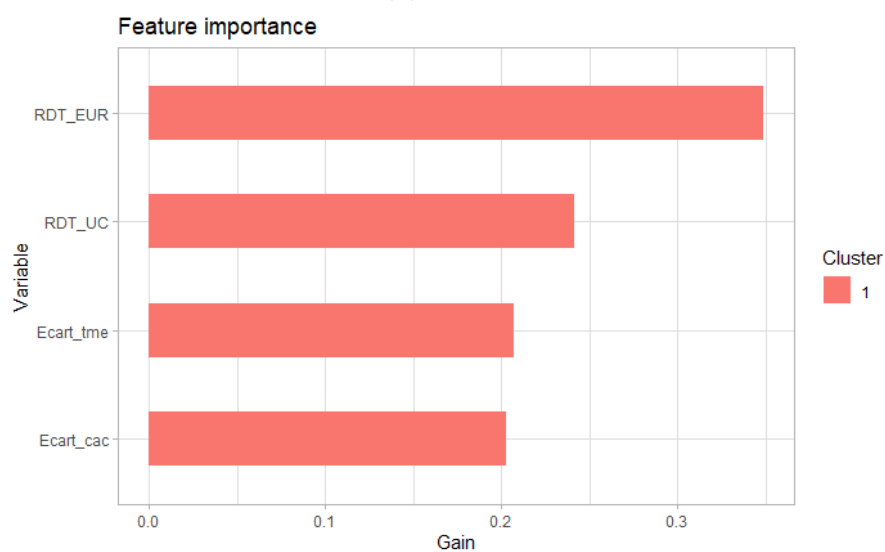
A priori, le modèle XGboost ne permet pas de sélectionner directement les variables et est impacté par la présence des variables corrélées dans le modèle. Donc, nous avons décidé d'enlever la variable `Ecart_eurstx` (corrélée avec `Ecart_cac`) et la variable `Ecart_tx10ans` (corrélée avec `Ecart_tme`). Nous avons aussi testé des paramètres candidats pour chaque variable et les paramètres optimaux sont obtenus par validation croisée à savoir : `n_rounds = 70`, `max_depth=3`, `col_sample_by_tree = 0.5`, `eta=0.1`, `gamma=0`, `min_child_weight = 1`, `subsample = 1`.



(a) Cover



(b) Weigth



(c) Gain

FIGURE 5.7 – Importance des variables dans le modèle XGboost conjoncturel

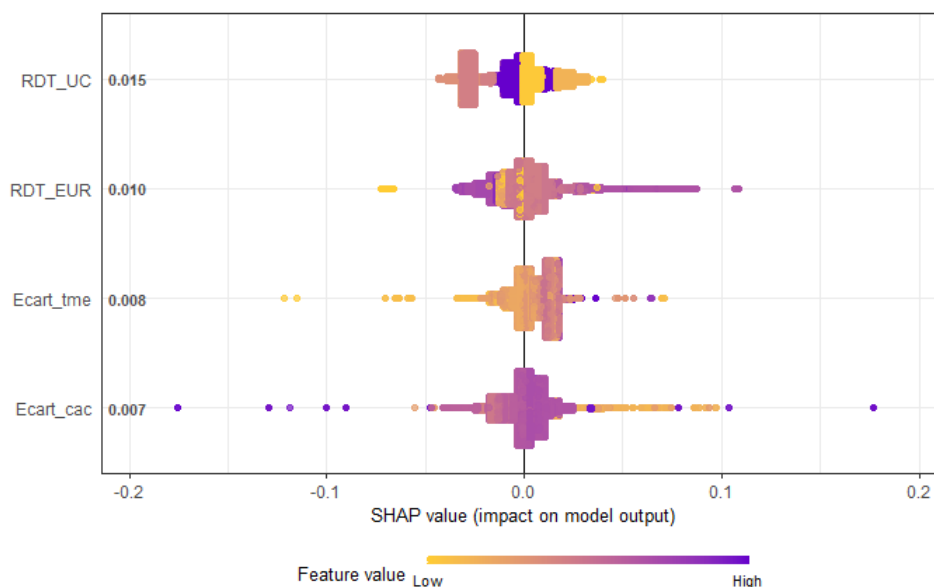


FIGURE 5.8 – Importance des variables dans le modèle XGboost conjoncturel par la méthode SHAP

Selon tous les critères d'importance des variables classiques, la figure 5.7 montre que le rendement en euro du contrat est la variable la plus utilisée pour la construction du modèle ce qui signifie que c'est cette variable qui influe le plus sur le comportement d'arbitrage dynamique des assurés en termes de taux selon le modèle XGBoost. Nous pouvons interpréter le résultat comme si les assurés accordent beaucoup plus d'importance à la sécurité qu'au rendement de leur contrat selon le modèle. Selon le critère d'importance SHAP, qui mesure la contribution à la prédiction absolue moyenne de la variable, le rendement du fonds en UC est la variable la plus importante pour la prédiction du taux d'arbitrage conjoncturel.

TABLEAU 5.5 – Performance des modèles taux d'arbitrages

Modèles	GLM tanh + LR	XGBoost
RMSE_train	0.4570	0.4027
RMSE_test	0.5958	0.5754
MAE	0.4967	0.2655
$R^2$	0.2787	0.5189

L'évaluation des modèles taux d'arbitrage est effectuée à la fois sur l'échantillon d'entraînement de 2013 à 2018 (RMSE, MAE et  $R^2$ ) et sur l'échantillon test de l'année 2019 à 2020 (RMSE et prédiction). Le tableau 5.5 montre que le modèle XGBoost est plus performant que le GLM tanh que ce soit en termes de pouvoir explicatif ou en termes de pouvoir prédictif. Ceci peut s'expliquer par le fait que le modèle XGBoost favorise une meilleure prise en compte des effets non-linéaires des variables continues. Bien que la discrétisation des variables conti-

nues dans le GLM permet de capter des non-linéarités, la plage des valeurs dans l'échantillon d'apprentissage pourrait différer de la plage dans l'échantillon test pour des variables hautement volatiles ; ainsi la discrétisation optimale obtenue sur la base des données historiques ne reflète plus l'information contenue dans la variable pour les années à venir (échantillon test). Par exemple, le rendement annuel du CAC entre 2013 et 2017 évolue entre -0.54 % et 17.99 % alors que le rendement du CAC en 2018 est de -10.96 % et celui de 2019 est 26,37 %. Notons aussi que le pouvoir explicatif  $R^2$  des deux modèles est moyen, ce résultat ne nous étonne vu la difficulté à cerner le comportement d'arbitrage des assurés et l'utilisation des variables qui approche certains facteurs jugés importants dans l'analyse des arbitrages comme le rendement des fonds en unités de compte par exemple.

Comme la valeur des taux d'arbitrage est proche de 0, la régression quantile donne des valeurs instables de quantiles. Nous avons utilisé la méthode percentile bootstrap pour construire l'intervalle de confiance.

TABLEAU 5.6 – Taux d'arbitrage annuel sur les données tests

Taux d'arbitrage	Réel	GLM	XGBoost
2019	-4.21 %	3.39 %	-4.19 %
2020	1.29 %	1.15 %	1.33%

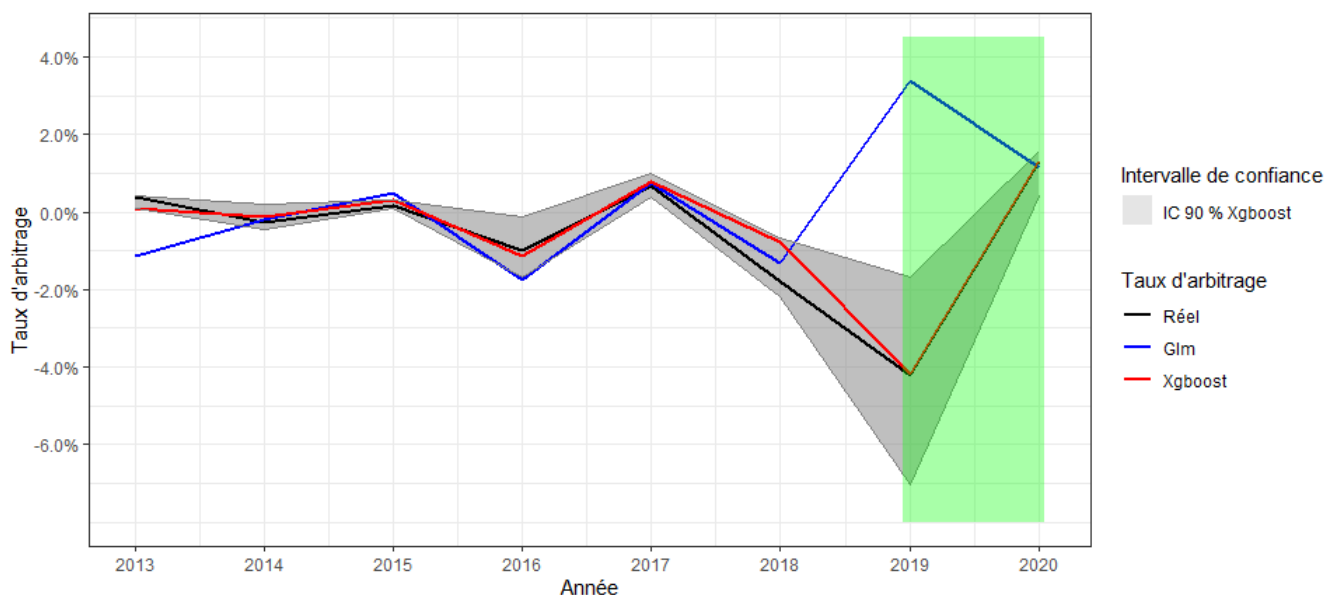


FIGURE 5.9 – Comparaison des taux d'arbitrage annuels

La partie blanche sur la figure 5.9 correspond aux taux d'arbitrage réel et estimés sur les données d'entraînement (2013 à 2018) et la partie verte correspond aux taux d'arbitrage réel et estimés sur les données tests (2019 et 2020). Il en ressort que c'est le modèle XGBoost qui



s'ajuste le mieux sur les données d'entraînement et sur les données tests. Notons aussi que la valeur du montant d'arbitrage net total est faible par rapport à la valeur de la provision mathématique, donc l'écart entre la valeur prédite du montant d'arbitrage et la vraie valeur du montant d'arbitrage devient petit lorsqu'on le ramène en termes de taux.

## ANALYSE DE L'IMPACT D'INTRODUCTION D'UNE LOI D'ARBITRAGE

La projection contrat par contrat du montant d'arbitrage est couteuse en termes de temps et très difficile à mettre en œuvre dans l'outil de gestion d'actif/passif de GENERALI (Prophet). Le présent chapitre est consacré à la modélisation par groupe de contrats des arbitrages et l'analyse de l'impact de cette loi sur la valeur économique de l'engagement de l'assureur.

### 6.1 Le modèle ALM

Le modèle ALM (Asset Liability Management) d'une compagnie d'assurance vie est un outil de projection de ses actifs et de ses passifs sur une période donnée. Il permet d'allouer de manière optimale l'actif de l'assureur en prenant en compte : la politique de gestion de la compagnie, le comportement de l'assuré, les engagements pris à l'égard de l'assuré, la réglementation et l'environnement économique.

Comme toute entreprise, les assureurs vie disposent des actifs et des passifs dans leur bilan. Pour les contrats d'épargne en assurance vie, l'évolution des rendements financiers des actifs de l'assureur a un impact sur leur passif et sur la valorisation des provisions mathématiques. En effet, comme nous l'avons énoncé dans le chapitre 1, les contrats d'épargne sur les fonds en euros sont valorisés chaque année en fonction du rendement des produits financiers de l'assureur. Si le taux de valorisation des contrats ne correspond pas au taux attendu par l'assuré, alors ce dernier peut racheter ou arbitrer son contrat. De même, pour les contrats d'épargne sur les unités de compte, si l'assuré anticipe la survenance d'une crise sur le marché financier, alors il peut arbitrer son contrat vers le fonds en euros.

Il est donc indispensable d'utiliser un modèle ALM pour modéliser les interactions entre l'actif et le passif des assureurs pour se prémunir contre les risques financiers et non financiers auxquels font face les assureurs. En effet, le modèle ALM permet d'avoir une vision prospective de l'interaction entre les éléments du bilan de l'assureur en fonction de l'environnement économique.

Le modèle de projection des actifs et des passifs chez GENERALI s'appelle le modèle ALS (Asset Liability Strategy).

#### 6.1.1 Fonctionnement du modèle

Afin de projeter les flux et les stocks d'actif et de passif de l'assureur, le modèle ALS prend à l'entrée des tables sur la situation de la compagnie d'assurance (table d'actif, table

du passif...), d'hypothèses (table de mortalité...) et des tables sur les scénarios économiques (rendement financier...).

### 6.1.2 Le Cash-Flow du Passif

L'une des tables à l'entrée du modèle ALS est la table des cash-flow déterministes du passif. Les flux déterministes du passif correspondent à l'engagement de l'assureur envers les assurés et qui se caractérisent par les modalités du contrat souscrit. Ils sont obtenus en faisant une projection de ces engagements sans prise en compte de l'environnement économique et sans considération des affaires nouvelles. Comme le nombre de contrats dans le portefeuille de l'assureur est très important, il est donc nécessaire de les regrouper en un groupe de contrats de risque homogène qu'on appelle « Model Point » pour optimiser la projection. En effet, une projection ligne à ligne serait très couteuse en termes de temps et en termes de complexité. L'approche utilisée pour la construction du model point consiste à regrouper les polices qui ont des caractéristiques similaires :

- Le produit, le réseau et la garantie souscrite ;
- L'âge et le sexe de l'assuré ;
- L'ancienneté du contrat ;
- Le taux garanti (contrats en euro, prévoyance), le support utilisé (UC/multisupports) ...

La projection des cash-flow déterministes du passif se déroule en trois étapes. La première étape consiste à projeter les conditions initiales du contrat, c'est-à-dire évaluer le montant de la prime par police et de la provision mathématique encours par police. La deuxième étape consiste à évaluer les prestations à travers l'évaluation du nombre de décès, de rachats, de maturité, du nombre de polices encours en fin de période et du nombre de police encours en début de période. La dernière étape correspond à l'évaluation de l'engagement probabiliste des contrats qui consiste à réévaluer le nouveau montant de la prime par police et de la provision mathématique par police après la réalisation des différentes prestations. La projection des cash-flow déterministes permet de générer deux tables qui sont utilisées comme input du modèle ALS :

- La table DET\_CF (Determinist cash-flow) qui comporte pour chaque année de projection :
  - \* Le montant de la provision mathématique ;
  - \* Le nombre de polices en fin de période ;
  - \* Le montant des rachats structurels ;
  - \* Le montant de rachats partiels ;
  - \* Le montant des arrérages ;
  - \* Le montant des maturités ;

- \* Les frais de gestion ;
- \* Les frais de gestion ;
- \* Les commissions sur encours, prime. . .

— La table de paramètre pour chaque produit (taux cible ou pas, TMG. . .)

Toutes les variables présentes dans les deux tables sont obtenues à la maille produit (regroupement de model point).

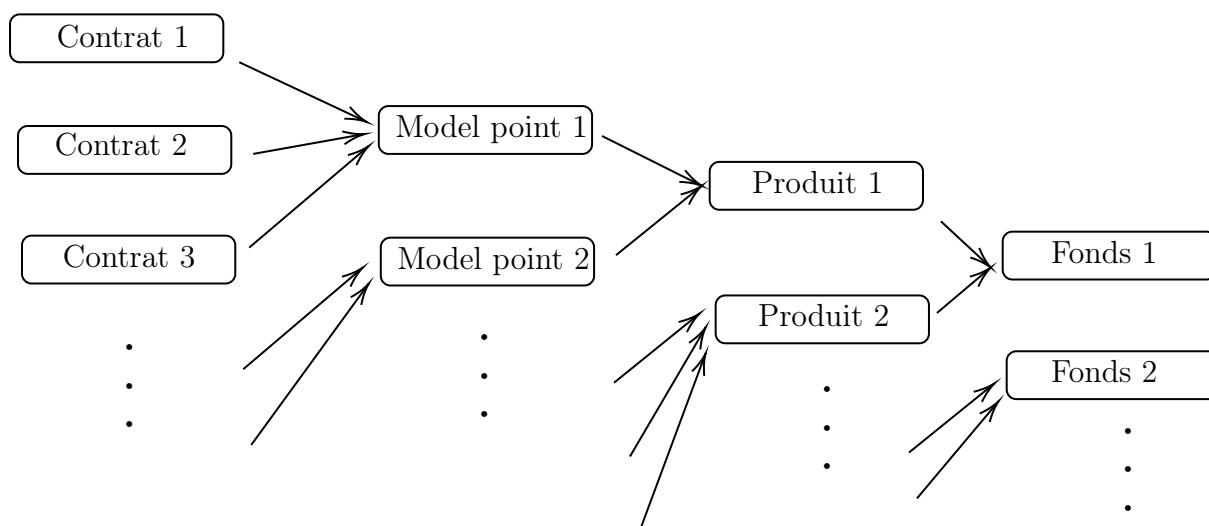


FIGURE 6.1 – Regroupement des contrats

### 6.1.3 Le Cash-flow de l'actif

La projection des cash-flows d'actif se fait aussi par model point. Mais cette partie ne sera pas détaillée dans ce mémoire.

### 6.1.4 Le scénario d'équivalent certain

Le calcul des flux dans le modèle ALS dépend de la conjoncture économique. Pour valider la convergence du modèle, il est souvent nécessaire de définir un scénario économique benchmark. Le scénario d'équivalent certain est le scénario de référence dans le modèle ALM, c'est un scénario dans lequel tous les actifs ont le même rendement que l'actif sans risques donc il n'y a pas un déclenchement des comportements dynamiques de l'assuré.

### 6.1.5 Les scénarios stochastiques

Les scénarios stochastiques utilisés dans le modèle ALS sont des scénarios économiques (taux d'inflation, taux de rendement des actifs. . .) obtenus à partir d'un générateur de scénario économique (GSE).

## 6.2 Le Best Estimate

Avant de définir le Best Estimate, nous allons rappeler le principe de la directive solvabilité 2. En effet, le Best Estimate est la valeur économique du passif dans le bilan prudentiel de solvabilité 2. La réforme Solvabilité 2 a pour objectif d'harmoniser et de renforcer la réglementation du secteur des assurances afin de garantir la stabilité de l'activité et de fournir une protection et une information appropriée pour les investisseurs et les assurés. Pour améliorer l'évaluation et le contrôle des risques, la réglementation solvabilité 2 a mis en place trois piliers : les exigences quantitatives, les exigences qualitatives et les exigences informationnelles. Afin d'atteindre cet objectif, Solvabilité 2 a instauré la notion de bilan prudentiel. Le bilan prudentiel permet d'avoir la valeur économique des différents éléments du bilan en utilisant les valeurs du marché ou des valeurs cohérentes avec celles du marché. Il permet d'évaluer la valeur des actifs et des passifs de l'assureur à leur juste valeur.

Le Best Estimate (BE) se définit comme : « la moyenne pondérée par leur probabilité des flux de trésorerie futurs compte tenu de la valeur temporelle de l'argent estimée sur la base de la courbe des taux sans risque pertinente, soit la valeur actuelle attendue des flux de trésorerie futurs » (Article R351-2 du Code des Assurances, transposition en droit français de l'article 77 de la Directive Solvabilité 2).

Le BE ne prend en compte que des engagements faisant partie des frontières du contrat. En effet, le calcul se fait en « run-off » c'est-à-dire que les contrats futurs sont exclus de la valorisation. Le BE tient compte aussi de tous les flux futurs entrants et sortants jusqu'à l'extinction des contrats en portefeuille.

Le Best Estimate se calcule comme suit :

$$BE = \mathbb{E}^{\mathbb{Q}} \left( \sum_{t=1}^{\infty} \frac{CF_t^{out} - CF_t^{in}}{(1 + r_t)^t} \right)$$

Avec :

- $\mathbb{Q}$  : la probabilité risque neutre ;
- $CF_t^{out}$  : les cash-flows sortants à la date  $t$  ;
- $CF_t^{in}$  : les cash-flows entrants à la date  $t$  ;
- $r_t$  : le taux sans risque à la date  $t$ .

Etant donné que le modèle ALS effectue des projections sur un horizon de temps finis, il s'avère qu'à la fin de la projection il reste encore des contrats d'assurance dans le portefeuille de l'assureur. Ainsi, pour un horizon de projection  $T$  fini le Best Estimate se calcule comme suit :

$$BE = \mathbb{E}^{\mathbb{Q}} \left( \sum_{t=1}^T \frac{CF_t^{out} - CF_t^{in}}{(1 + r_t)^t} + \frac{VF_T}{(1 + r_T)^T} \right)$$

Avec  $VF_T$  : liquidation finale de l'actif qui correspond à la valeur des contrats restants dans le portefeuille de l'assureur à l'instant T.

Dans le modèle ALS le calcul du Best Estimate se fait par Monte Carlo car il est impossible d'obtenir la valeur exacte du BE par cette formule. Le GES fournit N scénarios économiques indépendants qui permettent de calculer la valeur du Best Estimate.

Avec N scénarios économiques et un horizon de projection T la contrepartie empirique du Best Estimate dans le modèle ALS s'écrit comme suit :

$$BE = \frac{1}{N} \sum_{n=1}^N \left( \sum_{t=1}^T \frac{CF_t^{out} - CF_t^{in}}{(1+r_t)^t} + \frac{VF_T}{(1+r_T)^T} \right)$$

Les flux intervenants dans le calcul du Best Estimate sont :

- les frais ;
- les prestations ;
- les commissions ;
- les primes.

### 6.3 La PVFP

Le BE est assimilé à la part de richesse distribuée aux assurés, de même il est possible de calculer la valeur actuelle des profits ou pertes des actionnaires qu'on appelle PVFP (Present Value of Future Profit).

La PVFP se calcule comme suit :

$$PVFP = \mathbb{E}^{\mathbb{Q}} \left( \sum_{t=1}^T \frac{R_t}{(1+r_t)^t} \right)$$

$R_t$  : Les flux de trésorerie relatifs aux actionnaires (résultats versés, augmentation de capital ...)

Avec N scénarios économiques et un horizon de projection T la contrepartie empirique de la PVFP dans le modèle ALS s'écrit comme suit :

$$PVFP = \frac{1}{N} \sum_{n=1}^N \left( \sum_{t=1}^T \frac{R_t}{(1+r_t)^t} \right)$$

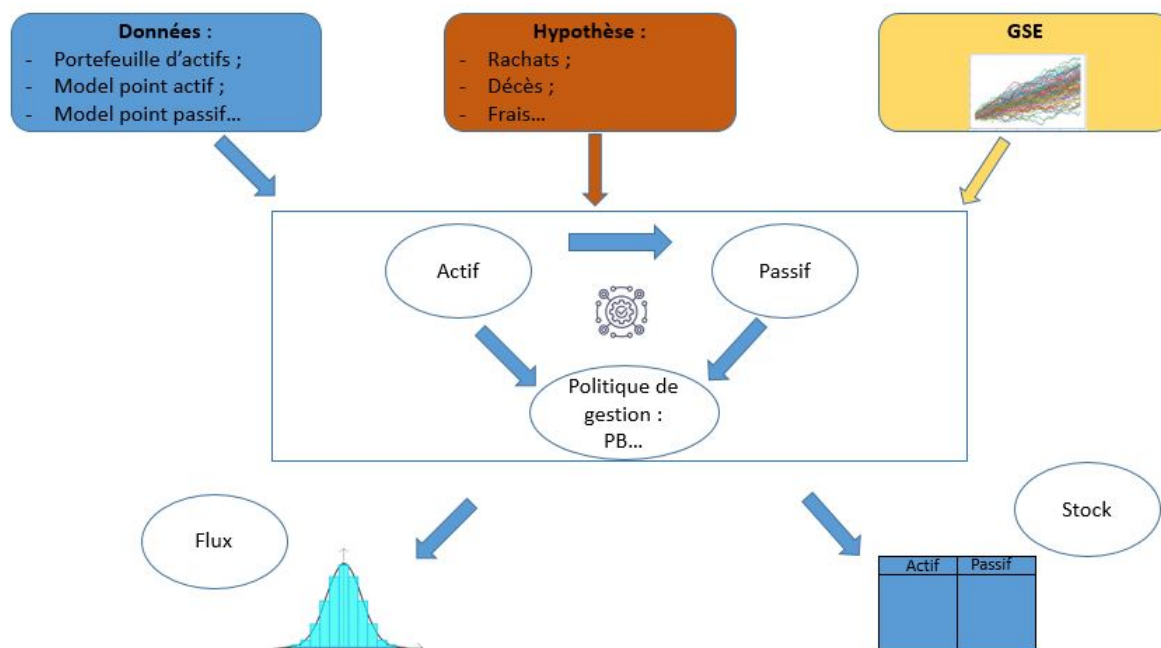


FIGURE 6.2 – Architecture d'un modèle de projection Actif/Passif

## 6.4 Modélisation du taux d'arbitrage par groupe de contrat

Il existe neuf (9) grandes familles de fonds en euros dans le portefeuille de GENERALI.

TABLEAU 6.1 – Poids des fonds

Fonds	Poids
1	34 %
2	36,93 %
3	1,97%
4	5,46 %
5	1,71%
6	11,15%
7	2,61%
8	0.01%
9	6,16 %

Dans cette partie, le taux d'arbitrage se calcule par fonds en euros et se définit comme :

$$\text{Taux arbitrage} = \frac{\text{Montant arbitré}}{\text{PM du fonds en euros}}$$

Si le taux est positif alors il y a un arbitrage du fonds en euros vers le fonds en UC et si le taux est négatif alors il y a un arbitrage du fonds en UC vers le fonds en euros.

### 6.4.1 Arbitrages structurels

Rappelons que les arbitrages structurels correspondent aux comportements d'arbitrage des assurés qui ne dépendent pas de la conjoncture économique mais des facteurs structurels comme l'âge, le sexe, l'ancienneté. . . Par la suite, nous allons supposer que les assurés qui se retrouvent dans un même produit ont le même comportement d'arbitrage. La modélisation des arbitrages structurels va donc se faire par produit. Pour des raisons de simplicité, le taux d'arbitrage structurel sera modélisé comme une moyenne mobile pondérée des taux d'arbitrage passé. En effet, afin de mieux représenter le comportement d'arbitrage des dernières années, il est plus raisonnable d'accorder un poids plus important aux arbitrages récents et un poids moins important aux arbitrages lointains. Nous allons définir une moyenne mobile spécifique afin de prendre en compte cet aspect dans notre modèle. Le poids du taux d'arbitrage en  $n-p$  dans le calcul du taux d'arbitrage en  $n$  est égal à  $\frac{1}{\alpha^{p-1}}$ . L'idée derrière est de dire que le poids du taux d'arbitrage en  $n-1$  est égal à 1 et ce poids va décroître de  $\frac{1}{\alpha}$  à chaque fois qu'on recule d'une année.

$$x_t = \sum_{i=1}^p \frac{\frac{1}{\alpha^{i-1}} x_{t-i}}{\sum_{j=1}^p \frac{1}{\alpha^{j-1}}}$$

La valeur de  $\alpha$  et l'ordre de la moyenne mobile sont déterminés ensuite sur la base de la RMSE. Nous disposons d'un historique du taux d'arbitrage annuel par produit entre 2007 et 2020. Nous calculons par moyenne mobile le taux d'arbitrage annuel par produit entre 2012 et 2020 et nous comparons la RMSE pour déterminer l'ordre et la valeur des paramètres optimaux.

TABLEAU 6.2 – RMSE par ordre et  $\alpha$ 

$\alpha$	Ordre			
	2	3	4	5
2	0.0073	0.0069	0.0066	0.0074
3	0.0074	0.0071	0.0069	0.0069
4	0.0075	0.0073	0.0072	0.0069
5	0.0076	0.0074	0.0074	0.0074

Il en ressort que  $\alpha = 2$  et l'ordre de la moyenne mobile égal à 4 qui minimise la valeur de la RMSE.

### 6.4.2 Loi QIS5

La première loi d'arbitrage dynamique que nous allons considérer est la loi QIS 5, c'est une loi préconisée par l'ONC pour modéliser les rachats dynamiques mais souvent utiliser aussi pour la modélisation des arbitrages dynamiques.

Cette loi modélise les arbitrages dynamiques qui se déclenchent en fonction de la différence entre le taux servi sur le fonds en euros (TE) et le taux servi sur le fonds en UC (TU) et adopte



une approche de modélisation linéaire. Dans sa formulation générale,

- la loi admet un taux maximum d'arbitrage qui est atteint lorsque l'écart de taux :  $TE - TU$  est en dessous du seuil  $\alpha$ . Ce n'est plus l'écart de taux qui explique le comportement des assurés ;
- les arbitrages dynamiques sont nuls lorsque l'écart se situe entre  $\beta$  et  $\gamma$  ;
- elle autorise également des arbitrages dynamiques négatifs lorsque cet écart excède un seuil  $\gamma$  ;
- enfin, la loi définit le minimum d'arbitrage dynamique qu'on peut atteindre lorsque le l'écart de taux excède un seuil  $\delta$ . Ce n'est plus l'écart de taux qui explique le comportement des assurés ;

$$ARB(TE, TU) = \begin{cases} ARB_{\max} & \text{si } TE - TU < \alpha \\ ARB_{\max} \frac{TE - TU - \beta}{\alpha - \beta} & \text{si } \alpha < TE - TU < \beta \\ 0 & \text{si } \beta < TE - TU < \gamma \\ ARB_{\min} \frac{TE - TU - \beta}{\alpha - \beta} & \text{si } \gamma < TE - TU < \delta \\ ARB_{\min} & \text{si } \delta < TE - TU \end{cases}$$

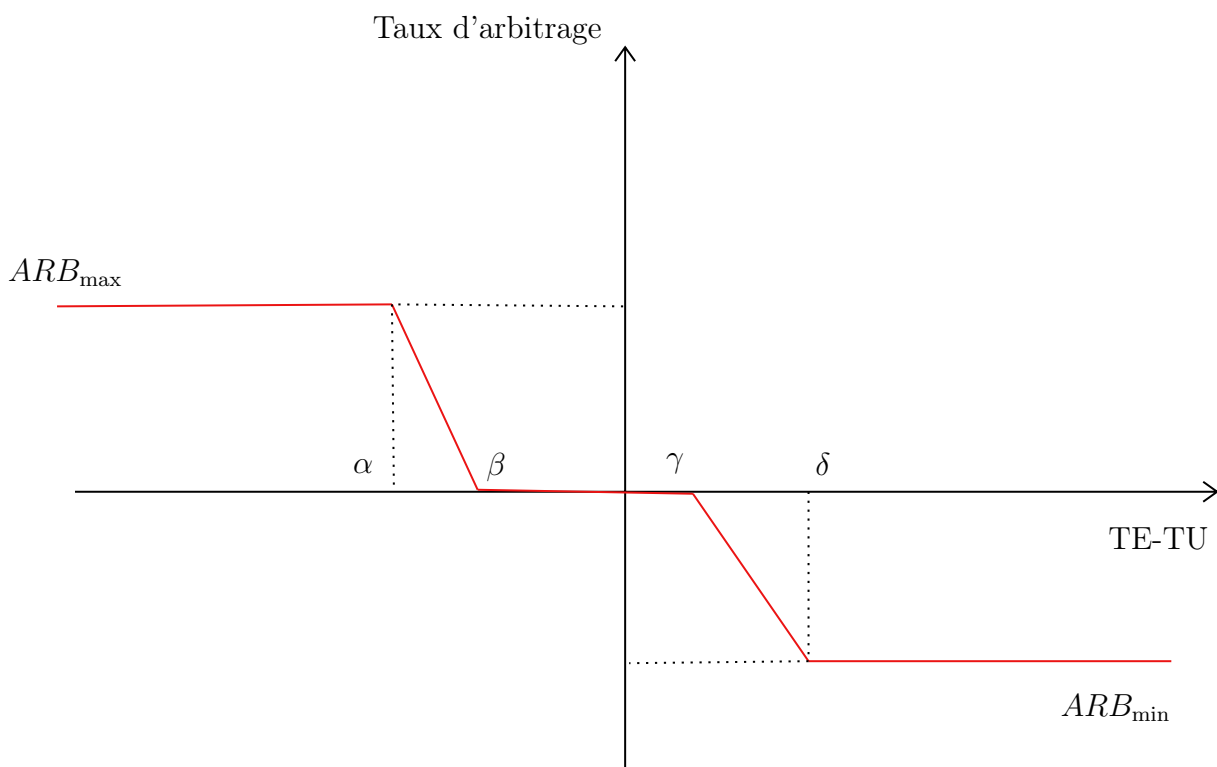


FIGURE 6.3 – Loi QIS

Les seuils de déclenchement et les taux extrémaux sont déterminés à partir de l'historique ou sur la base d'un modèle statistique que nous allons voir par la suite. Les seuils de stabilisations

sont obtenus à partir de l'historique (écart maximum et écart minimum entre les rendements sur l'historique).

### 6.4.3 Détermination des taux extrêmes

L'objectif de cette partie est de construire un modèle statistique pour déterminer les taux d'arbitrages maximums et minimums avec un niveau de risque donné. Comme la période d'observation des arbitrages n'est pas assez longue pour effectuer une inférence sur les données, dans cette partie, nous avons décidé de faire une analyse à pas de temps mensuel. Un coefficient de passage sera ensuite utilisé pour passer des taux extrêmes mensuels aux taux extrêmes annuels.

Comme pour la loi QIS 5, nous avons décidé de définir le taux d'arbitrage par rapport à un fond donné comme le rapport entre le montant entrant ou sortant sur ce fonds et sa provision mathématique.

Pour tout fonds en euros confondu, la figure 6.4 nous laisse penser l'existence d'un comportement de retour à la moyenne de la série de taux d'arbitrage mensuel du fonds en euros. Petar Radkov (2010) (17) a démontré dans son travail que la présence d'un comportement autorégressif (AR) dans une série et sa stationnarité équivaut à dire que la série présente une tendance de retour à la moyenne.

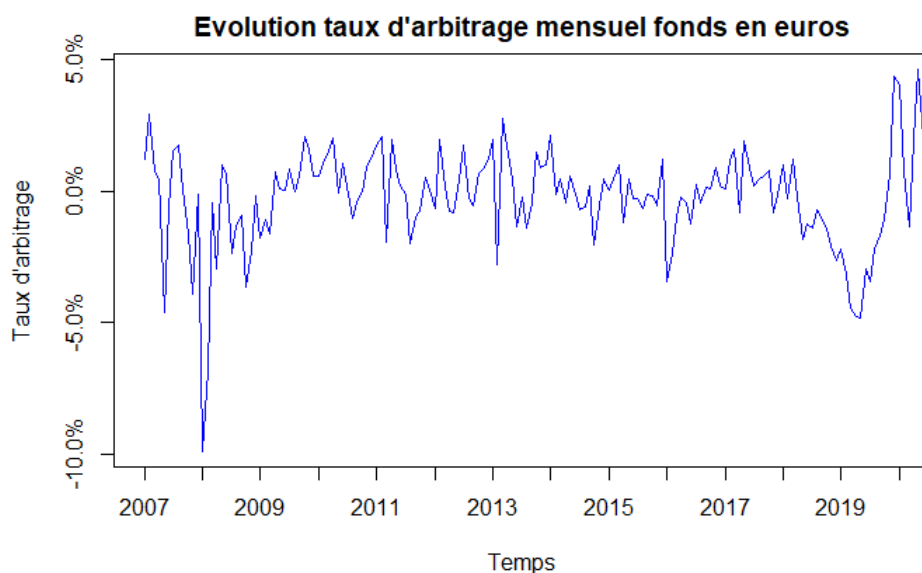


FIGURE 6.4 – Evolution du taux d'arbitrage mensuel du fonds en euros

D'après le graphique d'autocorrélation partielle du fonds en euros dans son ensemble, il existe une corrélation significative au niveau de décalage 1, suivie par des corrélations non significatives. Ce graphique indique donc que la série taux d'arbitrage mensuel possède est un terme autorégressif d'ordre 1 c'est-à-dire un AR(1).

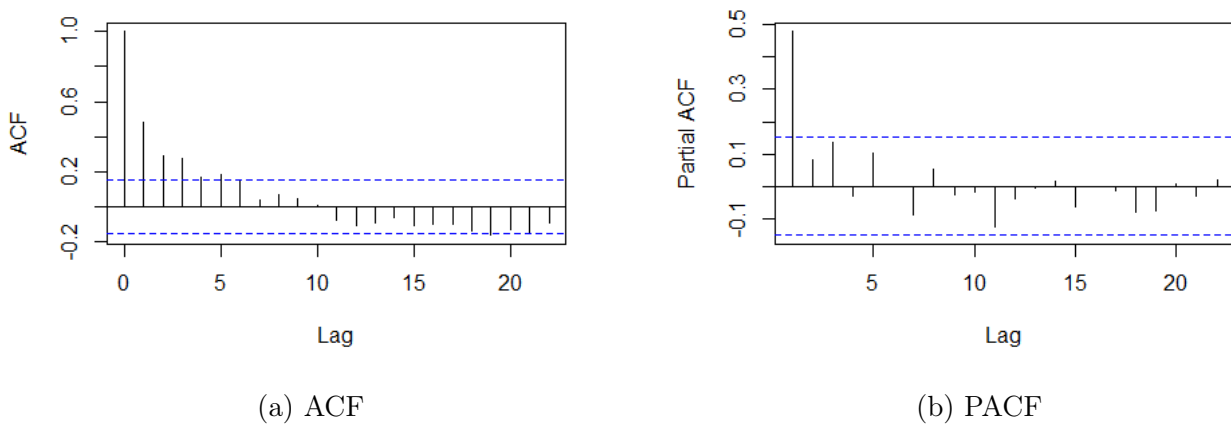


FIGURE 6.5 – ACF et PACF du fonds en euros

La propriété de retour à la moyenne est vérifiée pour un processus AR(1) si et seulement si son coefficient d'autorégression est inférieur ou égal à 1. Pour rappel, un processus  $x_t$  est AR(1) s'il peut s'écrire comme :

$$x_{t+1} = \mu + \alpha x_t + \sigma \varepsilon_{t+1} \Rightarrow \Delta x_{t+1} = (1 - \alpha) \left( \frac{\mu}{1 - \alpha} - x_t \right) + \sigma \varepsilon_{t+1}$$

Pour un processus AR(1),  $|\alpha| \leq 1$  est une condition nécessaire et suffisante pour la stationnarité. Notons que pour  $\alpha = 1$ , nous sommes en présence d'une marche aléatoire avec un drift constant.

Le test de Dickey-Fuller augmenté (ADF, l'hypothèse nulle du test est la présence d'une racine unitaire) David Dickey et Wayne Fuller (1981) (5) et le test de KPSS (l'hypothèse nulle du test est la nullité de la variance de la composante non stationnaire de la série) Denis Kwiatkowski et al (1992) (6) seront utilisés pour tester la stationnarité du taux d'arbitrage mensuel des fonds en euros dans son ensemble. Le test d'ADF donne un p-value = 0.03 et le test de KPSS donne un p-value = 0.1, donc nous rejetons l'hypothèse de non stationnarité de la série du taux d'arbitrage mensuel du fonds en euros.

Certains modèles stochastiques comme le modèle de Vasicek permettent de prendre en compte le comportement de retour à la moyenne d'une série et dont la discrétisation permet d'obtenir un processus AR(1). Donc, le modèle de Vasicek semble être adapté pour modéliser le taux d'arbitrage mensuel des fonds en euros. Ce modèle présente aussi l'avantage d'être facile à comprendre sur le plan théorique et présente des expressions analytiques moins complexes. Notons qu'il existe d'autres modèles stochastiques qui prend en compte l'effet de retour à la moyenne et dont la forme discrète correspond à un AR(1) comme le modèle CIR mais qui présente une forme analytique plus complexe.

Le modèle de Vasicek (proposé en 1977) est l'un des premiers modèles stochastiques de taux. On se place sur un espace de probabilité  $(\Omega, \mathcal{F}, \mathcal{P})$  muni d'une filtration  $(\mathcal{F}_t)_{0 \leq t \leq T}$ . On suppose que sous une probabilité risque-neutre  $\mathbb{Q}$ , le taux court instantané suit un processus

d'Ornstein-Uhlenbeck avec des coefficients constants. Le taux spot a la dynamique :

$$dr_t = \alpha(\theta - r_t)dt + \sigma dW_t$$

$\theta$  correspond à la moyenne du processus  $(r_t)_{t \geq 0}$ ,  $\alpha$  est la force de retour à la moyenne et  $\sigma$  est la volatilité, qui va donner plus ou moins d'importance au bruit  $(W_t)_{t \geq 0}$ .

La solution de cette équation différentielle stochastique s'écrit alors, pour  $0 < s < t$  :

$$r_t = r_0 e^{-\alpha t} + \theta (1 - e^{-\alpha t}) + \sigma e^{-\alpha t} \int_0^t e^{\alpha s} dW_s$$

avec

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}} [r_t | \mathcal{F}_s] &= r_s e^{-\alpha(t-s)} + \theta (1 - e^{-\alpha(t-s)}) \\ \text{Var}^{\mathbb{Q}} [r_t | \mathcal{F}_s] &= \frac{\sigma^2}{2\alpha} (1 - e^{-2\alpha(t-s)}) \end{aligned}$$

et  $r_t$  est un processus gaussien.

En régime stationnaire, c'est-à-dire lorsque  $t \rightarrow \infty$  on a :

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}} [r_t | \mathcal{F}_s] &= \theta \\ \text{Var}^{\mathbb{Q}} [r_t | \mathcal{F}_s] &= \frac{\sigma^2}{2\alpha} \end{aligned}$$

En discrétisant l'équation sur  $t_0, t_1, t_2 \dots$  avec un pas de temps  $\Delta t = t_i - t_{i-1} = 1$  (1 mois dans notre cas), on a :

$$r(t_i) = c + br(t_{i-1}) + \delta \varepsilon(t_i)$$

avec :

$$c = \theta (1 - e^{-\alpha \Delta t}) = \theta \tag{6.1}$$

$$b = e^{-\alpha \Delta t} = e^{-\alpha} \tag{6.2}$$

$\varepsilon(t)$  une variable normale  $N(0, 1)$ . La volatilité du terme d'innovation s'obtient par la formule d'Itô :

$$\delta = \sigma \sqrt{\left(\frac{1 - e^{-2\alpha \Delta t}}{2\alpha}\right)} = \sigma \sqrt{\left(\frac{1 - e^{-2\alpha}}{2\alpha}\right)} \tag{6.3}$$

Pour calibrer le modèle de Vasicek, la forme discrète du processus sera utilisée. La valeur des paramètres  $c$ ,  $b$  et  $\delta$  est déterminée en effectuant une régression linéaire de la variable  $r_t$  sur la variable  $r_{t-1}$ . La valeur des paramètres  $\theta$ ,  $\alpha$  et  $\sigma$  sera obtenue par la suite en résolvant les équations 6.1, 6.2 et 6.3.

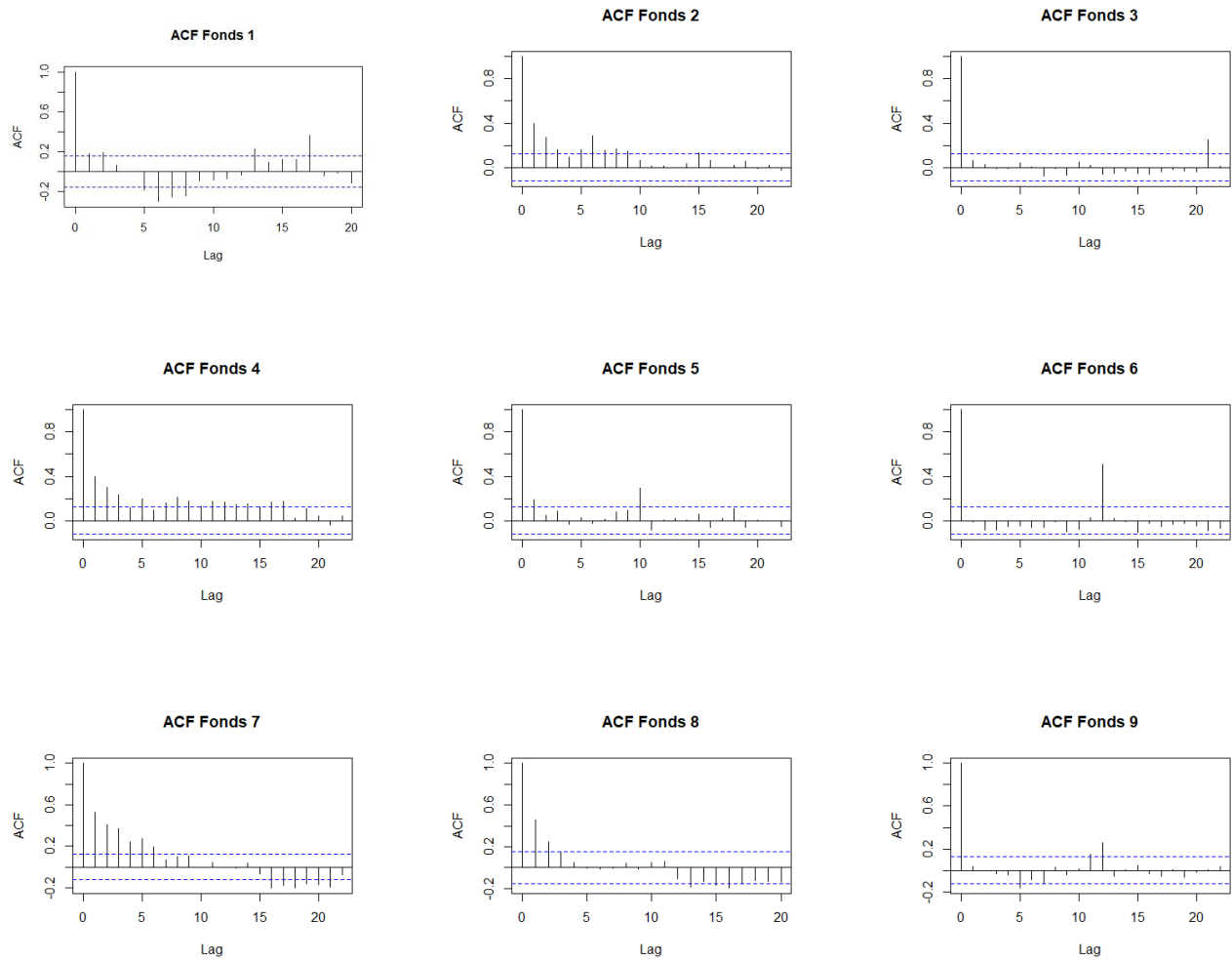
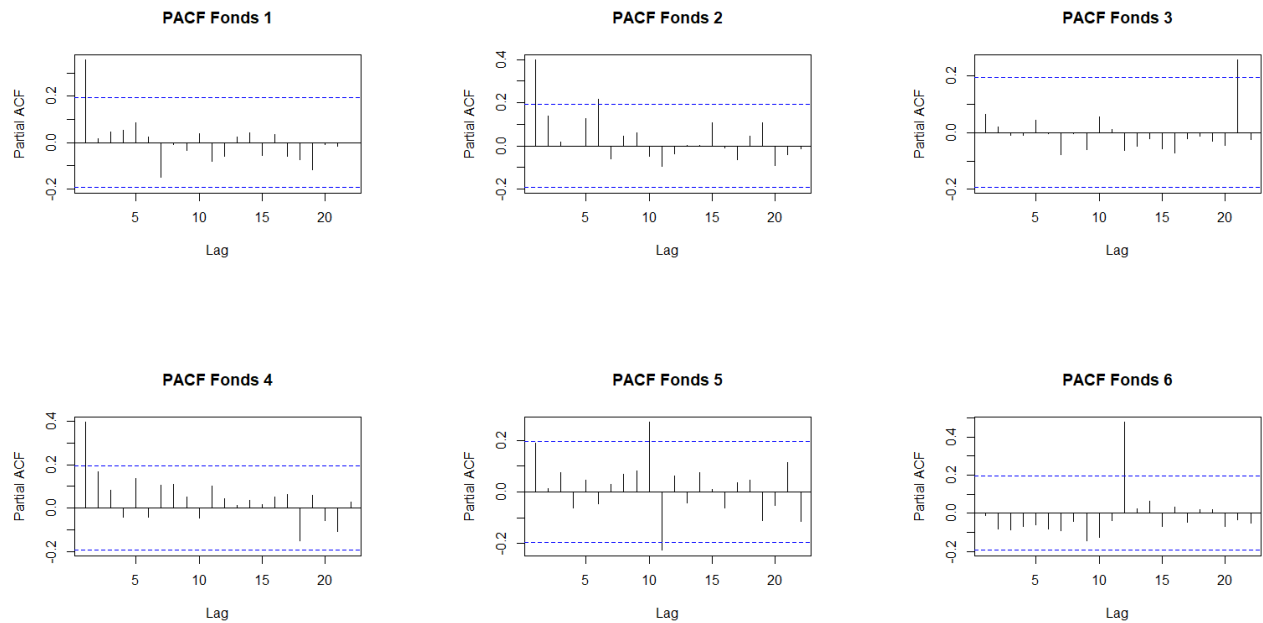


FIGURE 6.8 – ACF des fonds en euros



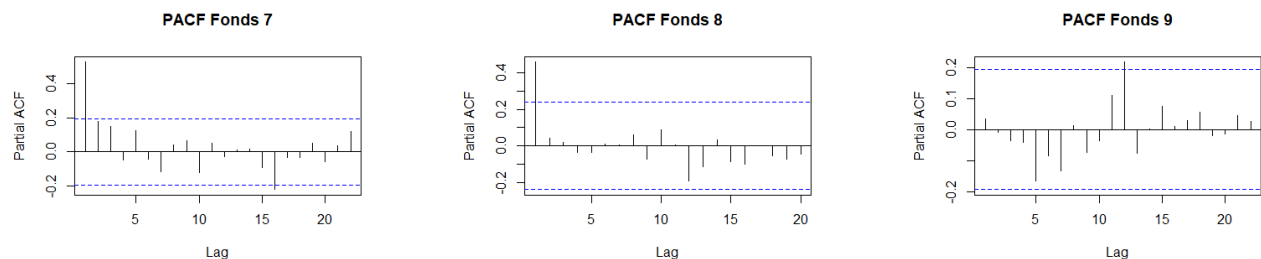


FIGURE 6.11 – PACF des fonds en euros

Une analyse par fonds des ACF et des PACF permet de voir que la plupart des fonds en euros présente un comportement AR(1). Le tableau suivant présente le résultat des tests d'ADF et KPSS par fonds en euros :

TABLEAU 6.3 – Test d'ADF et KPSS par fonds en euros

Fonds	ADF	KPSS
1	0.01	0.10
2	0.01	0.07
3	0.01	0.10
4	0.16	0.10
5	0.01	0.10
6	0.01	0.01
7	0.01	0.10
8	0.01	0.10
9	0.01	0.10

Il en ressort que d'après le test d'ADF, seul le taux d'arbitrage mensuel du fonds 4 n'est pas stationnaire et d'après le test de KPSS, seul le taux d'arbitrage mensuel du fonds 6 n'est pas stationnaire. Sur la base de ces résultats, on peut donc dire que le processus de chaque fonds en euros peut être modélisé comme un processus AR(1). Nous allons modéliser les séries de taux d'arbitrage mensuel par un modèle de Vasicek discret. Le tableau suivant présente la valeur des paramètres après estimation :

TABLEAU 6.4 – Paramètres par fonds en euros

Fonds	$\hat{c}$	$\hat{b}$
1	6.34e-05	0.36
2	-6.69e-04	0.52
3	-1.62e-04	0.46
4	1.18e-03	0.40
5	-1.89e-03	0.06
6	8.35e-05	0.39
7	2.71e-04	0.19
8	-4.61e-03	0.01
9	-3.88e-04	0.04

Les taux extremums sont obtenus en déterminant les quantiles d'ordre  $\beta$  pour un niveau de risque  $\beta$  donné. Selon la directive solvabilité 2, un évènement extrême devrait se produire une fois tous les 200 ans (2400 mois) donc nous avons choisis  $\beta = \frac{1}{2400}$ .

Pour le taux maximum on a :

$$\begin{aligned}
 P(r_t > M) &= \beta \\
 P\left(N\left(\theta, \frac{\sigma^2}{2\alpha}\right) > M\right) &= \beta \\
 P\left(N(0, 1) > \frac{M - \theta}{\frac{\sigma}{\sqrt{2\alpha}}}\right) &= \beta \\
 P\left(N(0, 1) \leq \frac{M - \theta}{\frac{\sigma}{\sqrt{2\alpha}}}\right) &= 1 - \beta \\
 M &= \theta + \frac{\sigma}{\sqrt{2\alpha}}q_{1-\beta}
 \end{aligned}$$

Pour le taux minimum on a :

$$\begin{aligned}
 P(r_t < m) &= \beta \\
 P\left(N\left(\theta, \frac{\sigma^2}{2\alpha}\right) < m\right) &= \beta \\
 P\left(N(0, 1) < \frac{m - \theta}{\frac{\sigma}{\sqrt{2\alpha}}}\right) &= \beta \\
 P\left(N(0, 1) \leq \frac{m - \theta}{\frac{\sigma}{\sqrt{2\alpha}}}\right) &= \beta \\
 m &= \theta + \frac{\sigma}{\sqrt{2\alpha}}q_{\beta}
 \end{aligned}$$

Pour passer des taux extremums mensuels aux taux extremums annuels, nous allons définir

un coefficient de passage :

$$1 + \text{Tx annuel} = (1 + \text{Tx mensuel})^\gamma$$

$$\gamma = \frac{\ln(1 + \text{Tx annuel})}{\ln(1 + \text{Tx mensuel})}$$

Ce taux est calculé pour chaque taux d'arbitrage mensuel pour tous les fonds sur l'historique. Le coefficient de passage maximal pour les taux d'arbitrage mensuel positif sera retenu comme coefficient de passage du taux d'arbitrage maximum mensuel au taux d'arbitrage maximum annuel pour chaque fonds. De même, le coefficient de passage pour les taux d'arbitrage négatif sera retenu comme coefficient de passage pour le taux d'arbitrage minimum par fonds.

Les tableaux suivants présentent les taux d'arbitrage extrémaux sur l'historique et obtenus par le modèle :

TABLEAU 6.5 – Taux extremums obtenus par le modèle de Vasicek

Fonds	Taux max	Taux min
1	4.08 %	-5.77 %
2	8.58 %	-9.48 %
3	10.59 %	- 7.01 %
4	5.74 %	-3.27 %
5	2.14 %	-15.84 %
6	3.78 %	-2.64 %
7	4.08 %	-2.86 %
8	4.37 %	-17.26%
9	2.6 %	-5.99 %

TABLEAU 6.6 – Taux extremums dans l'historique

Fonds	Taux max	Taux min
1	1.49 %	-1.19 %
2	2.44 %	-5.01 %
3	3.56 %	- 2.76 %
4	2.33 %	-0.35 %
5	0.47 %	-8.15 %
6	0.62 %	-1.26 %
7	1.84 %	-0.87 %
8	0.12 %	-9.34%
9	0.19 %	-1.49 %



### 6.4.4 Modèle GLM

Le premier modèle a l'inconvénient de ne pas être calibré sur l'historique. Dans cette partie, un modèle GLM (régression tanh) sera construit pour modéliser le taux d'arbitrage dynamique de chaque fonds en euros à partir des rendements de fonds en euros et les rendements des fonds en UC. Comme les rendements des fonds en euros sont à pas de temps annuel (2013-2020), alors le pas du temps du modèle est annuel. Le taux d'arbitrage dynamique historique s'obtient en calculant la différence entre le taux d'arbitrage historique et le taux d'arbitrage structurel obtenu par moyenne mobile. Pour prendre en compte l'effet non linéaire des taux de rendement sur le taux d'arbitrage dynamique, une discrétisation des variables a été réalisée. Ensuite, la méthode fused lasso a été appliquée pour trouver la discrétisation optimale.

Pour le choix des variables explicatives, nous allons considérer deux modèles : la première utilise l'écart des taux comme variable explicative et le deuxième utilise le taux de rendement des fonds en euros et le taux de rendement des fonds en UC séparément.

TABLEAU 6.7 – RMSE par méthode

Modèle	Écart des taux	Taux séparé
RMSE	0.1194	0.1204

Sur la base de la RMSE, il en ressort que c'est le modèle GLM avec l'écart des taux qui arrive à bien prédire le taux d'arbitrage par fonds en euros.

TABLEAU 6.8 – Résultat modèle GLM avec écart des taux

	Estimate	Std. Error	t value	Pr(>  t )
Ecart_rdt(.,-0.2]	0.0495	0.0309	1.60	0.1121
Ecart_rdt(-0.2,-0.1]	0.0319	0.0252	1.27	0.2084
Ecart_rdt(-0.1,-0.05]	0.0612	0.0343	1.78	0.0779
Ecart_rdt(-0.05,0.03]	-0.0093	0.0200	-0.46	0.6437
Ecart_rdt(0.03,0.12]	-0.0170	0.0617	-0.28	0.7833
Ecart_rdt(0.12,0.15]	-0.0659	0.0344	-1.92	0.0577
Ecart_rdt(0.15,.]	-0.0367	0.0873	-0.42	0.6755

Le résultat du modèle GLM avec écart des taux permet de constater que lorsque l'écart des taux de rendement évolue entre -5 % et 3 %, le taux d'arbitrage est de l'ordre de  $10^{-3}$  alors que ce taux est de l'ordre de  $10^{-2}$  pour les autres intervalles. Ce résultat permet de choisir les seuils de déclenchement des arbitrages dynamiques dans la loi QIS 5 à  $\beta = -5\%$  et  $\gamma = 3\%$ .

## 6.5 Impact sur le Best Estimate

Dans cette partie, nous allons évaluer l'impact de l'introduction d'une loi d'arbitrage structurel et dynamique dans le modèle ALS sur le BE et la PVFP. Rappelons que dans le scénario déterministe, il n'y a pas de comportement d'arbitrage dynamique. Un tel scénario permet donc de mesurer l'impact de la prise en compte d'une loi d'arbitrage structurel sur le BE et la PVFP. Une analyse de sensibilité des paramètres (ordre de la moyenne mobile et poids) a été réalisée afin de mesurer la robustesse de la loi d'arbitrage structurel. Le tableau suivant présente le BE et la PVFP en fonction des paramètres de la loi d'arbitrage structurel :

TABLEAU 6.9 – BE déterministe et PVFP déterministe par modèle

Modèle	Ordre	$\alpha$	BE	PVFP
Sans Arbitrage			73 237 299 179	3 494 599 795
Avec arbitrage structurel	3	2	73 316 849 463 (0.1086%)	3 415 052 404 (-2.2763%)
Avec arbitrage structurel	3	3	73 302 534 000 (0.0891%)	3 429 368 481 (-1.8666%)
Avec arbitrage structurel	3	4	73 300 637 551 (0.0865%)	3 431 264 623 (-1.8124%)
Avec arbitrage structurel	4	2	73 315 540 183 (0.1068%)	3 416 361 484 (-2.2388%)
Avec arbitrage structurel	4	3	73 296 473 849 (0.0808%)	3 435 428 950 (-1.6932%)
Avec arbitrage structurel	4	4	73 295 581 053 (0.0796%)	3 436 320 698 (-1.6677%)
Avec arbitrage structurel	5	2	73 312 637 244 (0.1029%)	3 419 264 156 (-2.1558%)
Avec arbitrage structurel	5	3	73 293 202 386 (0.0763%)	3 438 699 364 (-1.5996%)
Avec arbitrage structurel	5	4	73 291 357 508 (0.0738%)	3 440 545 231 (-1.5468%)

Le tableau 6.9 montre que l'introduction d'une loi d'arbitrage structurel fait augmenter le montant du BE déterministe et diminuer le montant de la PVFP déterministe. Pour un niveau de  $\alpha$  donné, le montant du BE déterministe diminue lorsque l'ordre de la moyenne mobile augmente et le montant du PVFP déterministe augmente. Cette tendance s'explique par un fort arbitrage vers le fonds en euros des assurés en 2019 et le taux d'arbitrage en 2019 est élevé par rapport aux taux d'arbitrage des autres années. En effet, plus on augmente l'ordre de la moyenne mobile, plus on dilue ce comportement d'arbitrage vers le fonds en euros ce

qui devrait faire baisser le montant de la BE et faire augmenter le montant de la PVFP. Notons qu'en général, pour un même montant d'épargne, le BE sur un contrat en euros est plus important que le BE sur un contrat en UC.

Pour un ordre de moyenne mobile donné, le montant du BE déterministe est une fonction décroissante de  $\alpha$  et la PVFP une fonction croissante de  $\alpha$ . En effet, une valeur de  $\alpha$  élevée signifie qu'on donne plus d'importance aux comportements d'arbitrage en 2020 (le plus récent et c'est un arbitrage euros vers UC).

Notons qu'il n'y a pas une forte variation du montant du BE et de la PVFP lorsqu'on fait varier les paramètres de la loi d'arbitrage structurel. Ce résultat permet de valider la robustesse de la loi d'arbitrage structurel par moyenne mobile pondérée.

Pour l'évaluation de l'impact de l'introduction d'une loi d'arbitrage dynamique, nous avons calculé le BE stochastique (moyenne des BE de 1000 simulations) et la PVFP stochastique. Le tableau suivant présente le montant du BE stochastique et de PVFP stochastique par loi d'arbitrage dynamique :

TABLEAU 6.10 – BE stochastique et PVFP stochastique par modèle

Modèle	Taux extremums	BE	PVFP
Sans Arbitrage		74 339 974 864	2 449 618 695
Loi QIS 5	Historique	74 621 081 315	2 168 579 806
		(0.3781%)	(-11.473%)
Loi QIS 5	Modèle	74 927 713 511	1 859 739 086
		(0.7906%)	(-24.080%)
GLM		74 643 142 923	2 145 520 527
		(0.4078%)	(-12.414%)

Il en ressort que le BE stochastique obtenu par la loi QIS 5 dépend fortement du choix des taux extremums. En plus d'avoir été calibré sur l'historique, l'impact du modèle GLM est assez similaire à l'impact de la loi QIS 5 avec des taux extremums de l'historique sur le BE stochastique et la PVFP stochastique.

---

## Conclusion

Le contexte de taux bas et la survenance des récentes crises financière et crise sanitaire met en difficulté les assureurs. En effet, une réorientation importante du placement des assurés peut avoir un impact sur la solvabilité et l'engagement des assureurs. La démarche mise en place dans cette étude pour le modèle individuel est l'approche fréquence/sévérité pour appréhender le comportement d'arbitrage des assurés. Le pas de temps utilisé dans la modélisation est annuel.

Plusieurs méthodologies ont été adoptées pour modéliser la fréquence (décision d'arbitrer ou non) tout au long de ce rapport de stage que ce soit du côté du choix du modèle que ce soit du côté de la méthode d'échantillonnage. Sur la base des indicateurs de choix de modèle (Taux de bien classés, Spécificité et Logloss), c'est le modèle XGBoost avec une base « oversampler » et « undersampler » qui présente le meilleur résultat. Le modèle met en avant à travers l'importance des variables, que ce sont surtout les variables qui sont liées à l'aversion au risque de l'assuré (Montant de provision mathématique, nombre de fonds en UC. . .) qui influent le plus la décision d'arbitrer. Les variables liées à la rationalité (Mode de gestion, Ancienneté. . .) semblent avoir moins d'impact sur le choix d'arbitrage. Notons aussi qu'aucun modèle de fréquence n'arrive à bien classer les assurés qui ont réalisé des arbitrages.

Pour le taux d'arbitrage, c'est le modèle XGboost en deux étapes qui arrivent à bien prédire le taux d'arbitrage des assurés. En effet, du point de vue explicatif et prédictif, le XGboost en deux étapes possède des meilleurs scores (RMSE,  $R^2$  et MAE) que la régression tanh. La première étape de la modélisation consiste à capter la composante structurelle des arbitrages à partir des variables liées aux caractéristiques du contrat et de l'assuré. Il s'avère que ce sont les variables montant de provision mathématique et Part en UC dans le contrat, c'est-à-dire les variables qui mesurent l'aversion au risque, qui influent le plus sur le montant ou taux d'arbitrage structurel de l'assuré. Sur la composante conjoncturelle, c'est le rendement du fonds en euros qui a le plus d'impact sur le taux d'arbitrage des assurés. Notons que lors de cette étude, le taux de rendement du fonds en euros utilisé est le véritable rendement des fonds en euros alors que celui du fonds en UC était la moyenne du rendement de l'indice CAC et de l'indice Eurostoxx, ce qui pourrait expliquer le résultat. Sur la base des résultats, nous pouvons dire qu'en général, l'aversion au risque des assurés a plus d'impact que sa rationalité sur son comportement d'arbitrage.

Pour la modélisation par groupe de contrats, les arbitrages structurels ont été modélisés par une moyenne mobile pondérée. Le Best Estimate déterministe est une fonction décroissante de l'ordre de la moyenne mobile et du facteur de décroissance du poids.

La loi QIS 5 et un modèle GLM ont été utilisés pour modéliser les arbitrages dynamiques. En plus d'être calibré sur l'historique, le modèle GLM a un impact similaire à la loi QIS 5 qui utilise les taux extrêmes de l'historique sur le Best Estimate stochastique et présente un résultat plus robuste. L'impact de la loi QIS 5 sur le Best Estimate stochastique dépend fortement de ses paramètres.

---

Toutefois, une modélisation au pas de temps mensuel et la prise en compte d'autres variables comme la notation ou le rating des obligations, la volatilité des fonds en unités de compte, le niveau d'éducation des assurés... aurait pu donner des meilleurs résultats et une meilleure compréhension du comportement d'arbitrage des assurés.

## BIBLIOGRAPHIE

- [1] Adrien Ehrhardt « *Formalization and study of statistical problems in Credit Scoring : Reject inference, discretization and pairwise interactions, logistic regression trees.* », Methodology [stat.ME]. Université de Lille, 2019. English. tel-02302691. <https://hal.archives-ouvertes.fr/tel-02302691/document>.
- [2] Camille Folgoas « *Modélisation des arbitrages sur les contrats d'assurance vie* », Mémoire d'actuariat de l'Institut des actuaires, 2017.
- [3] Cédric Asfa « *Le modèle Logit : Théorie et application* », 2016. INSEE.
- [4] Cem Ertur « *Méthodologies de test de la racine unitaire. [Rapport de recherche] Laboratoire d'analyse et de techniques économiques(LATEC).* », 36 p., Table, ref. bib. : 54 ref, 1998. hal-01527262.
- [5] David A. Dickey and Wayne A. Fuller, « *Likelihood Ratio Statistics for Autoregressive Time Series with a unit root* », Econometrica, Vol 49, N°, 1981. <http://www.u.arizona.edu/~rlo/readings/278800.pdf>
- [6] Denis Kwiatkowski, Peter Phillips , Peter Schmidt and Yongcheol Shin, « *Testing the null hypothesis of stationarity against the alternative of a unit root : How sure are we that economic time series have a unit root ?* », Journal of Econometrics, vol. 54, issue 1-3, 159-178, 1992.
- [7] Dimitri Delcaillau, « *Contrôle et Transparence des modèles complexes en actuariat* », Mémoire d'actuariat de l'Institut des actuaires, 2020.
- [8] FFA, « *Les contrats d'assurance en cas de vie* », Novembre 2020.
- [9] Guillaume Arquembourg, « *Modélisation de l'impact des arbitrages Euro/UC sur un portefeuille de contrats multisupports* », Mémoire d'actuariat de l'Institut des actuaires, 2014.
- [10] Introduction to Boosted Trees, <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>.
- [11] Karim Zennaf, « *Modèles d'arbitrages dynamiques dans le cadre des produits d'assurance-vie multisupports* », Mémoire d'actuariat de l'Institut des actuaires, 2012.

- 
- [12] Khaoula Lyoubi, « *Modélisation de la réponse des assurés à une incitation financière : Arbitrage entre fonds en euros et en unités de compte et politique de participation aux bénéfices* », Mémoire d'actuariat de l'Institut des actuaires, 2020.
- [13] Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone, « *Classification and Regression Trees* », Wadsworth & Brooks, 1984.
- [14] Meyez Abdellatif, « *Modèles des arbitrages dynamiques pour les contrats d'épargne multi-supports* », Mémoire d'actuariat de l'Institut des actuaires.
- [15] MIT « *Generalized Linear Models (GLMs)* », Statistics for Applications.
- [16] Omar Hamaoui, « *Modélisation du portefeuille de la gestion privée sous Solvabilité 2* », Mémoire d'actuariat de l'Institut des actuaires, 2012.
- [17] Petar Radkov « *The Mean Reversion Stochastic Processes Applications in Risk Management* », 2010. [https://www.researchgate.net/publication/247777389\\_The\\_Mean\\_Reversion\\_Stochastic\\_Processes\\_Applications\\_in\\_Risk\\_Management](https://www.researchgate.net/publication/247777389_The_Mean_Reversion_Stochastic_Processes_Applications_in_Risk_Management).
- [18] Sander Devriendt, Katrien Antonio, Tom Reynkens and Roel Verbelen « *Sparse regression with Multi-type Regularized Feature modeling* », Insurance : Mathematics and Economics, 96 :248–261, 2021. ISSN 01676687. <https://linkinghub.elsevier.com/retrieve/pii/S0167668720301608>.
- [19] Tianqi Chen and Carlos Guestrin « *XGBoost : A Scalable Tree Boosting System* », 2016. arXiv :1603.02754v3.
- [20] WikiStat « *Arbres binaires de décision* », Université de Toulouse.
- [21] WikiStat « *Agrégation de modèles* », Université de Toulouse.
- [22] Xavier D'Haultfoeuille et Pauline Givord « *La régression quantile en pratique* », Économie et statistique N° 471, 2014.
- [23] Yogesh J. Bagul « *A smooth transcendental approximation to  $|x|$*  », International J.of Math. Sci. Engg. Appls. (IJMSEA), Vol. 11 (II), pp.213 - 217, 2017. hal-01713196.

# ANNEXES

## Analyse bivariée

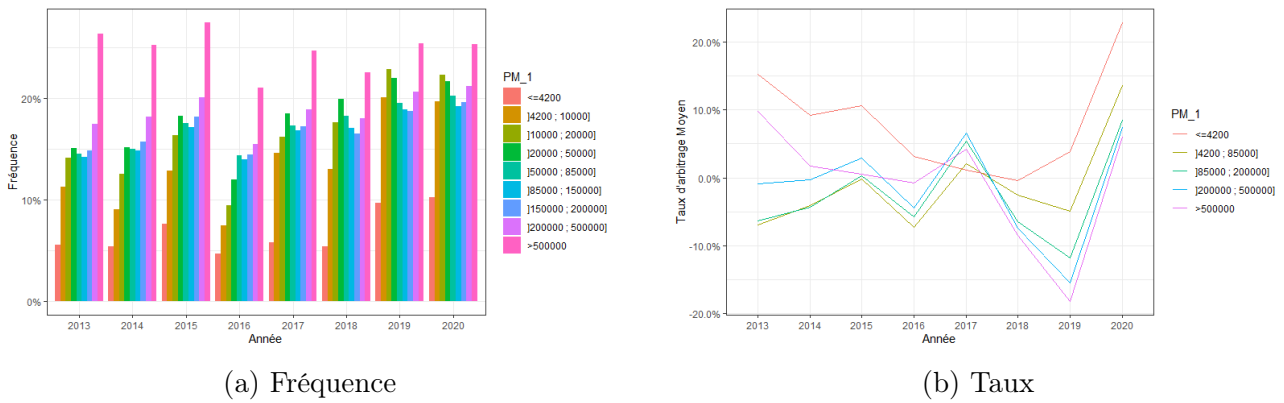


FIGURE 12 – Arbitrage par PM

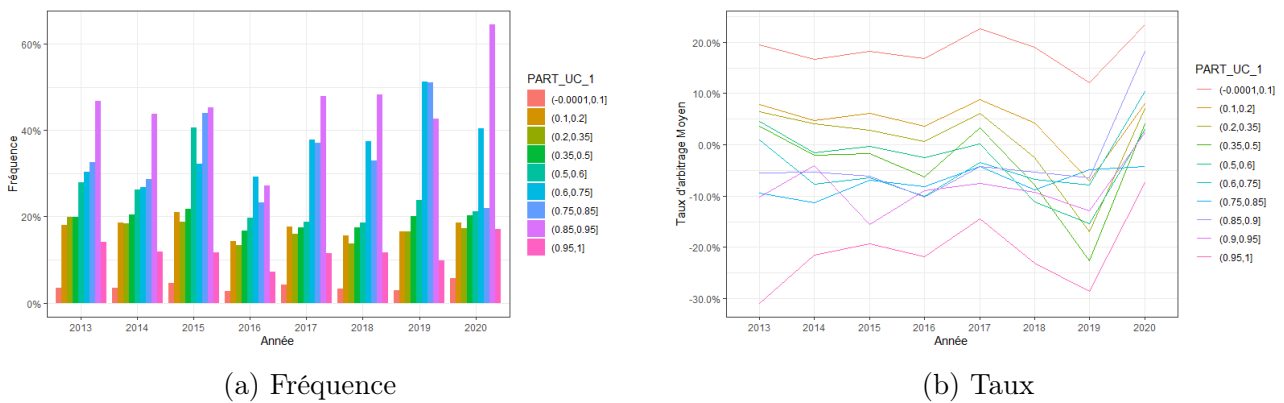


FIGURE 13 – Arbitrage par part en UC

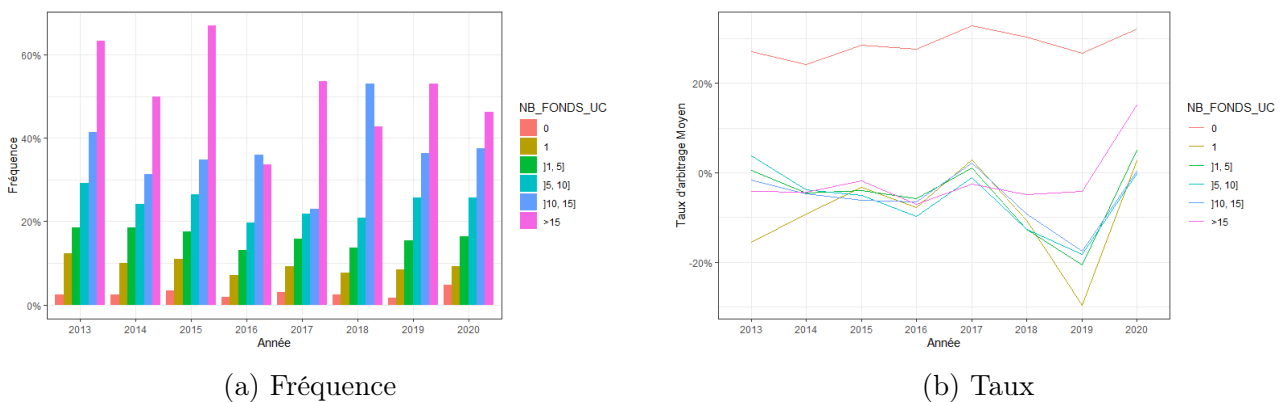
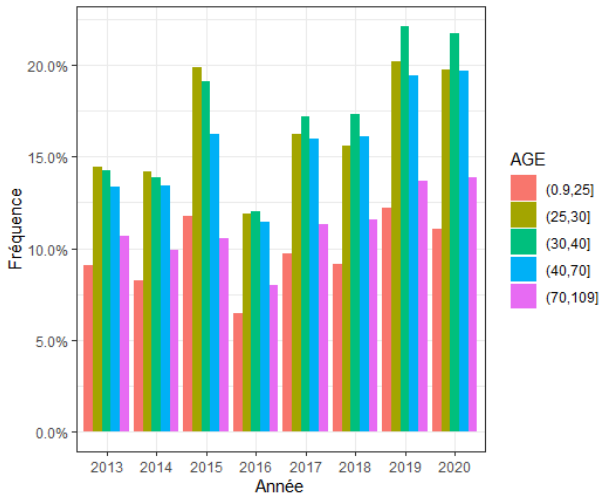


FIGURE 14 – Arbitrage par nombre de fonds en UC



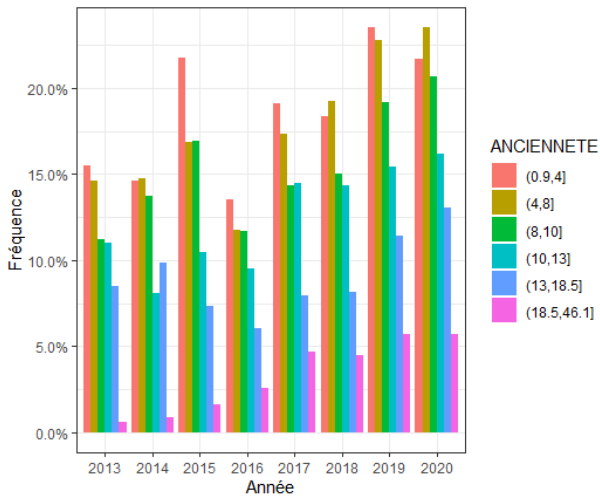


(a) Fréquence

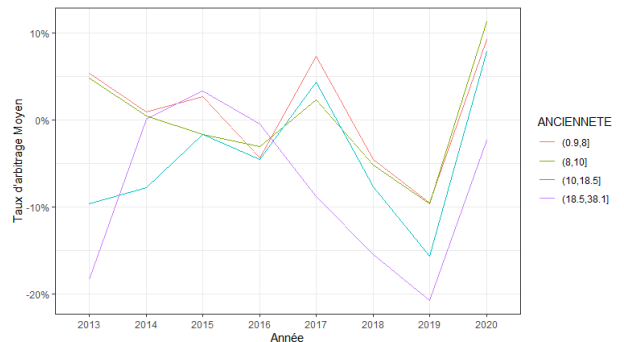


(b) Taux

FIGURE 15 – Arbitrage par âge



(a) Fréquence

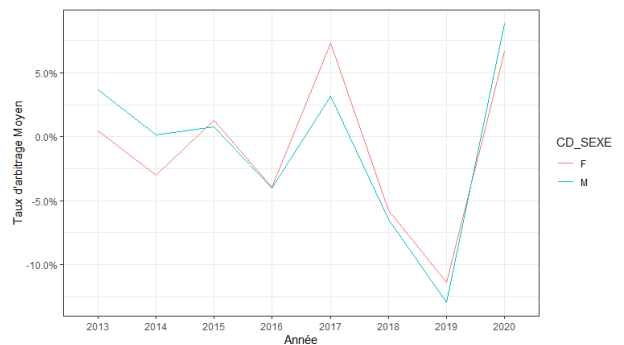


(b) Taux

FIGURE 16 – Arbitrage par ancienneté

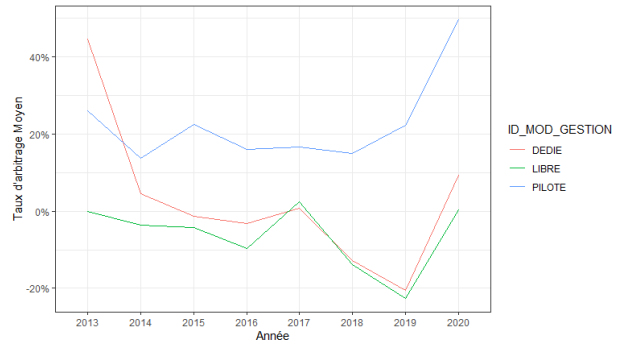
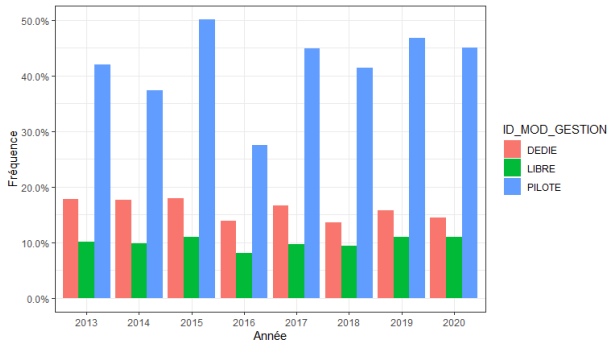


(a) Fréquence



(b) Taux

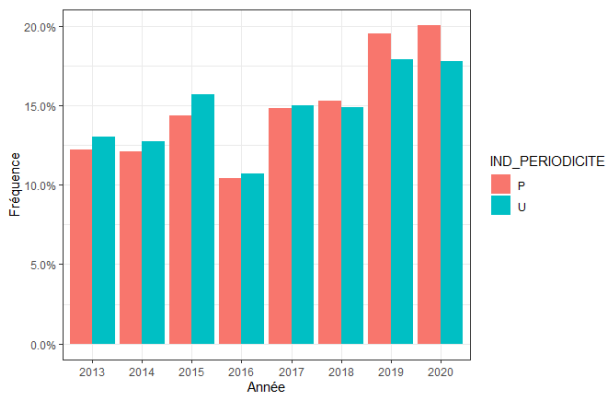
FIGURE 17 – Arbitrage par sexe



(a) Fréquence

(b) Taux

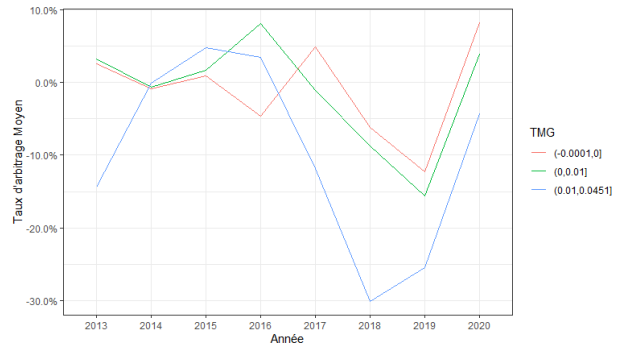
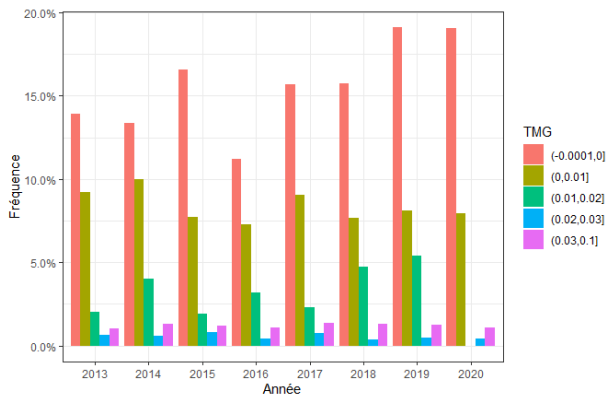
FIGURE 18 – Arbitrage par mode de gestion



(a) Fréquence

(b) Taux

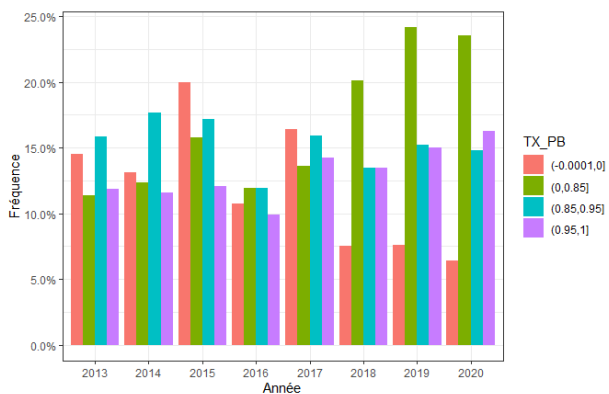
FIGURE 19 – Arbitrage par indice de périodicité des primes



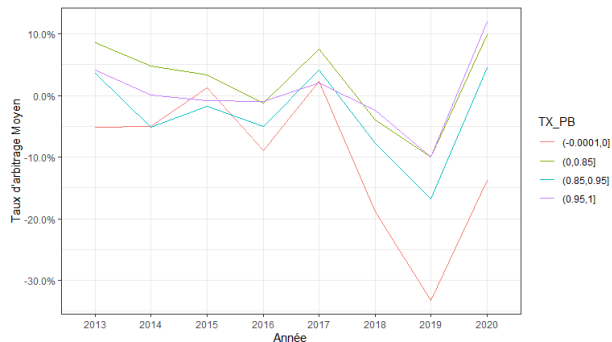
(a) Fréquence

(b) Taux

FIGURE 20 – Arbitrage par TMG



(a) Fréquence



(b) Taux

FIGURE 21 – Arbitrage par Taux de PB

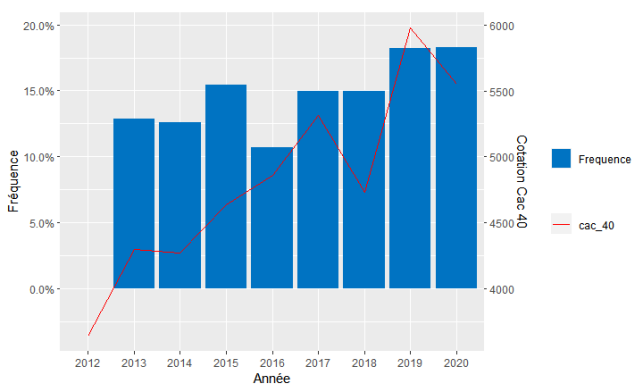
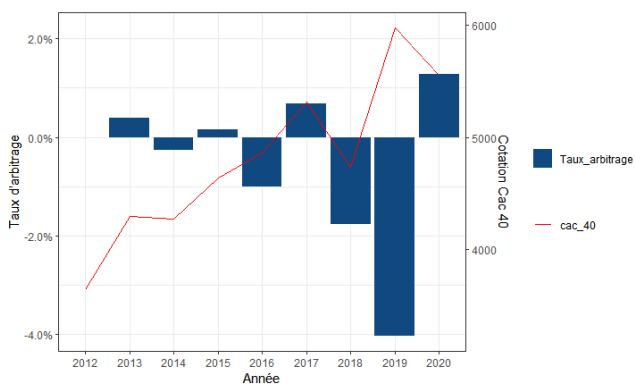


FIGURE 22 – Evolution des arbitrages et de la cotation du CAC 40

---

## Validation croisée

Le principe générale de la validation croisée est le suivant : On découpe l'échantillon d'apprentissage en  $L$  sous échantillons ( $L = 10$  pour se fixer les idées), Pour une valeur d'un paramètre  $\lambda$  , on estime le modèle sur  $L - 1$  sous échantillons et on mesure sa performance (RSME par exemple) sur le sous échantillon mis de côté. On réitère cette étape en faisant varier le sous échantillon de test, jusqu'à ce que les  $L$  sous échantillons aient été utilisé pour tester le modèle. On moyennise alors l'erreur de ces différentes étapes. On recommence tout le processus pour différentes valeurs de  $\lambda$ . Le  $\lambda$  retenu est celui qui minimise l'erreur moyenne de prédiction.

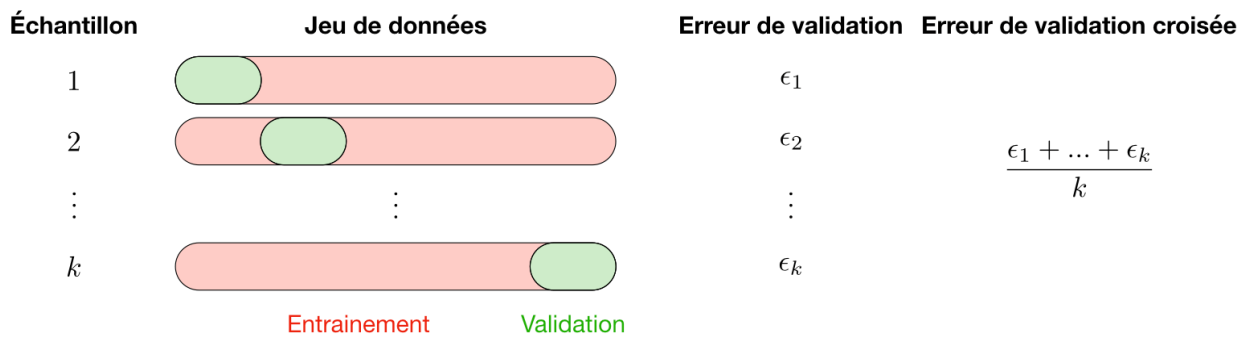


FIGURE 23 – Principe générale de la validation croisée

---

## Corrélation de Pearson

Le coefficient de Pearson est un indice reflétant une relation linéaire entre deux variables continues. Le coefficient de corrélation varie entre -1 et 1, 0 reflétant une relation nulle entre les deux variables, une valeur négative (corrélation négative) signifiant que lorsqu'une des variable augmente, l'autre diminue ; tandis qu'une valeur positive (corrélation positive) indique que les deux variables varient ensemble dans le même sens. Voici des exemples illustrant les 3 situations :

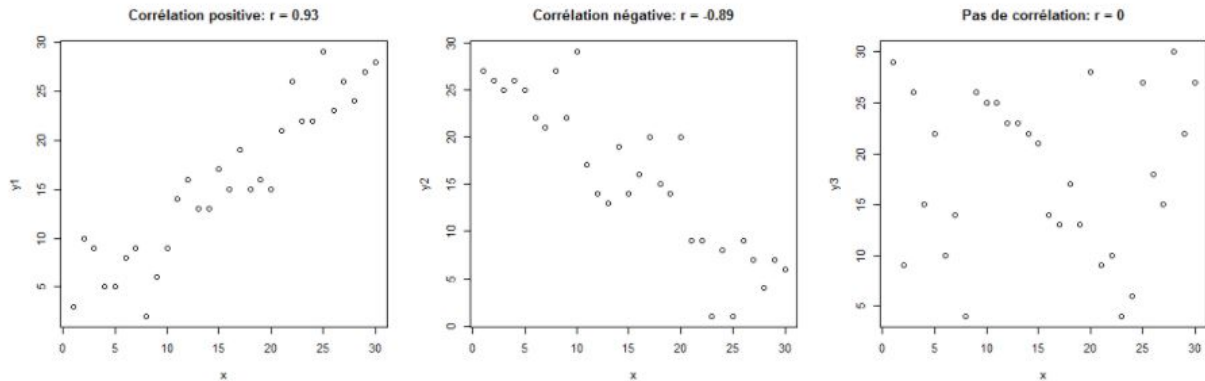


FIGURE 24 – Exemples de corrélation

Voici la formule pour calculer le coefficient de corrélation de Pearson :

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X}) (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Source : [http://www.biostat.ulg.ac.be/pages/Site\\_r/corr\\_pearson.html](http://www.biostat.ulg.ac.be/pages/Site_r/corr_pearson.html)

---

## Khi 2

On souhaite "mesurer" l'écart entre une distribution donnée et celle que l'on aurait, en théorie, si certaines hypothèses étaient vérifiées. Dans notre cas, l'hypothèse est que deux variables sont indépendantes, et on voudrait savoir si cette hypothèse est probablement vraie ou probablement fausse.

Le test du khi2 nous permet d'avoir une réponse. Ce test consiste à calculer un nombre à partir des deux distributions, réelle et théorique. Ce nombre est ensuite à comparer avec des tables, que l'on trouve dans tous les manuels de statistiques, et sans doute aussi sur le web . Selon la valeur de ce nombre, le nombre de modalités des variables, et le degré de confiance voulu, la table dit si l'hypothèse est statistiquement raisonnable ou non.

### Les formules mathématiques

Etant donné un tableau représentant la distribution jointe de deux variables, on utilise les notations suivantes :

- $n_{ij}$  est l'effectif contenu dans case repérée par la ligne  $i$  et la colonne  $j$ ,
- $n_{i.}$  est l'effectif marginal de la ligne  $i$ ,
- $n_{.j}$  est celui de la colonne  $j$ ,
- et  $n$  est l'effectif total (la taille de la population).

TABLEAU 11 – Exemple tableau de contingence

$n_{11}$	$n_{12}$	$n_{13}$	$n_{1.}$
$n_{21}$	$n_{22}$	$n_{23}$	$n_{.1}$
$n_{.2}$	$n_{.3}$	$n$	$n_{ij}$

D'autre part, on note  $n_{ij}^*$  l'effectif que l'on aurait si les variables étaient indépendantes. Dans ces conditions, le khi2 est donné par la formule suivante, où la somme porte sur toutes les lignes et toutes les colonnes du tableau (dans le tableau ci-dessus, on a  $i = 2$  et  $j = 3$ ) :

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

Par définition de l'indépendance , on sait que

$$n_{ij}^* \times n = n_{i.} \times n_{.j}$$

En remplaçant et avec un petit peu de calcul, la formule du khi2 précédente peut s'écrire sous

---

cette forme un peu plus simple, et surtout beaucoup plus facile à programmer sur tableur :

$$\chi^2 = n \left( \sum_{i,j} \frac{n_{ij}^2}{n_i \cdot n_j} - 1 \right)$$

Source : [https://www.univ-montp3.fr/miap/ens/AES\\_IDS/TD4/Exo3.html](https://www.univ-montp3.fr/miap/ens/AES_IDS/TD4/Exo3.html)

---

## V de Cramer

Contrairement au  $\chi^2$ , il reste stable si l'on augmente la taille de l'échantillon dans les mêmes proportions inter-modalités. Il est basé sur le  $\chi^2$  maximal que le tableau de contingence pourrait théoriquement produire : ce dernier aurait alors une seule case non nulle par ligne ou par colonne (selon que le tableau a plus de lignes ou plus de colonnes). Ce  $\chi_{\max}^2$  théorique est égal à l'effectif multiplié par le plus petit côté du tableau (nombre de lignes ou de colonnes) moins 1. Par exemple un tableau de  $2 \times 3$  avec un effectif de 100 a pour  $\chi_{\max}^2$  :  $100 \times (2 - 1) = 100$ .

Le  $V$  de Cramer est la racine carrée du  $\chi^2$  divisé par le  $\chi_{\max}^2$

$$v = \sqrt{\frac{\chi^2}{\chi_{\max}^2}} = \sqrt{\frac{\chi^2}{n \times [\min(l, c) - 1]}}$$

Plus  $V$  est proche de zéro, plus il y a indépendance entre les deux variables étudiées. Il vaut 1 en cas de complète dépendance puisque le  $\chi^2$  est alors égal au  $\chi_{\max}^2$  (il prend une valeur comprise entre 0 et 1).

Source : <http://www.jybaudot.fr/Inferentielle/associations.html>



# SHAP pour un individu donné dans la base

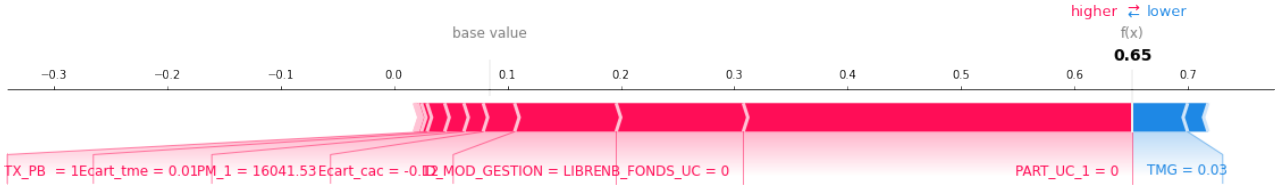


FIGURE 25 – Force plot SHAP