Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Madame / Monsieur Lucas GILMAS

Titre du mémoire : Modélisation des arbitrages sur un portefeuille d'épargne patrimoniale

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 : Signature :	Entreprise : Nom : A MONDIALE PARTENAIRE 14-16, Bld Malesherbes 75379 PARIS CEDEX 08 S.A. au capital de 73 413 150 €uros R.C.S. Paris B 313 689 713 Entreprise régie par le Code des Assurances Signature :
	Directeur de mémoire en entreprise Nom : GOBLET Marion Signature :
Membres présents du jury de l'Institut des Actuaires : Signature :	Invité : Nom : Signature :
	Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité) Signature du responsable entreprise : Signature du candidat :

RESUME

La proportion des contrats d'assurance vie multisupports n'a fait qu'augmenter depuis leur introduction en 1996. Ils offrent la possibilité aux assurés d'arbitrer entre le fond Euro et les fonds UC, dont les performances varient et dépendent du contexte économique.

L'objectif de ce mémoire est de comprendre et d'anticiper ces mouvements d'arbitrage pour permettre à l'assureur d'améliorer sa solvabilité et ses prédictions lors des modélisations prospectives, mais aussi d'éviter à l'assureur de subir des pertes liées à des ventes d'actifs en situation de moins-value.
Pour cela, différentes méthodes statistiques ont été mises en place, des modèles linéaires jusqu'aux agrégations de modèles non linéaires d'apprentissage automatique. Tous les modèles implémentés se sont faits à la maille contrat avec un pas de temps annuel, la variable cible étant le montant arbitré (net vers UC). Les variables classiques telles que l'âge, l'ancienneté, la PM se sont avérées insuffisantes pour comprendre et prédire ces mouvements. Des données de différentes natures ont dû être utilisées, en commençant par des variables construites telles que le nombre d'arbitrages déjà effectué, les plus ou moins-values réalisées l'année précédente jusqu'aux variables conjoncturelles tels que le taux de l'OAT France 10 ans et la valeur du CAC 40.

Pour éviter le sur-apprentissage, les paramètres des modèles ont été fixés en utilisant la validation croisée et les modèles ont été testés sur les données de 2020 et 2021 qui n'ont pas servi à la construction de ces derniers.

La modélisation de ces comportements a ensuite été intégrée dans les modélisations prospectives pour analyser les différents impacts, notamment sur la rentabilité de l'assureur et la valeur intrinsèque de son portefeuille.

Mots clés : Arbitrage, fond Euro, fond UC, Machine Learning, forêts aléatoires, Embedded Value

ABSTRACT

The proportion of multiple unit-linked policies has only increased since their introduction in 1996. They offer the possibility for insured to arbitrate between the Euro fund and the UL funds, whose performance varies and depends on the economic context.

The objective of this thesis is to understand and anticipate these arbitrage movements to enable the insurer to improve its solvency and its predictions during prospective modeling, but also to prevent the insurer from suffering losses related to sales of assets at a loss.

For this, different statistical methods have been implemented, from linear models to aggregations of non-linear machine learning models. All the models implemented are made at the contract level with an annual time step, the target variable being the arbitrated amount (net to UL). Conventional variables such as age, seniority, provision reverse, etc. have proven insufficient to understand and predict these movements. Different type of data had to be used, starting with constructed variables such as the number of arbitrages already carried out, the capital gains or losses realized the previous year, etc. up to short-term variables such as France 10-Year Bond Yield, the value of the CAC 40, …

To avoid over-fitting, model parameters were fixed using cross-validation and models were tested on 2020 and 2021 data that were not used to build them.

The modeling of these behaviors was then integrated into the prospective modeling to analyze the various impacts, in particular on the profitability of the insurer and the intrinsic value of its portfolio.

Keywords : Arbitrage, Euro fund, UL fund, machine learning, random forests, Embedded Value

REMERCIEMENTS

Ce mémoire ayant été réalisé dans le cadre de mon alternance au sein de la Direction Actuariat Produits et Finances (DAPF) de AG2R LA MONDIALE, j'aimerais remercier toutes les personnes qui ont contribué à sa réalisation.

Je tiens tout d'abord à remercier ma tutrice Marion GOBLET, pour sa confiance, son aide, sa pédagogie et son soutien dans les différentes missions qu'elle m'a confiées ainsi que dans la réalisation de ce mémoire.

De même je souhaite remercier Sidick SIMRICK, responsable du service actuariat de la DAPF, et plus globalement tous les membres de l'équipe, pour m'avoir accueilli de cette manière, mais également pour leur disponibilité et leur précieux conseils lors de la réalisation de ce mémoire.

Je remercie également Gaëlle COSTILLE, directrice de la DAPF, pour ses remarques pertinentes, et aussi pour sa bonne humeur et sa bienveillance qui sont à l'image de cette direction.

Aussi, je remercie Nicolas BOUSQUET, mon tuteur académique, pour avoir répondu à mes questions tout au long de la rédaction de cette étude, ainsi que tous mes professeurs à l'ISUP et à l'UGE pour m'avoir transmis leurs connaissances et pour la qualité de leurs enseignements.

Pour finir, je remercie ma famille et mes proches pour leur soutien.

NOTE DE SYNTHESE

Depuis leur introduction en 1996, les contrats multisupports n'ont cessé de se développer et représentent aujourd'hui la quasi-totalité des contrats d'assurance vie souscrits chez AG2R LA MONDIALE en 2021.

Avec l'apparition de ces contrats sont apparus de nouveaux risques pour l'assureur, plus particulièrement les risques liés à l'option d'arbitrage. Cette option offre la possibilité à l'assuré de transférer tout ou une partie de son épargne d'un support à un autre.

Or, en cas de réorientation conséquente de l'épargne des assurés vers un fond, l'assureur est amené à céder précipitamment des actifs financiers pour pouvoir les réinvestir sur le(s) support(s) choisi(s) par l'assuré. Ce faisant, l'assureur s'expose au risque de devoir acheter et/ou vendre ses actifs au mauvais moment et causer des pertes importantes. De plus, les marges techniques et financières ne sont pas les mêmes suivant les fonds, impactant la valeur du portefeuille et les modélisations prospectives. Pour finir, comme les fonds déposés sur l'Euro sont garantis, cela nécessite pour l'assureur une immobilisation du capital plus importante que les fonds déposés sur l'UC.

L'étude porte donc sur la modélisation de ces comportements d'arbitrage. Elle sera opérée à la maille contrat dans un soucis de précision et d'exhaustivité. Le pas de temps retenu sera annuel, offrant le meilleur compromis entre précision et parcimonie des données. Le modèle sera testé sur les données de 2020 et 2021, qui n'auront pas été utilisées dans la construction des modèles. L'objectif étant de prédire le montant net arbitré par année, que l'on définit simplement comme la différence entre la somme des montants arbitrés sur l'UC et la somme des montants arbitrés sur l'Euro dans l'année.

$$Montant\ net\ arbitré_n = \sum Montant\ arbitré\ EU \rightarrow UC_n - \sum Montant\ arbitré\ UC \rightarrow EU_n$$

Notamment dû à une procédure d'arbitrage extrêmement lourde, en moyenne sur une année, seulement 6% des contrats effectuent un arbitrage. Ainsi, la modélisation directe de cette variable à la maille contrat s'est avérée impossible (problème de sparsity).

La méthode de modélisation qui s'est avérée être la plus efficace a été d'utiliser une approche par fréquence/sévérité probabiliste. L'idée derrière est de modéliser l'espérance du montant arbitré et non pas le montant lui-même. Cette méthode nous permet ainsi de décomposer le problème. En effet, soit M_i le montant net arbitré de l'individu i pour l'année à venir et A_i la variable qui prend la valeur 1 si l'individu i arbitre l'année à venir et 0 sinon, on a par le théorème de l'espérance totale :

$$\begin{aligned}
\mathbb{E}(M_i) &= \mathbb{E}(\mathbb{E}(M_i|A_i)) \\
&= \sum_a \mathbb{E}(M_i|A_i = a)\mathbb{P}(A_i = a) \\
&= \mathbb{E}(M_i|A_i = 0)\mathbb{P}(A_i = 0) + \mathbb{E}(M_i|A_i = 1)\mathbb{P}(A_i = 1) \\
&= \mathbb{E}(M_i|A_i = 1)\mathbb{P}(A_i = 1)
\end{aligned}$$

La simplification de la dernière étape provenant du fait que l'espérance du montant arbitré conditionné au fait qu'il n'y a pas d'arbitrage est évidemment nul.

Il s'est également avéré plus simple de modéliser le pourcentage de la PM arbitré plutôt que le montant lui-même.

$$Taux\ arbitrage\ net = \frac{Montant\ net\ arbitré}{PM}$$

En effet, le taux d'arbitrage net est borné entre -1 et 1. -1 et 1 correspondant respectivement à une réorientation complète de l'épargne vers le fond Euro (resp UC). La PM étant déterministe, l'équation précédente devient :

$$\mathbb{E}(M_i) = \mathbb{E}(PM_i Y_i) = PM_i \mathbb{E}(Y_i) = PM_i \mathbb{E}(Y_i | A_i = 1)\mathbb{P}(A_i = 1)$$

Avec Y_i le taux d'arbitrage net de l'individu i.

L'espérance du taux net arbitré conditionné au fait d'arbitrer a été estimée par un modèle de régression construit uniquement sur les données où il y a effectivement eu un arbitrage. La probabilité d'arbitrer a été estimée par un modèle de classification. Dans les deux cas, les modèles retenus sont des forêts aléatoires dont les paramètres (nombre d'arbres, profondeur maximal, …) ont été fixés en utilisant la *cross validation*.

Tableau 1 : Performances des modèles de classification sur les données test

Méthode	Hyperpamètres optimaux	Taux de précision	Log Loss	AUC Precision-Recall
GLM	Pénalité : Ridge Lambda : 0.1	93,78%	**0,221**	**0,26**
CART	Profondeure maximale : 8 Fonction d'hétérogénéité : Indice de Gini Nombre d'observation minimale sur une feuille : 190	94,26%	**0,183**	**0,32**
Bagging	Nombre d'arbre : 50 Taille des échantillions boostrap : 75% Profondeure maximale : 8 Fonction d'hétérogénéité : Entropie Nombre d'observation minimale sur une feuille : 400	94,28%	**0,178**	**0,33**
Random Forest	Nombre d'arbre : 75 Nombre de variable : p/2 Profondeure maximale : 8 Fonction d'hétérogénéité : Indice de Gini Nombre d'observation minimale sur une feuille : 100	94,32%	**0,171**	**0,36**

Le tableau 1 montre que le modèle de classification qui a les meilleures performances est la forêt aléatoire (la meilleure *Log Loss* étant la plus faible). L'analyse de l'importance des variables indique que ce sont les variables liées aux comportements d'arbitrage passé (nombre d'arbitrage en $n-1$ et depuis $n-1$), ainsi que la PM et la proportion d'UC qui impactent le plus la probabilité d'effectuer un arbitrage.

Tableau 2: Performances des modèles de régression sur les données de test

Méthode	Hyperpamètres optimaux	Erreur absolue moyenne train	Erreur absolue moyenne test	R^2
GLM	Pénalité : aucune Lambda : 0	0,272	**0,277**	**0,052**
CART	Profondeure maximale : 7 Fonction d'hétérogénéité (fixée) : MSE Nombre d'observation minimale sur une feuille : 220	0,23	**0,267**	**0,132**
Bagging	Nombre d'arbre : 150 Taille des échantillions boostrap : 80% Profondeure maximale : 9 Fonction d'hétérogénéité (fixée) : MSE Nombre d'observation minimale sur une feuille : 50	0,222	**0,259**	**0,165**
Random Forest	Nombre d'arbre : 150 Nombre de variable : √p Profondeure maximale : 9 Fonction d'hétérogénéité (fixée) : MSE Nombre d'observation minimale sur une feuille : 50	0,167	**0,184**	**0,290**

Le tableau 2 montre également que parmi les différents modèles de régression implémentés, le meilleur est la forêt aléatoire. L'analyse de l'importance des variables indique que la proportion d'UC, l'écart entre les plus-values réalisées et le taux de l'OAT France 10ans en n-1 et plus globalement les variables liées aux plus ou moins-values réalisées en $n-1$ sont celles qui influencent le plus le taux net arbitré.

En modélisant l'espérance du taux arbitré et non pas le taux lui-même, on obtient naturellement des prédictions individuelles relativement mauvaises, mais la loi des grands nombres nous permet d'observer une convergence quand on agrège les résultats à l'année.

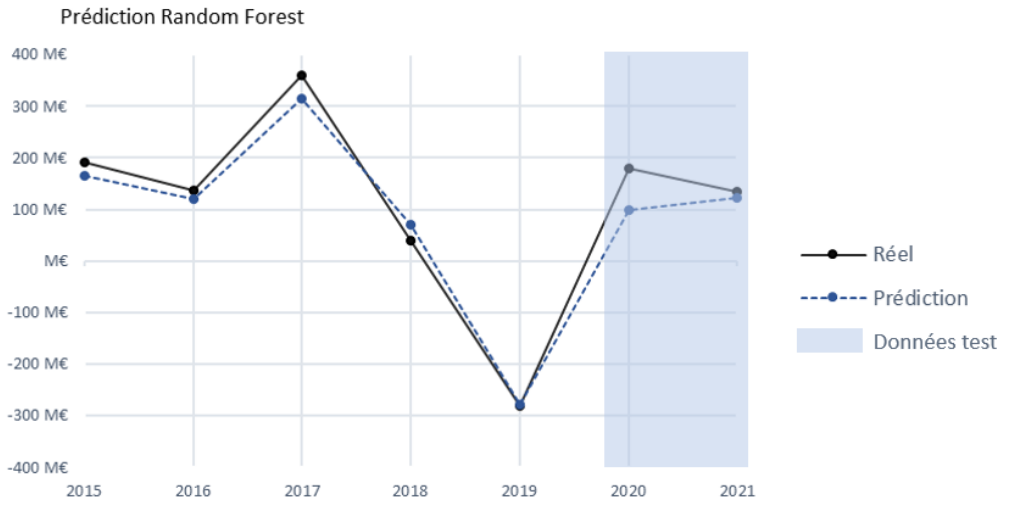


Figure 1 : Prédiction du montant net annuel arbitré par l'agrégation des modèles

Un modèle simplifié a ensuite été implémenté dans nos modélisations prospectives dans le but d'étudier l'impact des arbitrages sur l'Embedded Value (EV) de la compagnie ainsi que sur son Taux de Rendement Interne (TRI).

Tableau 3 : Impacts sur la modélisation prospective

	Sans arbitrage	Avec arbitrage
VA produit assurance (1)	1740,46	1737,56
VA impots sur les sociétés (2)	184,21	197,12
VA charges (3)	1027,28	974,42
VA résultats futurs (4)=(1)-(2)-(3)	528,97	566,02
MS initial (5)	501,60	502,96
VA dotation MS (6)	-149,36	-156,65
VA intérêt ms (7)	207,96	203,81
VA IS intérêt ms (8)	53,71	52,64
Cout immo fp (9)=(5)+(6)-(7)+(8)	198,00	195,15
EV (10)=(3)-(9)	384,60	423,76
TRI (11)	8,47%	8,98%

Le tableau 3 indique que ces indicateurs augmentent lorsqu'on intègre notre loi. Ce résultat provient d'un enchainement des phénomènes liés aux prédictions faites par notre loi, qui sont majoritairement dans le sens de l'UC. Ceci implique une baisse de la PM Euro au profit de la PM UC, faisant baisser les charges de réassurance ainsi que l'immobilisation de fonds propres.

EXECUTIVE SUMMARY

Since their introduction in 1996 multiple unit-linked policies have grown steady representing almost the totality of the policies sold by AG2R LA MONDIALE today.

New risks have emerged because of them particularly through the arbitrage option which allows policyholders to transfer some or all their savings from one fund to another.

A massive reorientation of the policyholders' savings toward a given fund will lead the insurer to sell his financial assets to be able to reinvest on the selected fund(s). By doing so, the insurer might sell and/or buy his assets at the wrong time resulting in important losses. Moreover, technical, and financial margins are not equal on both funds impacting the portfolio value and the prospective modeling. Finally, amounts invested in euro fund are guaranteed which need more allocated capital than UL funds.

The study is about modeling arbitrage behaviour and more precisely the annual net amount arbitrated which is simply define as the difference between the sum of all the amount arbitrated towards UL funds and the sum of all the amount arbitrated towards euro fund in the year.

$$Net\ amount\ arbitrated_n = \sum Amount\ arbitrated\ Euro \rightarrow UL_n - \sum Amount\ arbitrated\ UL \rightarrow Euro_n$$

Partly because of an extremely heavy arbitrage procedure, on average over a year, only 6% of the contracts make an arbitrage. Therefore, direct modeling of this variable to the contract mesh was found impossible (sparsity problem).

The modeling method the most efficient was to model the expectation of the arbitrated amount instead of the amount itself. This method helps us splitting the issue. Indeed, let M_i the arbitrated amount of the person i for each coming year and A_i the variable which takes the value 1 if the person i arbitrate the coming year and 0 otherwise. Given by the law of total expectation we have:

$$\begin{aligned}
\mathbb{E}(M_i) &= \mathbb{E}(\mathbb{E}(M_i|A_i) \\
&= \sum_a \mathbb{E}(M_i|A_i = a)\mathbb{P}(A_i = a) \\
&= \mathbb{E}(M_i|A_i = 0)\mathbb{P}(A_i = 0) + \mathbb{E}(M_i|A_i = 1)\mathbb{P}(A_i = 1) \\
&= \mathbb{E}(M_i|A_i = 1)\mathbb{P}(A_i = 1)
\end{aligned}$$

The final step simplification coming from the fact that the expectation of the arbitrated amount conditional upon that there is no arbitrage is obviously null.

Furthermore, it has been simpler to model the percentage of the Mathematical Provisions (MP) arbitrated instead of the amount itself.

$$Net\ arbitrage\ rate = \frac{Net\ amount\ arbitrated}{MP}$$

Indeed, the net arbitration rate is bounded between -1 and 1. -1 and 1 corresponding respectively to a complete reorientation of savings towards the Euro fund (resp UL). The MP being deterministic, the previous equation becomes:

$$\mathbb{E}(M_i) = \mathbb{E}(PM_iY_i) = PM_i\mathbb{E}(Y_i) = PM_i\mathbb{E}(Y_i|A_i = 1)\mathbb{P}(A_i = 1)$$

With Y_i the net arbitration rate of the person i.

The net arbitration rate expectation conditional on the fact of arbitrating was estimated by a regression model built only on the data where there was indeed an arbitrage. The probability of arbitrating was estimated by a classification model. In both cases, the models retained are random forests whose parameters (number of trees, maximum depth, etc.) have been set using cross validation.

Tableau 4 : Classification model performances on test data

Method	Optimal hyperparameters	Accuracy	Log Loss	AUC Precision-Recall
GLM	Penality : Ridge Lambda : 0.1	93,78%	**0,221**	**0,26**
CART	Maximum depth : 8 Splitting function : Indice de Gini Minimum number of samples at a lead node : 190	94,26%	**0,183**	**0,32**
Bagging	Number of trees : 50 Size of boostrap sample : 75% Maximum depth : 8 Splitting function : Entropie Minimum number of samples at a lead node : 400	94,28%	**0,178**	**0,33**
Random Forest	Number of trees : 75 Number of features : p/2 Maximum depth : 8 Splitting function : Indice de Gini Minimum number of samples at a lead node : 100	94,32%	**0,171**	**0,36**

Table above shows that the classification model which has the best results is the random forest (the best *Log Loss* being the lowest). The analysis of the importance if the variables indicates that it is the variables linked to past arbitrage behavior (number of arbitrages in $n-1$ and since $n-1$), as well as the MP and the proportion of UL that impact the higher the probability of performing an arbitrage.

Tableau 5 : Regression model performances on test data

Method	Optimal hyperparameters	MAE train	MAE test	R^2
GLM	Penality : none Lambda : 0	0,272	**0,277**	**0,052**
CART	Maximum depth : 7 Splitting function (fixed) : MSE Minimum number of samples at a lead node : 220	0,23	**0,267**	**0,132**
Bagging	Number of trees : 150 Size of boostrap sample : 80% Maximum depth : 9 Splitting function (fixed) : MSE Minimum number of samples at a lead node : 50	0,222	**0,259**	**0,165**
Random Forest	Number of trees : 150 Number of features : √p Maximum depth : 9 Splitting function (fixed) : MSE Minimum number of samples at a lead node : 50	0,167	**0,184**	**0,290**

Table 5 also shows that among the different regression models implemented the best is the random forest. The analysis of the importance of the variables indicates that the UL proportion, the difference between the capital gains realized and the rate of the France 10-year bond yield in $n-1$ and more generally the variables linked to the capital gains or losses realized in $n-1$ are those that influence the net arbitrated rate the most.

By modeling the expectation of the net arbitrated rate and not the rate itself, we naturally obtain relatively poor individual predictions, but with the law of large numbers we observe a convergence when aggregating the results over the year.

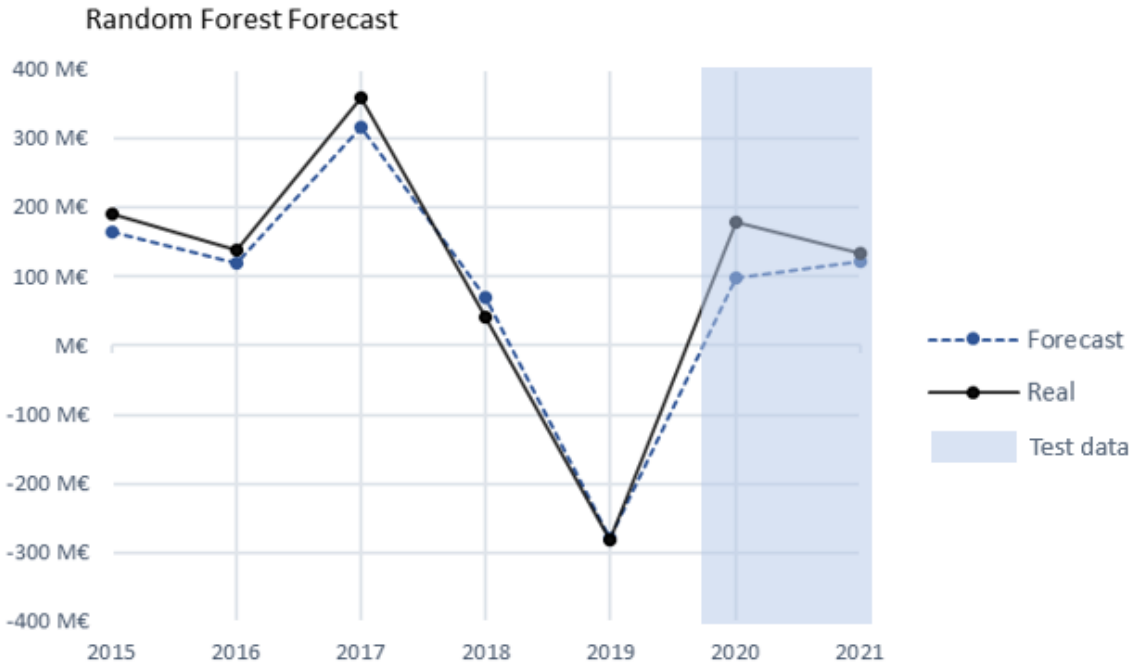


Figure 2: Annual net amount arbitrated forecast by models aggregation

A simplified model was then implemented in our prospective modeling to study the impact of arbitrages on the company's Embedded Value (EV) as well as on its Internal Rate of Return (IRR).

Tableau 6 : Impacts on prospective modelisation

	Without arbitrage	With arbitrage
AV insurance revenues (1)	1740,46	1737,56
AV corporate taxes (2)	184,21	197,12
AV expenses (3)	1027,28	974,42
AV future outcomes (4)=(1)-(2)-(3)	528,97	566,02
MS initial (5)	501,60	502,96
AV dotation MS (6)	-149,36	-156,65
AV intest ms (7)	207,96	203,81
AV IS interest ms (8)	53,71	52,64
Cost allocated capital (9)=(5)+(6)-(7)+(8)	198,00	195,15
EV (10)=(3)-(9)	384,60	423,76
IRR (11)	8,47%	8,98%

Tab 6 shows that those indicators increase when integrating our law. This result comes from a chain of reactions induced by the predictions of our law which are mainly toward UL funds. This resulting in a reduction of the Euro MP in favor of the UL MP which automatically decrease reinsurances expenses as well as allocated capital.

TABLE DES MATIERES

RESUME 3
ABSTRACT 4
REMERCIEMENTS 5
NOTE DE SYNTHESE 6
EXECUTIVE SUMMARY 10
TABLE DES MATIERES 14
LISTE DES TABLEAUX 17
LISTE DES FIGURES 18
INTRODUCTION 20
PARTIE 1 : CADRE DE L'ETUDE 21
1. Eléments de contexte 21
1.1. L'entreprise 21
1.2. L'assurance vie et le contexte économique 22
1.2.1. Les différents types de contrats 22
1.2.2. La fiscalité 24
1.2.3. Les options et garanties 25
1.2.4. Contexte de l'étude 26
1.3. Spécificité du portefeuille LMP 26
2. Objectif 26
3. Données 28
3.1. Création de la base 28
3.2. Variables requises pour la modélisation 29
3.2.1. Variables endogènes 30
3.2.2. Variables exogènes 36
3.3. Nettoyage et retraitements des données 42
3.3.1. Au niveau arbitrage 42
3.3.2. Au niveau contrat 43
3.4. Base finale 44
3.5. Analyse des corrélations entre les variables explicatives 44
PARTIE 2 : MODELISATION 47
1. Mise en place de la modélisation 47

1.1. Méthodes de modélisation 47
1.1.1. Modélisation en une étape 47
1.1.2. Modélisation par fréquence sévérité 47
1.1.3. Modélisation par fréquence sévérité probabiliste 48
1.2. Sur apprentissage 49
1.3. Base d'entrainement et base de test 49
1.4. Détermination des hyperparamètres par validation croisée 50
2. Scores de performances 51
2.1. Modèles de classification 51
2.2. Modèles de régression 52
3. Les différents modèles utilisés 54
3.1. Modèle linéaire généralisé 54
3.1.1. Généralités sur le modèle linéaire généralisé 54
3.1.2. Pénalisations 56
3.1.3. Discrétisation 57
3.1.4. Régression Logistique 57
3.1.5. Régression Linéaire 59
3.1.6. Importances des variables 59
3.1.7. Limites 59
3.2. Modèle CART 60
3.2.1. Généralités sur le modèle CART 60
3.2.2. Fonctions d'hétérogénéité 61
3.2.3. Prédictions 62
3.2.4. Construction du modèle CART 62
3.2.5. Importances des variables 63
3.2.6. Limites 64
3.3. Agrégation de modèles 64
3.3.1. Bagging 64
3.3.2. Forêt aléatoire 67
PARTIE 3 : RESULTATS 69
1. Implémentation des modèles 69
1.1. Modèles de classification 69
1.1.1. Modèle GLM de classification 69
1.1.2. Modèle CART de classification 71

1.1.3. Modèle Bagging de classification 73
1.1.4. Modèle Forêt aléatoires de classification 74
1.1.5. Conclusion et interprétations 76
1.2. Modèles de régression 77
1.2.1. Modèle GLM de régression 78
1.2.2. Modèle CART de régression 79
1.2.3. Modèle Bagging de régression 80
1.2.4. Modèle Forêt aléatoires de régression 81
1.2.5. Conclusion et interprétations 83
2. Agrégation des deux modèles et prédictions 84
2.1. Modélisation par fréquence sévérité 84
2.2. Modélisation par fréquence sévérité probabiliste 86
3. Interprétation et limites 87
PARTIE 4 : PROJECTION 89
1. Objectif 89
2. Modification du modèle 90
3. Résultat du modèle simplifié 91
4. Implémentation des variables 92
5. Projections 94
5.1. Résultat 94
5.2. Calcul de sensibilité 95
6. Impacts sur la modélisation prospective 96
CONCLUSION 98
BIBLIOGRAPHIE 100
ANNEXES 101

LISTE DES TABLEAUX

Tableau 1 : Performances des modèles de classification sur les données test.....7
Tableau 2: Performances des modèles de régression sur les données de test.....8
Tableau 3 : Impacts sur la modélisation prospective.....9
Tableau 4 : Classification model performances on test data.....11
Tableau 5 : Regression model performances on test data.....11
Tableau 6 : Impacts on prospective modelisation.....12
Tableau 7 : Illustration de la base de modélisation (maille contrat x année).....29
Tableau 8 : Variables de la base de modélisation.....30
Tableau 9 : Extrait de la base de modélisation.....44
Tableau 10 : GLM fonction de lien canonique.....54
Tableau 11 : valeurs des paramètres pour mettre les lois usuelles sous la forme canonique exponentielle 55
Tableau 12 : Information sur les données test (1).....69
Tableau 13 : Importances des variables dans le GLM fréquence (simple).....70
Tableau 14 : Résultats des GLM fréquence.....71
Tableau 15 : Importance des variables dans les modèles de classification CART.....72
Tableau 16 : Résultats des modèles de classification CART.....72
Tableau 17 : Importance des variables dans les modèles de classification Bagging.....74
Tableau 18 : : Résultats des modèles de classification Bagging.....74
Tableau 19 : Importance des variables dans les modèles de classification Random Forest.....75
Tableau 20 : Résultats des modèles de classification Random Forest.....76
Tableau 21 : Synthèse des performances obtenues par les modèles de classification.....77
Tableau 22 : Information sur les données test (2).....77
Tableau 23 : Performances des modèles triviaux de régression.....78
Tableau 24 : Importances des variables GLM taux d'arbitrage.....78
Tableau 25 : Résultats du modèle GLM taux d'arbitrage.....79
Tableau 26 : Importance des variables modèle CART taux d'arbitrage.....80
Tableau 27 : Résultats du modèle CART taux d'arbitrage.....80
Tableau 28 : Importance des variables modèle Bagging taux d'arbitrage.....81
Tableau 29 : Résultats du modèle Bagging taux d'arbitrage.....81
Tableau 30 : Importance des variables modèle Random Forest taux d'arbitrage.....82
Tableau 31 : Résultats du modèle Random Forest taux d'arbitrage.....83
Tableau 32 : Illustration effet de la convergence.....87
Tableau 33 : illustration problème intervalle de confiance.....88
Tableau 34 : Impacts sur la modélisation prospective.....97

LISTE DES FIGURES

Figure 1 : Prédiction du montant net annuel arbitré par l'agrégation des modèles 8
Figure 2: Annual net amount arbitrated forecast by models aggregation 12
Figure 3: Organisation du groupe 21
Figure 4 : Cotisations en UC des 15 plus gros assureurs vie (source : L'Argus de l'assurance) 22
Figure 5 : Performance moyenne des fonds euros (source : ACPR) 23
Figure 6 : Performance des supports UC (sources : Euronext, France Assureurs) 24
Figure 7 : Valeur du PFL selon l'ancienneté du contrat (Source : impots.gouv.fr) 25
Figure 8 : Illustration de la méthodologie de calcul de la variable cible 27
Figure 9 : Montant annuel arbitré (net vers UC) 27
Figure 10 : Schéma simplifié de la construction de la base 28
Figure 11 : Moyenne des taux arbitrés en fonction de l'âge de l'assuré 31
Figure 12 : Proportion d'arbitrage en fonction de l'âge de l'assuré 31
Figure 13 : Moyenne des taux arbitrés en fonction de l'ancienneté du contrat 32
Figure 14 : Proportion d'arbitrage en fonction de l'ancienneté du contrat 33
Figure 15 : Proportion d'arbitrage en fonction du nombre d'arbitrage l'année précédente 34
Figure 16 : Moyenne des taux arbitrés en fonction du comportement d'arbitrage l'année précédente 34
Figure 17 : Moyenne des taux arbitrés en fonction de la proportion d'UC dans le contrat 35
Figure 18 : Proportion d'arbitrage en fonction de la proportion d'UC dans le contrat 36
Figure 19 : Montant mensuel arbitré (net vers UC) 36
Figure 20 : Valeurs mensuelles du montant net arbitré et du CAC40 37
Figure 21: Valeurs mensuelles du montant net arbitré et de la volatilité du CAC40 38
Figure 22 : Valeurs mensuelles du montant net arbitré et du taux de l'OAT France 10 ans 38
Figure 23 : Valeurs mensuelles du montant net arbitré et de l'ICM 39
Figure 24 : Moyenne des taux arbitrés en fonction de l'écart avec le CAC40 40
Figure 25 : Proportion d'arbitrage en fonction de l'âge de l'écart avec le CAC40 40
Figure 26 : Moyenne des taux arbitrés en fonction de l'écart avec l'OAT France 10 ans 41
Figure 27 : Proportion d'arbitrage en fonction de l'âge de l'écart avec l'OAT France 10 ans 41
Figure 28 : Proportion d'arbitrage automatique 42
Figure 29 : Montant annuel arbitré (net vers UC) avec et sans les arbitrages automatiques 42
Figure 30 : Impact des arbitrages automatiques sur les variables cibles 43
Figure 31 : Différence corrélation Pearson / Spearman 45
Figure 32 : Matrice de corrélation (calculée avec le coefficient de Spearman) 46
Figure 33 : Illustration du réechantillonnage 48
Figure 34 : Illustration du sur/sous apprentissage 49
Figure 35 : Illustration de la validation croisée 50
Figure 36 : Matrice de confusion 51
Figure 37 : Aire sous la courbe précision/rappel 52
Figure 38 : Norme L1 et L2 56
Figure 39 : Arbre de décision 60
Figure 40 : Calcul de la probabilité estimée avec un arbre de décision 62
Figure 41 : Illustration élagage d'un arbre de décision 63

Figure 42 : Illustration des différentes étapes dans un modèle de Bagging 65
Figure 43 : Illustration des différentes étapes dans un modèle de forêt aléatoire 68
Figure 44 : Prédiction du montant net annuel arbitré par la modélisation Fréquence/Sévérité 85
Figure 45 : Prédiction du montant net annuel arbitré par la modélisation fréquence/sévérité probabiliste 87
Figure 46 : Illustration du problème pour implémenter le modèle sur plusieurs années (1) 90
Figure 47 : Illustration du problème pour implémenter le modèle sur plusieurs années (2) 90
Figure 48 : Illustration du problème pour implémenter le modèle sur plusieurs années (3) 91
Figure 49 : Prédiction du modèle simplifié 92
Figure 50 : Prédiction du taux d'arbitrage (net vers UC) sur 60 ans 94
Figure 51 : Prédiction du montant arbitré (net vers UC) sur 60 ans 95
Figure 52 : Analyse de la sensibilité aux performances des fonds 96
Figure 53 : Formulaire à remplir pour effectuer un arbitrage (1) 101
Figure 54 : Formulaire à remplir pour effectuer un arbitrage (2) 102
Figure 55 : Variables cible en fonction de la plus ou moins-value sur l'UC en n-1 103
Figure 56 : Variables cible en fonction de la plus ou moins-value en n-1 103
Figure 57 : Allure du modèle fréquence CART 104
Figure 58 : Allure du modèle fréquence rééchantillonné CART 104
Figure 59 : Résultat des prédictions avec la modélisation fréquence/sévérité rééchantillonné 70/30 (sortie python) 105
Figure 60 : Résultat des prédictions avec la modélisation fréquence/sévérité rééchantillonné 75/25 (sortie python) 105

INTRODUCTION

L'assurance vie est le deuxième placement préféré des Français derrière le livret A et le placement le plus important en termes d'encours. Les raisons de son succès sont multiples ; fiscalité avantageuse, transmission du capital, rémunération attractive, diversité des placements, …

Les contrats d'assurance vie multisupports offrent la possibilité aux assurés d'arbitrer entre le fond Euro et les fonds UC, dont les performances varient et dépendent du contexte économique. Les arbitrages ont un impact sur la rentabilité et la solvabilité de l'assureur. Dès lors, la question se pose de savoir ce qui motive ces mouvements ? Quels sont les facteurs déterminants dans la décision d'arbitrer ? Comment les anticiper ? A quel point ces derniers impactent-ils la solvabilité et la rentabilité de l'assureur ?

C'est pour répondre à ces questions que nous allons tenter de modéliser ces flux. Nous opterons pour une modélisation à la maille contrat avec un pas de temps annuel. Deux approches seront détaillées : l'approche par fréquence/sévérité dite « classique » et l'approche par fréquence/sévérité dite « probabiliste ». Ces approches sont souvent utilisées en assurance non-vie pour des modèles individuels.

Pour cela, nous devrons d'abord créer la base qui nous permettra une telle modélisation. Une fois cette base créée, il faudra s'interroger sur les potentielles raisons qui poussent les individus à arbitrer afin d'intégrer des variables explicatives pertinentes. Ces raisons peuvent être propres à l'assuré (l'âge, l'ancienneté du contrat, la provision mathématiques, …) ou bien propres au contexte économique (valeur du CAC40, valeur de l'OAT France 10 ans, l'indice de confiance des ménages, …). Plusieurs modèles seront ensuite implémentés, en commençant par des modèles linéaires jusqu'aux agrégations de modèles non linéaires d'apprentissage automatique. Pour éviter le sur-apprentissage, les hyperparamètres seront fixés en utilisant la validation croisée et les modèles seront testés sur les données de 2020 et 2021 qui n'ont pas servi à leur construction. L'évaluation de ces modèles se fera à l'aide de métriques pertinentes et adaptées à notre problématique. Elle nous permettra de retenir le meilleur modèle parmi tous ceux implémentés.

Le modèle retenu sera alors légèrement modifié pour permettre de le laisser évoluer sur 60 années. Les projections seront ensuite intégrées dans un outil de projection d'actif/passif pour permettre d'évaluer l'impact de notre modélisation des arbitrages sur la rentabilité et la solvabilité de l'assureur.

PARTIE 1 : CADRE DE L'ETUDE

1. Eléments de contexte

1.1. L'entreprise

AG2R La Mondiale est un organisme français de protection sociale et patrimoniale dont la gouvernance repose sur le paritarisme et le mutualisme. Il fut fondé en 1905 par 7 industriels du Nord mais a beaucoup évolué au fil des années du fait des nombreux partenariats.

En 1951, est créée l'Association Générale de Retraite par Répartition (AGRR) qui a finalement été renommée AG2R en 1992. Puis, en 2005, AG2R se rapproche de Prémalliance pour former un partenariat. En 2008 est créée la SGAM (Société de Groupe d'Assurance Mutuelle) AG2R La Mondiale. En 2014, VIASANTE Mutuelle rejoint AG2R La Mondiale puis en 2015 c'est au tour de Réunica de rejoindre le groupe. En janvier 2019, le groupe AG2R La Mondiale et la Matmut présentent la nouvelle entité née de leur collaboration, AG2R La Mondiale Matmut. Le nouveau groupe a pour ambition de devenir « le premier assureur de l'économie sociale à développer une approche globale en direction des professionnels et entreprises » et « l'assureur de référence des seniors ». Cependant, courant 2019, le groupe nouvellement créé se sépare à la suite du retrait dans le processus de rapprochement de La Mondiale.

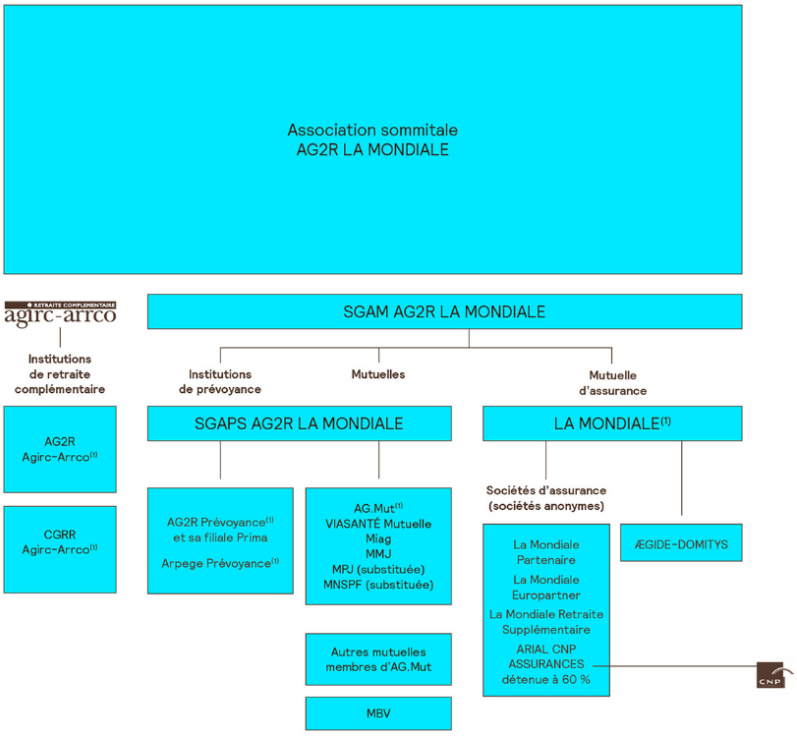


Figure 3: Organisation du groupe

AG2R La Mondiale assure 15 millions de clients particuliers et 500 000 entreprises. En 2021, d'après le classement réalisé par l'Argus de l'Assurance, le groupe occupe la treizième place au classement des plus gros assureurs vie avec un encours de près de 40 milliards d'euros et 3.1 milliards d'euros de cotisation en affaire direct, dont 48% réalisés en UC (contre environ 40% en moyenne pour le top 15).

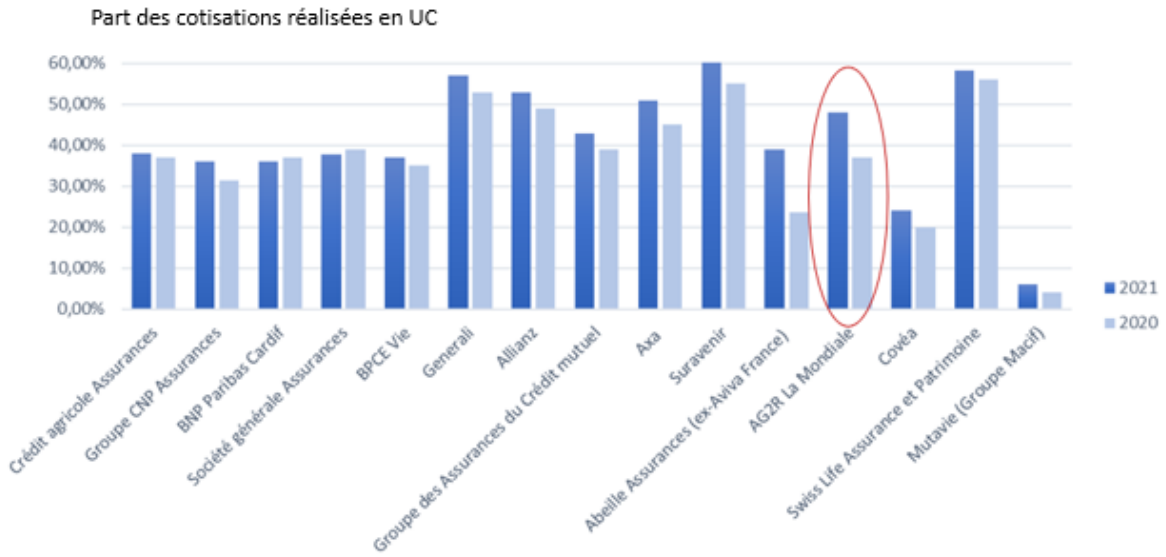


Figure 4 : Cotisations en UC des 15 plus gros assureurs vie (source : L'Argus de l'assurance)

1.2. L'assurance vie et le contexte économique

Un contrat d'assurance vie est un outil d'épargne qui permet à un souscripteur de construire et valoriser son capital. Il se distingue des autres placements financiers par deux aspects : une fiscalité avantageuse et la possibilité de transmettre ou non le patrimoine du souscripteur à un ou plusieurs bénéficiaires choisis lors de la signature du contrat. De plus, ce placement offre historiquement une rémunération attractive du capital.

C'est pour ces trois raisons notamment, qu'en 2021, près de 18 millions de français en détiennent au moins une, soit plus de 40% des ménages, positionnant l'assurance vie comme le deuxième placement préféré des Français après le Livret A.

Néanmoins, avec un encours s'élevant à 1876 milliards d'euros fin 2021 (+4.4% par rapport à 2020) contre moins de 350 milliards d'euros pour le Livret A, l'assurance vie est de loin le placement le plus important en France.

Dans cette partie, nous nous attarderons sur l'assurance vie et ses spécificités en évoquant les différents types de contrats, la fiscalité et les différentes options et garanties.

1.2.1. Les différents types de contrats

Lors de la souscription à un contrat d'assurance vie, le souscripteur a la possibilité de choisir entre différents types de contrat. Il peut choisir de placer son épargne sur :

- Le fond EURO, qui est le support financier le plus sécurisé car c'est l'assureur qui prend le risque. En effet, le capital investi est généralement placé sur des placements financiers peu risqués comme des obligations (BFT, OAT etc.). L'assureur lui garantit la totalité des fonds investis et s'engage à les revaloriser tous les ans au Taux Minimum Garanti (TMG) qui est inscrit dans le contrat. Ce taux garanti est plafonné et réglementé par le Code des assurances. A ce TMG s'ajoute également la Participation aux Bénéfices (PB) qui s'élève au minimum à

85% en ce qui concerne le résultat financier et 90% en ce qui concerne le résultat technique. On parle alors de contrat monosupport EURO.

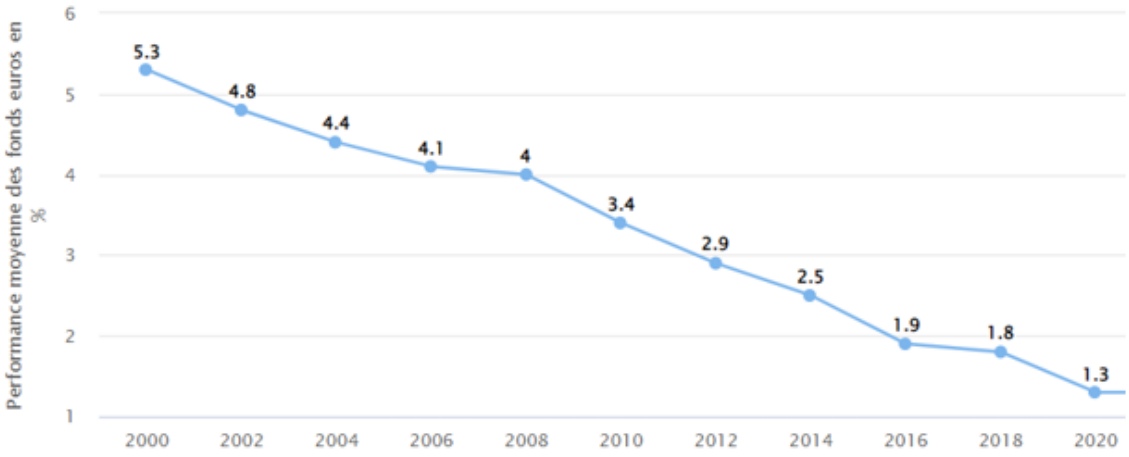


Figure 5 : Performance moyenne des fonds euros (source : ACPR)

Cependant, bien que les fonds Euros soient les placements les plus sécurisés pour les souscripteurs, on remarque que leur rendement diminue au fil des années pour atteindre 1.3% en 2020. Dû à la baisse historique des taux d'intérêts qui a eu lieu entre les années 80 jusqu'à fin 2021, les performances des fonds euros n'ont fait que de baisser, incitant certains assurés à se tourner vers un autre type de support, le fond UC.

- Le fond UC, est un support financier plus risqué que le fond Euro car c'est l'assuré qui porte le risque de marché. En effet, lors de la signature du contrat, l'assureur attribue au souscripteur un nombre de parts d'UC en échange de son capital investi. Il peut être investi sur différents placements financiers tels que des actifs immobiliers, des actions ou encore des obligations. L'assureur garanti donc au souscripteur le nombre de parts d'UC mais pas la valeur de ces dernières qui, elle, fluctue en fonction du marché. En contrepartie de cette prise de risque, l'assuré peut espérer un rendement plus important qu'avec le support Euro. On parle alors de contrat monosupport UC. Ces contrats sont très rares car peu de gens optent pour cette option.

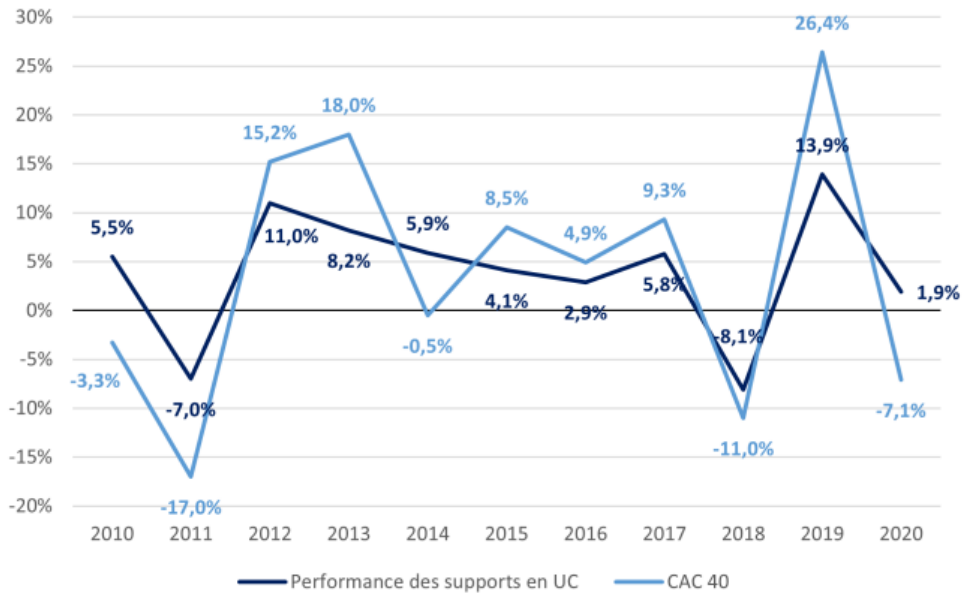


Figure 6 : Performance des supports UC (sources : Euronext, France Assureurs)

Comme on peut le constater sur la figure ci-dessous, les performances des supports UC sont historiquement fortement corrélés avec les performances des principaux indices boursier, notamment celui du CAC40.

- Les deux fonds en même temps. Dans les proportions qu'il souhaite, l'assuré peut placer une partie de son épargne sur le fond EURO et l'autre partie sur le fond UC. A travers l'option d'arbitrage décrite plus bas, l'assuré se garde la possibilité de changer son épargne de fond. On parle alors de contrat multisupport. En 2021, la quasi-totalité des contrats souscrit de notre portefeuille sont des contrats multisupports, le monosupport EURO étant fermé à la commercialisation.

1.2.2. La fiscalité

Les intérêts perçus par les contrats d'assurance vie sont soumis, comme tout investissement financier, aux cotisations sociales et à l'imposition. Cependant, la façon de calculer les cotisations sociales diffère entre les différents supports d'investissement. En effet, pour le support Euro, les intérêts produits sont abattus de 17,2% chaque année avant d'être comptabilisés dans la provision mathématiques (PM) du contrat. Pour le support en UC, c'est seulement à l'issu du contrat que la PM est déduite de 17,2% de la somme des intérêts produits chaque année.

A la clôture du contrat ou lorsque l'assuré souhaite récupérer tout ou partie de son capital revalorisé, deux possibilités s'offrent à lui quant à l'imposition sur les intérêts perçus : soit il peut décider de les intégrer à sa déclaration de revenu, soit il peut décider de les déduire d'un Prélèvement Forfaitaire Libératoire (PFL) qui se présente comme suit :

Figure 7 : Valeur du PFL selon l'ancienneté du contrat (Source : impots.gouv.fr)

Etant donné que la fiscalité devient plus avantageuse sur les plus-values lorsque l'ancienneté du contrat augmente, on observe des pics de contrats rachetés après la 4ème et la 8ème année de la durée de vie du contrat.

1.2.3. Les options et garanties

Le souscripteur d'une assurance vie possède plusieurs options sur son contrat. Les principales sont les suivantes :

- Les rachats : Le souscripteur peut, à tout moment, récupérer tout ou une partie de son épargne. On appelle ça un rachat. Il est dit « total » si la totalité de l'épargne est retirée, « partiel » sinon. Cette liberté dans le choix de rachat lui assure l'avantage d'une liquidité de son épargne. Cependant, cela peut en contrepartie avoir un coût pour l'assureur.
  En effet, quand du capital investi sur le fond Euro est racheté, l'assureur doit revendre les actifs sur lesquels est investi ce capital pour pouvoir verser le rachat demandé. L'assureur s'expose donc à des ventes en situation de moins-value. Or comme le capital est garanti sur le fond Euro, c'est l'assureur qui doit payer la différence.

- Les versements : Il en existe trois types :
  - Les versements initiaux : ce sont les montants versés à l'ouverture du contrat
  - Les versements programmés : ce sont les montants versés à une périodicité définie.
  - Les versements libres : ce sont les montants versés selon la volonté de l'assuré peu importe la période.

  Il est important de noter que, pour certains contrats, un montant minimal de versements est exigé, sans distinction de leur type. Aussi, le fait de réaliser des versements programmés n'empêche pas de réaliser des versements libres.

- Les arbitrages : Ils sont utilisés pour un contrat multisupport. Ils permettent à l'assuré de modifier la répartition de son épargne entre les supports Euro et les supports UC.

1.2.4. Contexte de l'étude

Notre périmètre d'étude portera sur tous les contrats multi-supports du portefeuille LMP. En effet, il ne fait pas sens de prendre en compte dans notre étude les contrats mono-support étant donné que ces derniers ne peuvent pas effectuer d'arbitrage.

De plus nous nous restreignons aux contrats souscrits entre le 01 janvier 2001 et le 31 décembre 2021. Les données n'étant pas ou peu disponible avant cette date.

Ce périmètre d'étude représente 209 088 contrats dont 122 353 encore ouverts au 31 décembre 2021, pour une Provision Mathématiques (PM) totale de 28 057 millions € (15 177 millions € sur l'EURO et 12 880 millions sur l'UC). Par rapport au portefeuille total de LMP, cela représente 74.22% des contrats et 84.08% de la PM.

1.3. Spécificité du portefeuille LMP

Le portefeuille LMP est un portefeuille d'épargne patrimonial. Il diffère d'un portefeuille d'épargne classique du fait de la clientèle ciblée. Avec un versement initial minimum de 30 000€ et une Provision Mathématique (PM) moyenne de 200 000€, les produits s'adressent uniquement à une clientèle fortunée.

Ceci présente plusieurs impacts potentiels sur notre étude.

- Premièrement, on s'attend à ce que les clients aient une certaine connaissance financière. Raison pour laquelle l'offre d'UC est en général plus large que sur un portefeuille classique. On peut également légitimement penser que les comportements d'arbitrages lors d'évènements majeurs (crises financières par exemple) peuvent différer d'un portefeuille classique.
- Deuxièmement, comme le ratio Provision Mathématiques sur Nombre de clients est bien plus élevé que sur un portefeuille classique, le poids de chaque client est plus important. Raison pour laquelle les clients doivent, pour chaque procédure (rachats, arbitrages, …), prendre rendez-vous avec leur conseiller pour remplir un formulaire de demande sous format papier afin qu'il soit envoyé à un « approbateur » pour y être soumis à validation. On peut donc penser que cette procédure lourde et longue va également modifier les comportements d'arbitrages lors d'évènements majeures en limitant les mouvements de panique par exemple.

2. Objectif

L'objectif principal de ce mémoire est de pouvoir modéliser le montant d'arbitrage net vers l'UC dans une année. On définit ce montant comme étant la différence entre la somme des arbitrages vers l'UC et la somme des arbitrages vers l'EURO.

$$Montant\ net\ arbitré_n = \sum Montant\ arbitré\ EU \rightarrow UC_n - \sum Montant\ arbitré\ UC \rightarrow EU_n$$

Ci-dessous un exemple illustratif pour un individu :

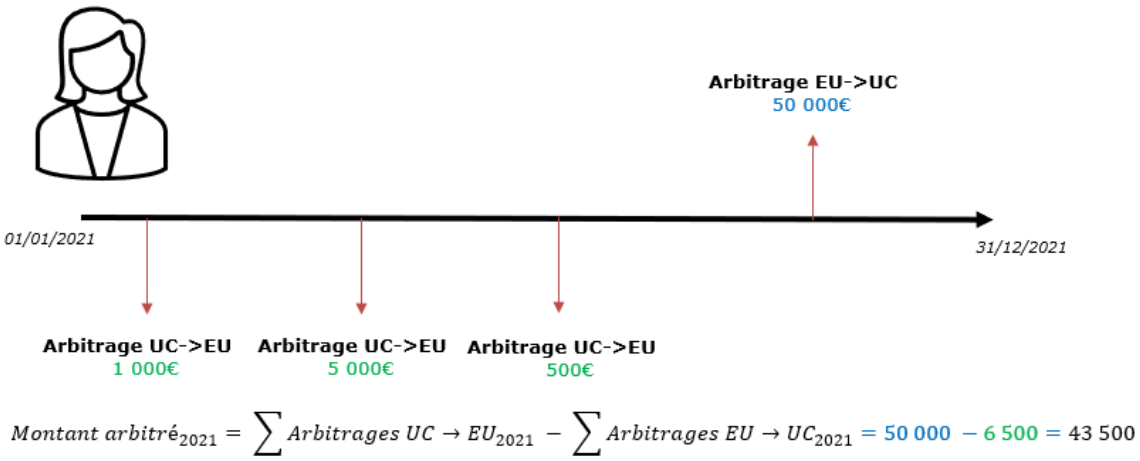


Figure 8 : Illustration de la méthodologie de calcul de la variable cible

On procède de cette manière pour tous les contrats de l'année en question (2021 dans notre exemple) et on somme le tout.

Un résultat négatif signifie donc qu'il y a eu plus d'arbitrages sortants de l'UC que d'arbitrages entrants. Pour avoir un ordre de grandeur, voici les montants nets arbitrés de 2015 à 2021 sur notre périmètre d'étude.

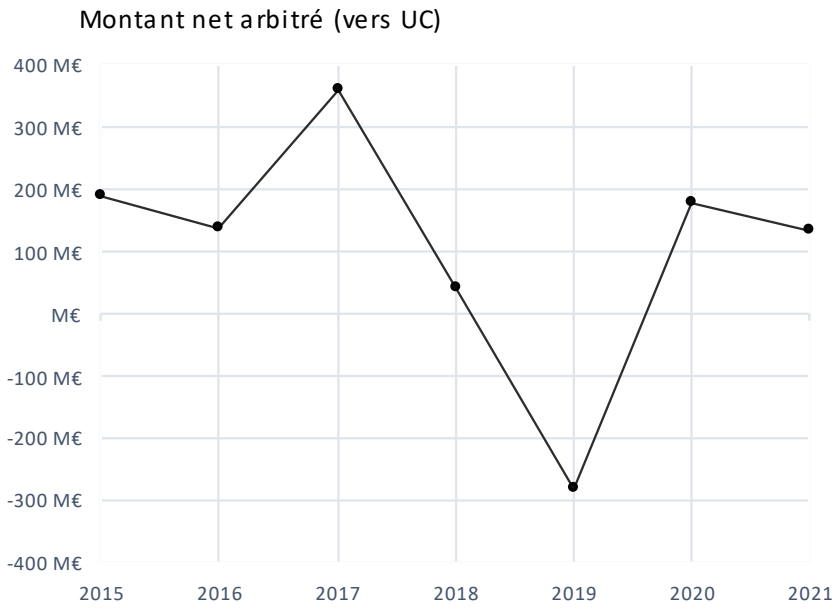


Figure 9 : Montant annuel arbitré (net vers UC)

Les intérêts de pouvoir prédire ce montant sont multiples.

Premièrement, en cas de réorientation conséquente de l'épargne des assurés vers un fond, l'assureur se voit dans l'obligation de céder précipitamment des actifs financiers pour pouvoir réinvestir sur le(s) support(s) choisi(s) par l'assuré. Ceci peut conduire l'assureur à acheter et/ou céder ses actifs au mauvais

moment et causer des pertes conséquentes. Anticiper ce mouvement permettra à l'assureur d'avoir un temps de recul plus important pour acheter et/ou céder ses actifs au moment le plus opportun.

Deuxièmement, les marges techniques et financières, ainsi que les charges (réassurance notamment) ne sont pas les mêmes suivant les fonds. Les arbitrages ont donc un impact[1] sur la valeur du portefeuille et les modélisations prospectives.

Pour finir, les sommes placées par les assurés sur le fond euro sont garanties par l'assureur. C'est donc l'assureur qui porte le risque, et ceci fait augmenter son besoin en fonds propres.

En somme, la modélisation des arbitrages permet à l'assureur d'améliorer la connaissance du risque auquel il est exposé. Malgré tout, bien que reconnus par les compagnies d'assurance, les flux d'arbitrages ne restent souvent pas ou peu modélisés par ces dernières. La raison à cela provient de la grande complexité à modéliser ces mouvements, se trouvant être à la fois d'origines endogènes et exogènes.

3. Données

3.1. Création de la base

Pour effectuer notre modélisation, nous avons voulu créer un jeu de données à la maille contrat X année. Autrement dit, avoir une ligne par contrat et par année. L'intérêt est de construire les modèles sur les données jusqu'à 2019 inclus, puis de les tester sur 2020 et 2021.

Cette base n'existant pas, nous avons dû la construire. Ci-dessous se trouve un schéma de la construction de la base.

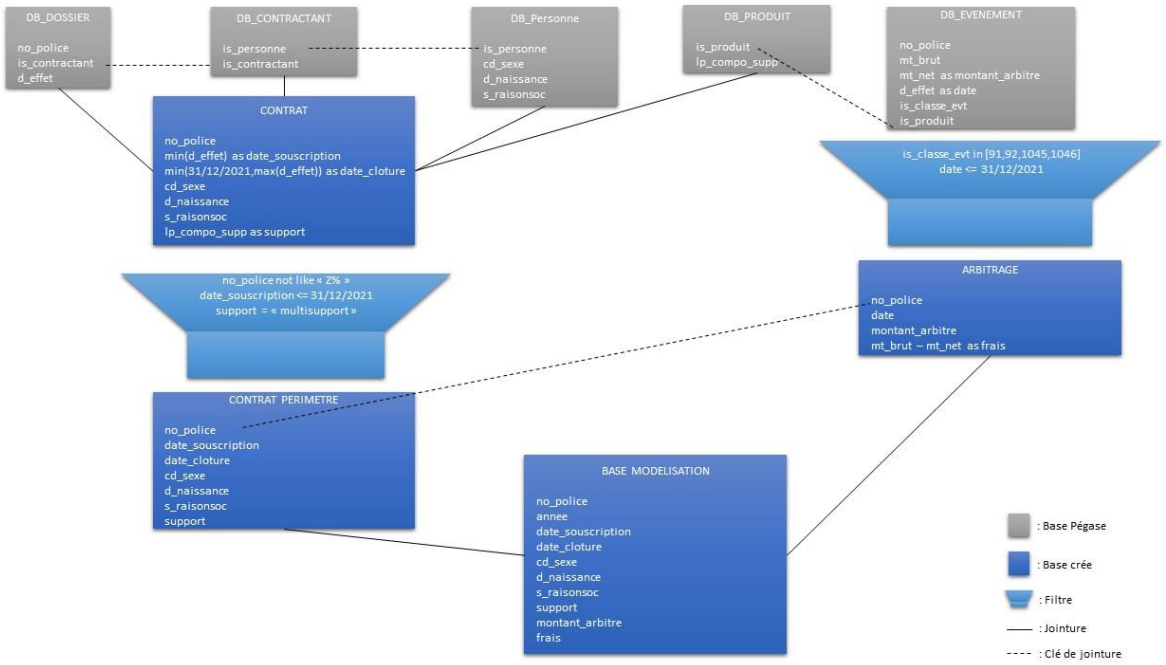


Figure 10 : Schéma simplifié de la construction de la base

Voici une illustration du résultat obtenu

[1] La réalisation de ce mémoire fait suite à une demande du commissaire aux comptes souhaitant connaitre et quantifier cet impact.

Tableau 7 : Illustration de la base de modélisation (maille contrat x année)

No_police	Annee	Genre	Age	Ancienneté	Montant	Frais
0001	2019	F	48	2	- €	- €
0001	2020	F	49	3	12 000,00 €	- €
0001	2021	F	50	4	11 000,00 €	- €
0002	2017	M	34	2	- 40 000,00 €	400,00 €
0002	2018	M	35	3	- €	- €
0002	2019	M	36	4	- €	- €

D'autres variables explicatives seront ensuite rajoutées directement sur ce jeu de données.

3.2. Variables requises pour la modélisation

Comme expliqué précédemment, on cherche à modéliser le montant net arbitré dans une année. On peut également définir le taux d'arbitrage net de la façon suivante :

$$Taux\ arbitrage\ net = \frac{Montant\ net\ arbitré}{PM}$$

La PM étant déterministe, modéliser l'un revient à modéliser l'autre. On privilégiera la modélisation du taux car c'est une valeur comprise entre -1 et 1, ce qui est plus pratique à manipuler.

Modéliser les flux arbitrés commence déjà par se questionner sur les variables qui peuvent potentiellement, de près ou de loin, impacter les comportements d'arbitrages.

Les recherches qui ont porté sur la modélisation des arbitrages ont montré que ces derniers sont impactés par deux types de grandeurs. Les grandeurs endogènes bien sûr, propres aux assurés (l'âge de l'assuré, l'ancienneté du contrat, …). Mais aussi les grandeurs exogènes, propres à la période (indicateurs économiques, boursiers, …).

Au total, près d'une vingtaine de variables ont été utilisées pour la création de ce modèle.

Tableau 8 : Variables de la base de modélisation

Variables	Signification
No_police	Identifiant du contrat
Annee	Année observée
Type_personne	Homme/Femme/Personne morale
Age	Âge de l'assuré
Anciennete	Ancienneté du contrat
Nb_supports	Nombre de supports UC accessibles à l'assuré
Versement_ini	Versement initial
Pm	Provision Mathématiques au 31/12/n-1
Proportion_UC	Proportion du contrat placé sur l'UC au 31/12/n-1
Taux_pb	Taux de pb versé en n-1
Pmv_prop	Taux de plus ou moins value en n-1
Pmv_eu_prop	Taux de plus ou moins value sur l'EURO en n-1
Pmv_uc_prop	Taux de plus ou moins value sur l'UC en n-1
Part_uc_pmv	Proportion de la plus value n-1 réalisé par l'UC
Ecart_pmv_uc_eu	Différence entre le taux de plus ou moins value sur l'UC en n-1 et celle sur l'EURO
Moyenne_cac	Valeur moyenne du CAC 40 en n-1
Std_cac	Volatilité du CAC 40 en n-1
Tx_rdmt_cac	Taux de rendement global du CAC 40 sur l'année n-1
Oat_fra_10a	Taux de l'OAT France sur 10 ans
Ecart_cac	Différence entre le taux de plus ou moins value du contrat et le taux de rendement global du CAC 40 en n-1
Ecart_oat	Différence entre le taux de plus ou moins value du contrat et le taux de l'OAT France 10 ans en n-1
Arbitrage_dum	Prend la valeur 1 si l'assuré a arbitré sur l'année en question, 0 sinon
Taux_arbitrage	La proportion du contrat arbitré
Montant_arbitre	Le montant arbitré

Ci-dessous sont détaillés certaines d'entre elles, notamment celles qui ceux sont révélés comme étant les plus importantes.

3.2.1. Variables endogènes

- L'âge de l'assuré

Souvent discriminant dans des modèles individuels en assurance, il parait naturel de penser que plus il sera avancé, plus les arbitrages se feront rares et d'avantage vers l'Euro (dans une optique de sécurisation de l'épargne).

Néanmoins on remarque graphiquement que cette conjecture s'avère peu vérifiée sur nos données, si ce n'est pour les âges très avancés. En effet, si on regarde par exemple la moyenne des taux arbitrés [2] en fonction des différentes classes d'âges, on s'aperçoit que seuls les plus de 85 ans s'écartent fortement de la moyenne globale.

[2] Moyenne uniquement calculée sur les lignes où il y a eu arbitrage.

Figure 11 : Moyenne des taux arbitrés en fonction de l'âge de l'assuré

De plus si on regarde cette fois, pour chaque âge, la proportion des contrats qui n'ont pas effectué d'arbitrage, on constate que cette proportion varie peu. Parmi tous les assurés de 30 ans, 94% n'ont pas fait de mouvement d'arbitrage. Cette proportion descend à 92% pour les assurés de 50 ans et monte à 98% pour ceux de 90 ans.

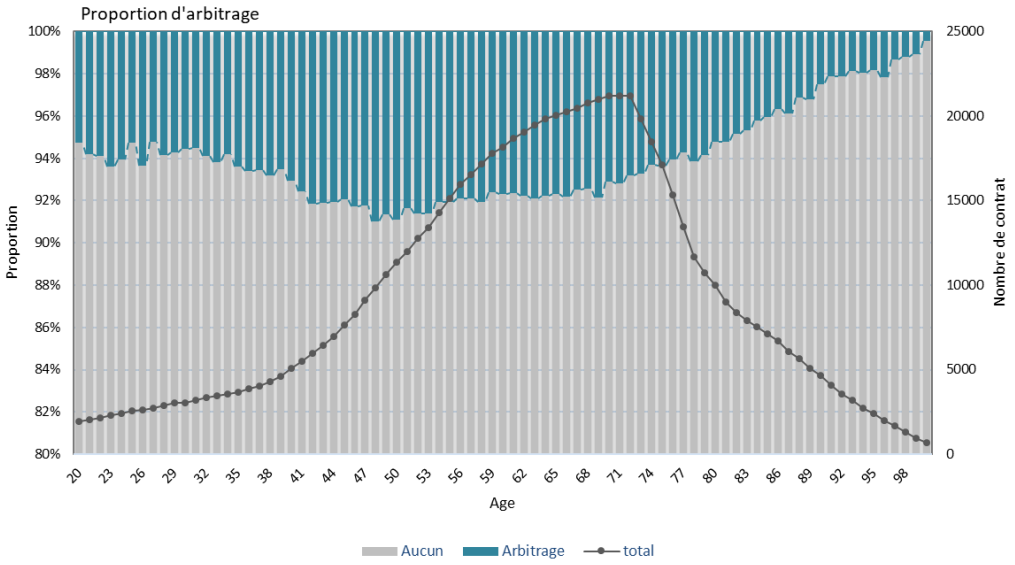


Figure 12 : Proportion d'arbitrage en fonction de l'âge de l'assuré

- L'ancienneté du contrat

Très discriminante dans les flux de rachats du fait des avantages fiscaux liés à l'ancienneté, il n'y a à priori pas de raison pour qu'elle joue un rôle significatif dans les flux d'arbitrage.

Néanmoins lorsque l'on regarde sur nos données nous observons une grosse différence dans le taux moyen arbitré en fonction de l'ancienneté. Il apparaît assez nettement que plus le contrat est récent plus les arbitrages vers l'UC sont conséquents.

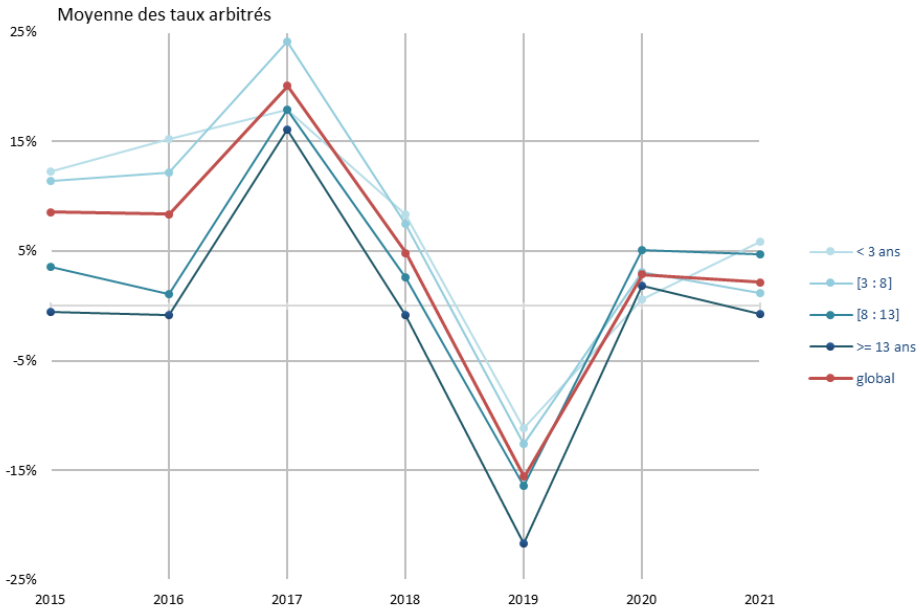


Figure 13 : Moyenne des taux arbitrés en fonction de l'ancienneté du contrat

De plus on remarque également une légère décroissance dans la proportion de contrats qui effectuent un arbitrage en fonction de l'ancienneté.

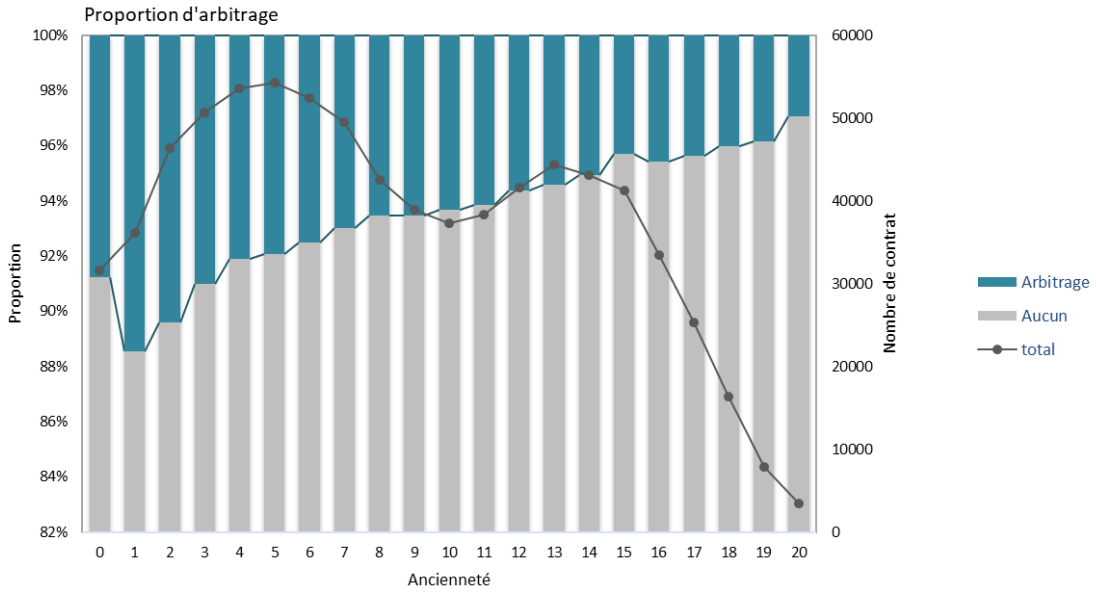


Figure 14 : Proportion d'arbitrage en fonction de l'ancienneté du contrat

- Le nombre d'arbitrage l'année d'avant

« *Qui a bu boira* » dit le célèbre proverbe. Il est donc légitime de penser qu'un assuré ayant effectué un arbitrage l'année précédente est plus à même d'en effectuer un l'année d'après.

C'est en effet vérifié dans nos données. Il apparaît même que cette variable est la plus discriminante de toutes en ce qui concerne la proportion de personne ayant arbitrée.

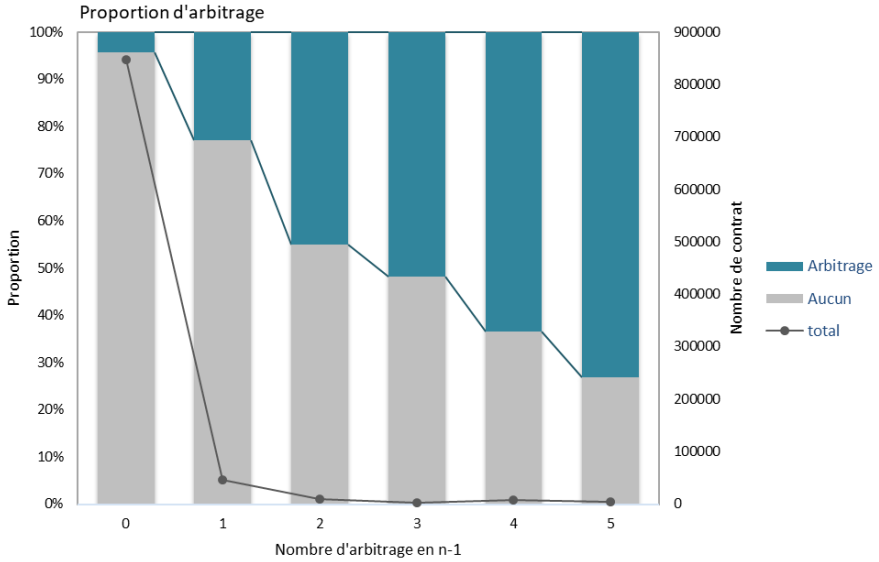


Figure 15 : Proportion d'arbitrage en fonction du nombre d'arbitrage l'année précédente

La tendance est très marquée. La proportion d'arbitrage chez les assurés qui n'ont pas arbitrés l'année d'avant est de 4 %. Cette proportion ne fait qu'augmenter plus le nombre d'arbitrage effectué l'année d'avant est élevé, jusqu'à atteindre plus de 70% chez les assurés ayant effectués 5 arbitrages l'année précédente.

Néanmoins, arbitrer beaucoup en fréquence ne signifie pas arbitrer beaucoup en montant. Et quand on regarde le taux moyen arbitré on s'aperçoit que c'est en réalité l'inverse. Celui des assurés ayant arbitrés en n-1 est significativement plus faible que les autres.

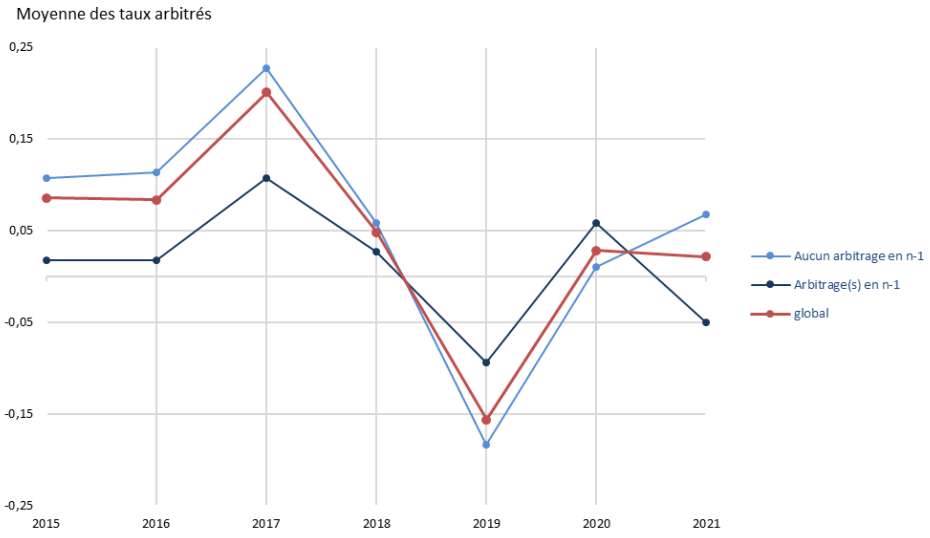


Figure 16 : Moyenne des taux arbitrés en fonction du comportement d'arbitrage l'année précédente

- Le nombre d'arbitrages cumulés

Basé sur le même principe, il ne s'agit pas seulement de regarder le nombre d'arbitrages effectué l'année d'avant, mais le nombre d'arbitrages depuis la création du contrat.

La même tendance est observée, légèrement moins marquée mais qui présente l'avantage d'avoir des classes plus équilibrées.

- La proportion d'UC dans le contrat au début de l'année

Comme expliqué précédemment, l'assuré peut répartir son épargne comme il le souhaite entre le fond EURO et les fonds UC. L'idée est de prendre cette répartition au 31/12 de l'année n-1. Il est à priori difficile d'émettre une conjecture de l'impact de cette variable. Pourtant, il apparaît clairement qu'elle est la plus discriminante de toutes pour ce qui concerne le taux arbitré.

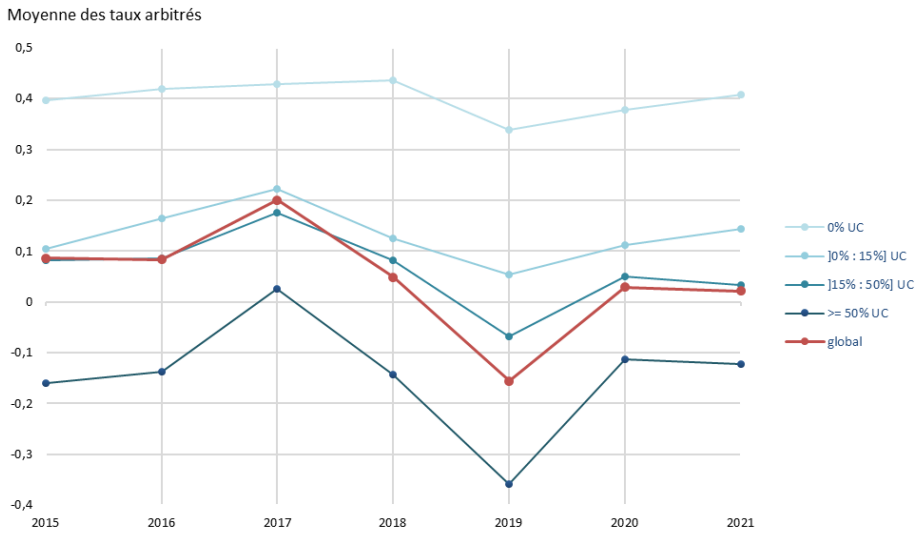


Figure 17 : Moyenne des taux arbitrés en fonction de la proportion d'UC dans le contrat

En réalité, cela se comprend aisément. Premièrement les contrats ayant 0% d'UC ne peuvent qu'arbitrer vers l'UC. Or, au 31/12/2021, cela représente un tiers de nos contrats.

La logique reste vraie dans le sens inverse, même si peu de contrat sont 100% UC, un assuré possédant 90% d'UC a peu de chance d'arbitrer vers de l'UC. Globalement, on remarque que, quand il y a arbitrage, c'est en direction du fond minoritaire.

Pour ce qui concerne la fréquence d'arbitrage, on remarque que ce sont les contrats les plus équilibrés entre EURO/UC qui arbitrent davantage.

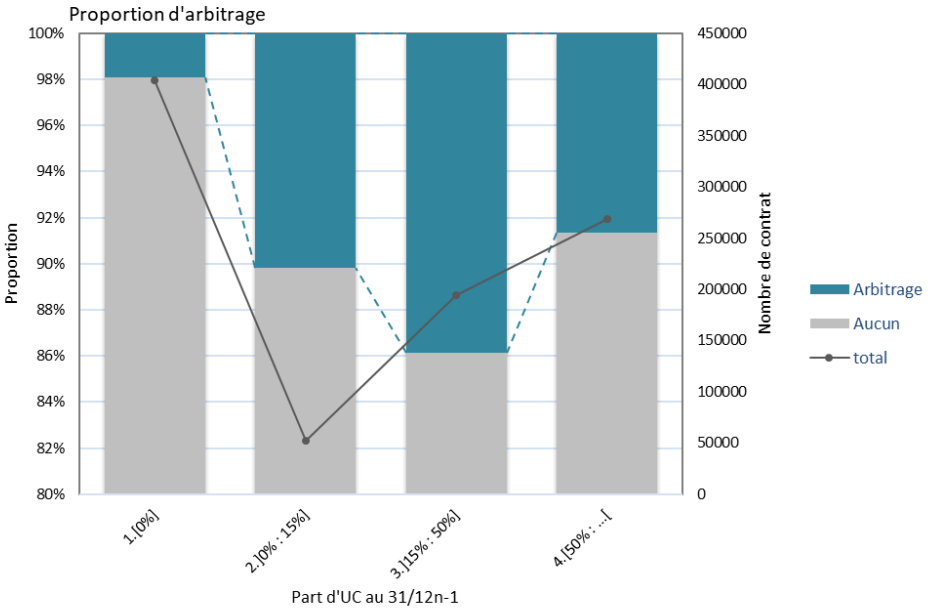


Figure 18 : Proportion d'arbitrage en fonction de la proportion d'UC dans le contrat

3.2.2. Variables exogènes

En observant le montant net arbitré (vers l'UC) mensuel, on observe à certaines dates de grosses fluctuations. Ceci nous conforte alors dans l'idée que le contexte économique et financier influence les arbitrages.

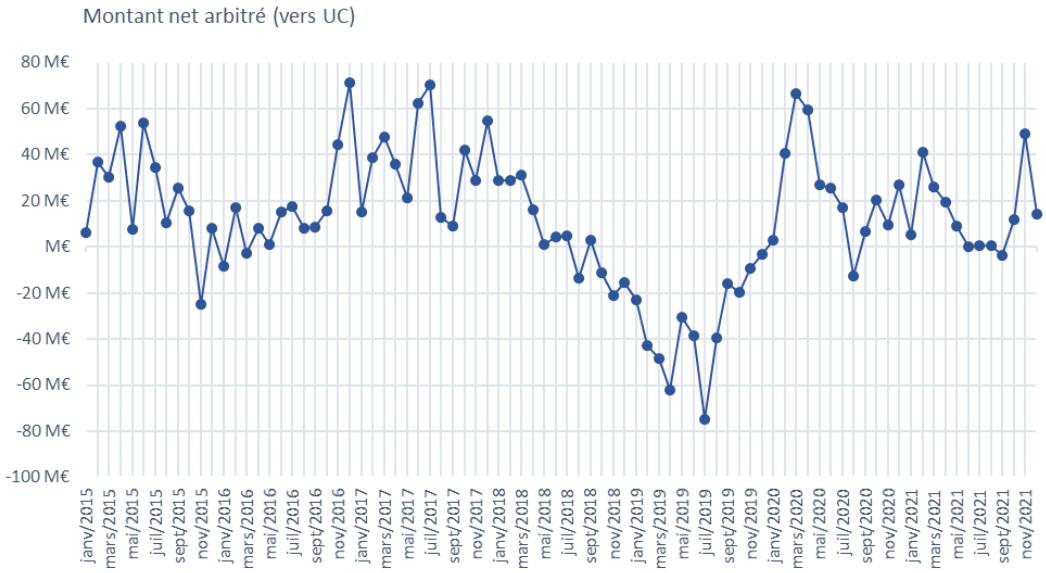


Figure 19 : Montant mensuel arbitré (net vers UC)

- Le CAC 40

Tous les experts s'accordent à dire que le CAC 40 est l'indice boursier le plus suivi par les épargnants français qui souhaitent avoir des indications sur l'évolution des marchés actions. De plus, nous savons que les performances du CAC 40 impactent directement les performances de nombreuses UC.

Néanmoins, graphiquement on ne distingue aucune corrélation particulière entre la valeur du CAC 40 et les flux d'arbitrages.

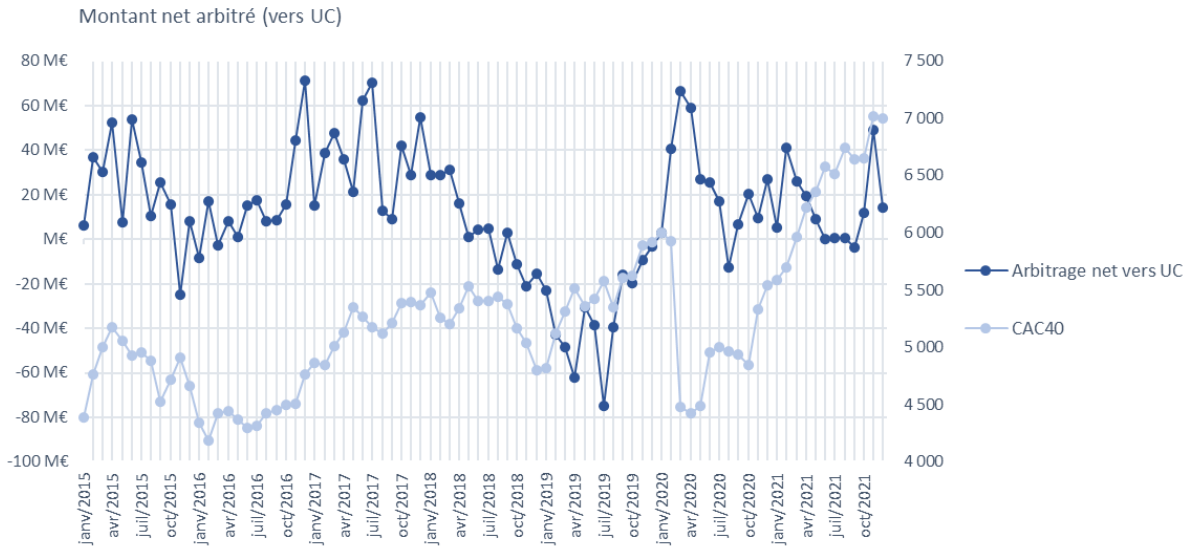


Figure 20 : Valeurs mensuelles du montant net arbitré et du CAC40

Cependant, au moment de la crise COVID, on observe simultanément une forte chute du CAC 40 et une forte hausse des arbitrages vers l'UC. Peut-être est-ce donc d'avantage la volatilité du CAC 40 qui impacterai les comportements d'arbitrages ?

- Volatilité du CAC 40

On décide donc cette fois-ci de regarder le montant net arbitré avec la volatilité du CAC 40 pour chaque mois. Une nouvelle fois, aucune corrélation ne semble apparaître.

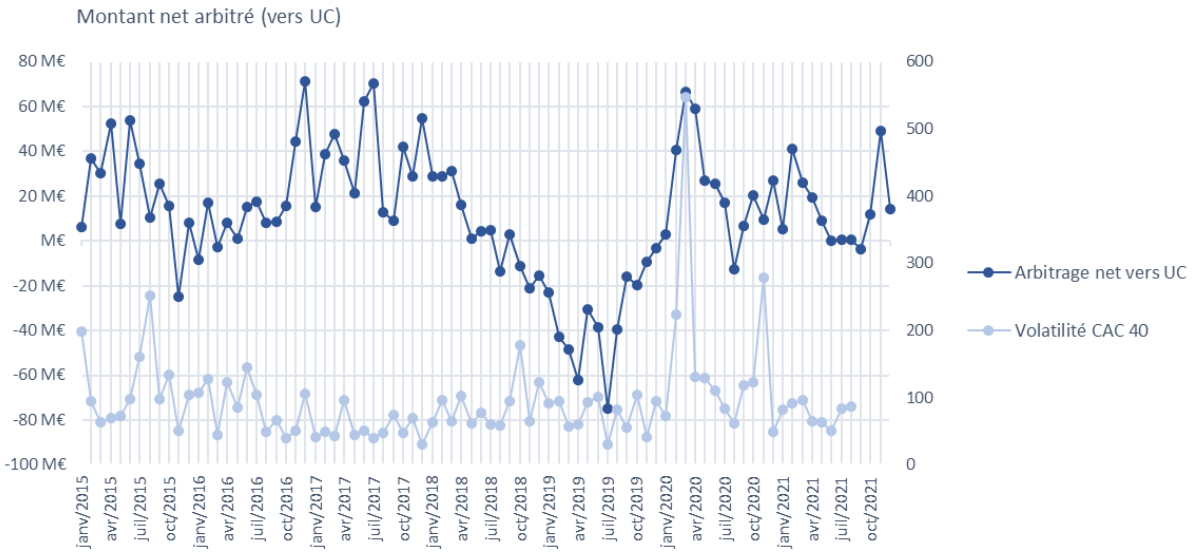


Figure 21: Valeurs mensuelles du montant net arbitré et de la volatilité du CAC40

- Taux OAT France 10 ans

Il représente en général le rendement à long terme des obligations émises par l'Etat. Souvent considéré comme représentatif du rendement sans risque à long terme, il est peut-être utilisé comme tel par nos assurés.

Graphiquement on ne distingue pas de corrélation nette, si ce n'est entre janvier 2018 et juillet 2019 où on observe simultanément une baisse quasi continue des deux grandeurs.

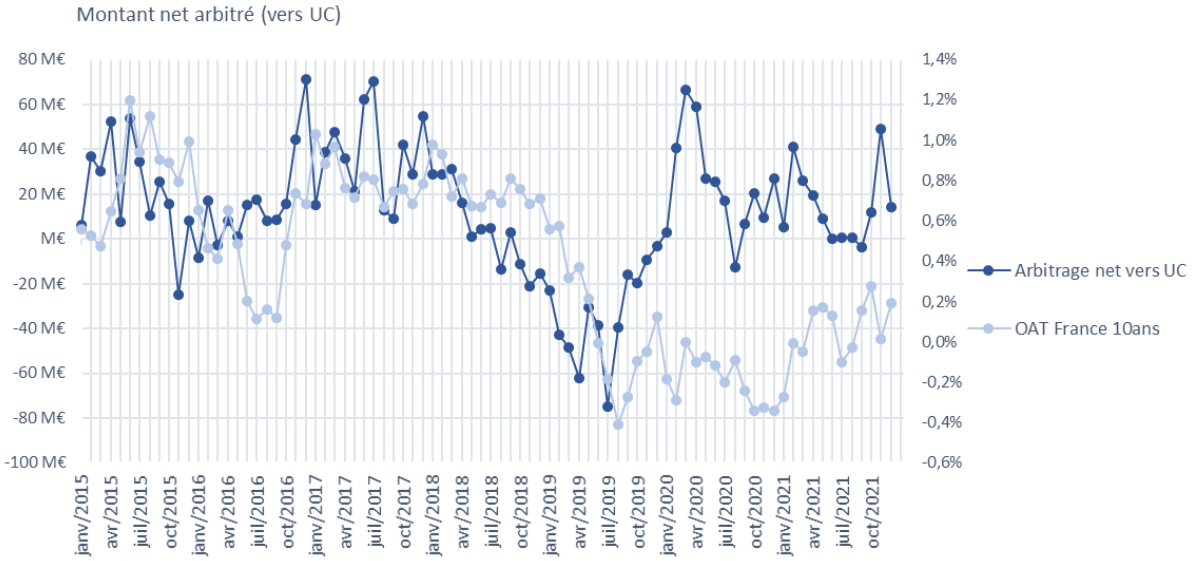


Figure 22 : Valeurs mensuelles du montant net arbitré et du taux de l'OAT France 10 ans

- Indice de confiance des ménages

Calculé mensuellement par l'INSEE, cet indicateur a pour objectif de synthétiser l'opinion des Français quant à la situation économique du moment. A priori on peut penser que plus les Français sont optimistes et confiants par rapport à la situation économique (score élevé), plus ils auront tendance à se diriger vers de l'UC.

Une nouvelle fois, on ne distingue pas de corrélation nette entre les deux grandeurs.

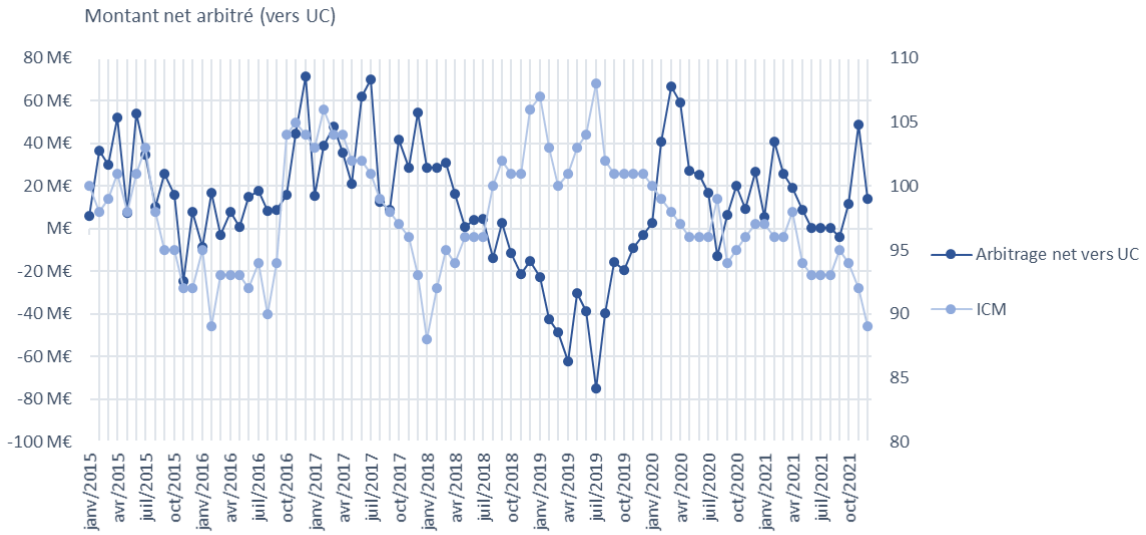


Figure 23 : Valeurs mensuelles du montant net arbitré et de l'ICM

- Ecart entre le rendement du contrat et le taux du CAC 40 l'année d'avant

Comme nous l'avons constaté, il semble que les variables financières en soient, ne soit pas pertinentes. Il est raisonnable de penser qu'en réalité, l'assuré va plutôt comparer le taux qu'on lui a servis pour ses placements avec les performances du fond en question (ici le CAC 40). Ainsi on modélise cette variable comme étant la différence entre le taux de plus-value réalisé sur son contrat en n-1 et le taux de rendement du CAC 40 en n-1 également. Il est alors raisonnable de penser que si un assuré observe que l'année précédente, les placements de son assurance vie ont significativement moins rapporté qu'un fond répliquant parfaitement les performances du CAC40, il serait tenté d'arbitrer davantage vers de l'UC.

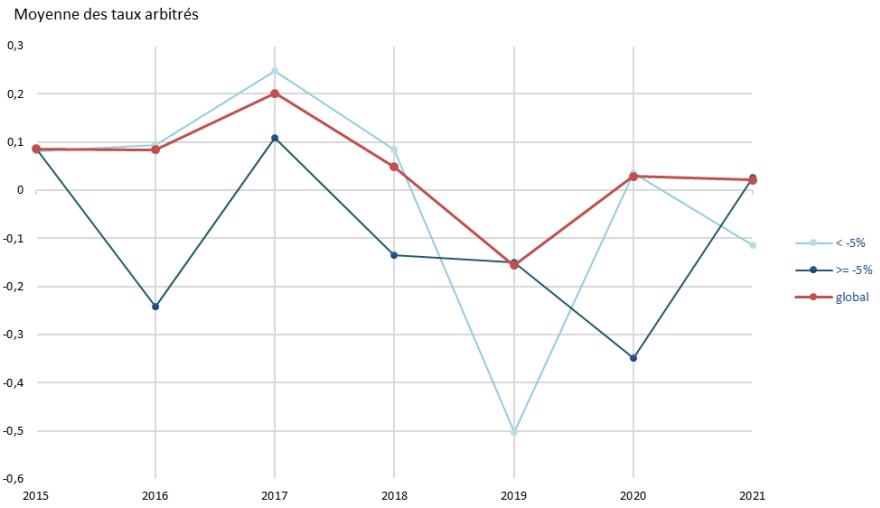


Figure 24 : Moyenne des taux arbitrés en fonction de l'écart avec le CAC40

[3]

C'est effectivement ce qui est observé en pratique, excepté pour les années 2019 et 2021 mais qui ne sont pas représentatives car la moyenne des taux arbitrés ces années a été calculée avec très peu de données.

Ceci s'explique aisément par le fait qu'en n-1 (respectivement 2018 et 2020), les performances annuelles du CAC 40 ont été mauvaises avec des rendements annuels négatifs. Rare sont donc les personnes ayant fait significativement moins.

En revanche, on constate que la fréquence d'arbitrage ne diffère pas entre les deux classes.

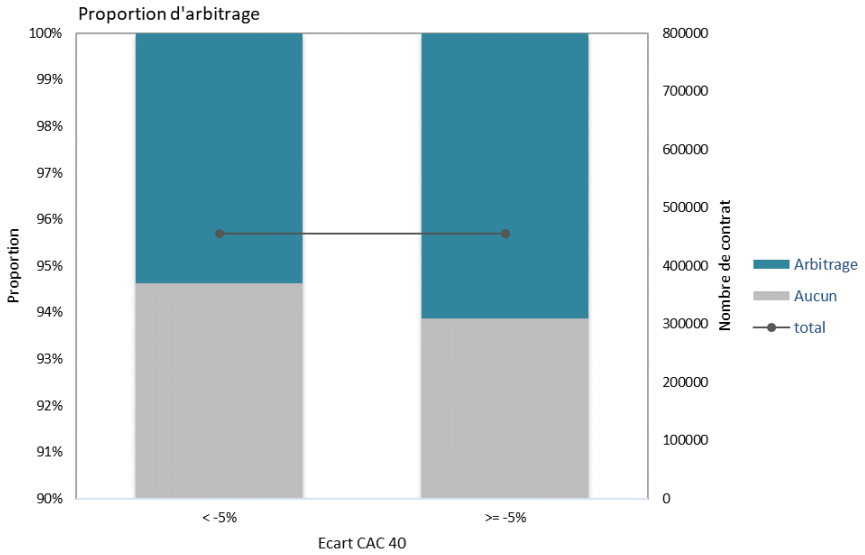


Figure 25 : Proportion d'arbitrage en fonction de l'âge de l'écart avec le CAC40

- Ecart entre le rendement du contrat et le taux de l'OAT France 10 ans l'année d'avant

[3] Le seuil de -5% correspond à l'écart médian.

On procède de la même manière. A priori on peut penser qu'un assuré observant en fin d'année que les performances de son contrat d'assurance vie sont inférieures au taux OAT 10 ans (souvent considéré comme le taux sans risque à long terme) sera tenté d'arbitrer vers l'UC l'année suivante.

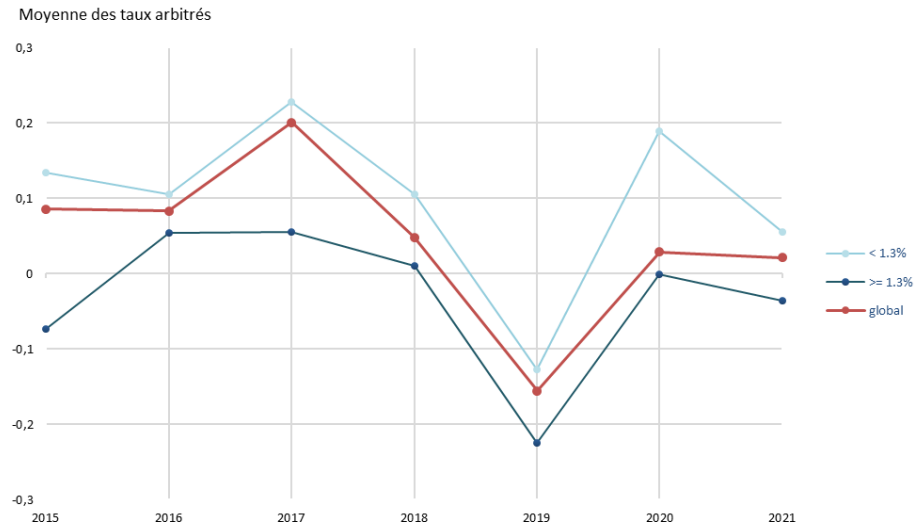


Figure 26 : Moyenne des taux arbitrés en fonction de l'écart avec l'OAT France 10 ans

[4]

C'est effectivement ce qui est observé. Cependant, on observe ici aussi que la fréquence d'arbitrage entre ces deux classes est très similaire.

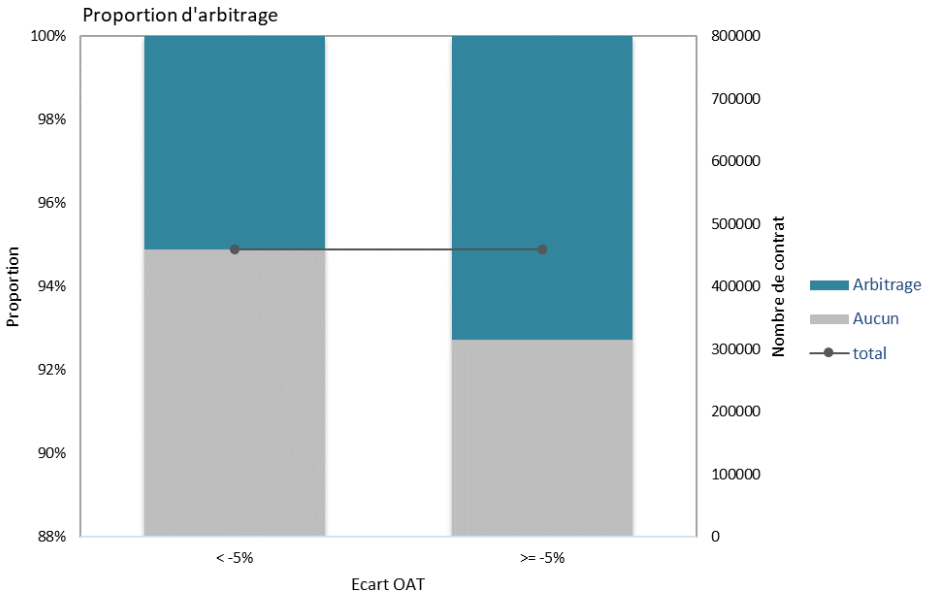


Figure 27 : Proportion d'arbitrage en fonction de l'âge de l'écart avec l'OAT France 10 ans

[4] Le seuil de 1.3% correspond à l'écart médian.

3.3. Nettoyage et retraitements des données

3.3.1. Au niveau arbitrage

Lorsqu'un assuré place une partie (ou la totalité) de son épargne sur de l'UC, il se voit généralement proposer des garanties optionnelles telles que « Garantie Sécu de la plus-value » ou encore « Garantie Stop Loss ».

Leur fonctionnement est en général assez simple, l'assureur se propose d'effectuer automatiquement, sans consultation de l'assuré, un arbitrage lorsque des conditions bien précises sont remplies. Ces conditions sont définies contractuellement au préalable avec l'assuré.

Par exemple pour une garantie Stop Loss, dès lors que l'épargne investie sur l'unité de compte devient inférieure au Floor (plancher) fixé par le souscripteur, un arbitrage automatique est effectué du support source vers le support cible (l'Euro la plupart du temps).

Ces arbitrages automatiques se sont beaucoup démocratisés ces dernières années. En 2015, ils représentaient environ 10% des arbitrages effectués, alors qu'en 2021, ils représentent près d'un tier des arbitrage effectués.

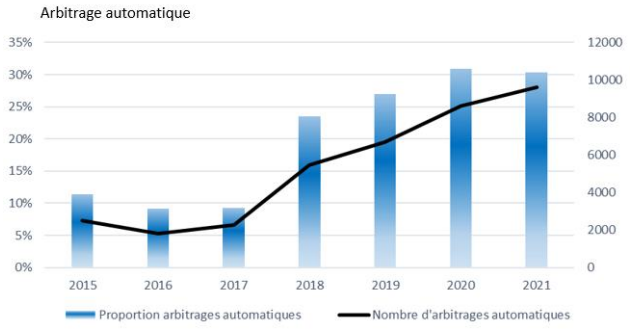


Figure 28 : Proportion d'arbitrage automatique

Ils ont néanmoins la particularité d'être d'un très faible montant (en proportion de la PM). En les retraitant, l'impact sur notre variable finale (montant annuel d'arbitrage net vers UC) est quasiment nul.

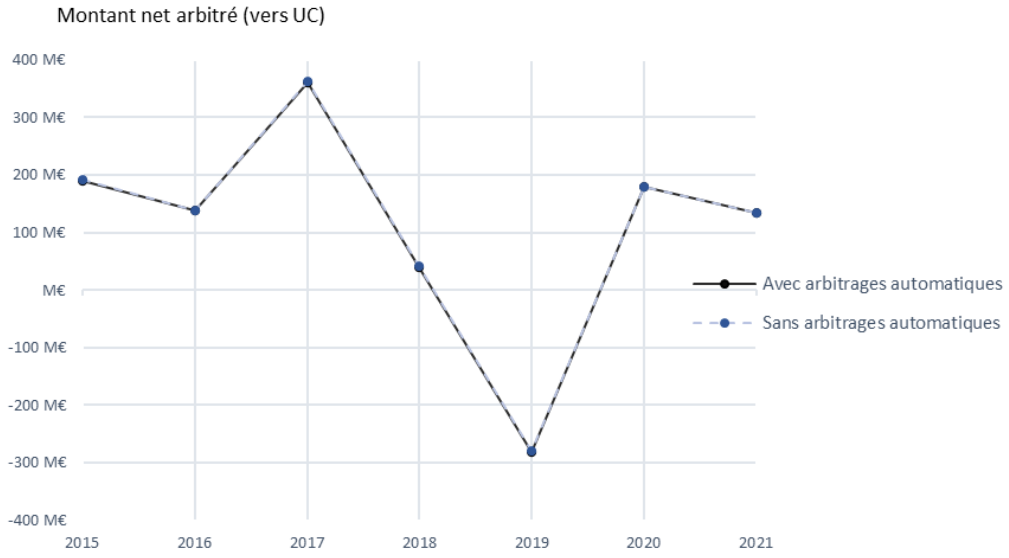


Figure 29 : Montant annuel arbitré (net vers UC) avec et sans les arbitrages automatiques

En revanche, comme nous pouvons le constater sur le tableau ci-contre, l'impact sur la structure de nos données est très important. Grâce à ce retraitement, nous avons pu améliorer considérablement nos prédictions qui se trouvaient alors fortement perturbées par ces arbitrages automatiques (sous-estimation des montants arbitrés prédits).

	Avec arbitrage automatique	Sans arbitrage automatique
Pourcentage de contrat qui arbitre / an	6,83%	5,37%
Montant moyen arbitré	12 492 €	15 112 €

Figure 30 : Impact des arbitrages automatiques sur les variables cibles

3.3.2. Au niveau contrat

209 088 contrats se trouvent dans notre périmètre d'étude (Portefeuille LMP souscrit entre le 01/01/2001 et le 31/12/2021). Certains retraitements et certaines modifications sont nécessaires.

- Les souscriptions annulées

209 088 correspond au nombre de souscriptions à l'intérieur de notre périmètre. À la suite d'une souscription, l'assuré possède 30 jours calendaires révolus pour se rétracter. [5] Si l'assuré décide d'exercer ce droit, le contrat est alors clôturé.

On décide donc naturellement de les retirer de notre jeu de données, soit 6 801 contrats.

- Dates aberrantes

Il peut arriver que certaines données aberrantes apparaissent, liées la plupart du temps à une erreur de saisie. Dans notre cas, on retrouve 2 contrats où la date de souscription est plus ancienne que la date de naissance. On retraite ces contrats.

- Données manquantes

Beaucoup de modèles que nous tenterons d'implémenter ne tolèrent pas les valeurs manquantes. Il est donc nécessaire de traiter ces données. Pour ce faire, il est important de se questionner sur la cause de leur manque dans nos données. Est-ce simplement dû au hasard ? Ou bien leur manque résulte-t-il d'un phénomène non-aléatoire ? La manière de traiter les données manquantes doit tenir compte de la réponse à ces questions.

Dans notre cas, il s'agissait soit de valeurs manquantes apparaissant de manière aléatoire, probablement dû à des erreurs de saisies, soit de valeurs manquantes apparaissant car certaines modalités de variables sont incompatibles entres elles.

Voici la liste des traitements effectués :

- Le versement initial de 720 contrats est manquant. On remplace la valeur manquante par la moyenne calculée sur les contrats du même produit.
- Le nombre de supports UC est manquant pour 389 contrats. On remplace également la valeur manquante par la moyenne calculée sur les contrats du même produit.
- L'âge est manquant pour toutes les personnes morales car ces assurés ne possèdent pas de date de naissance. L'information sur l'ancienneté de l'entité n'est pas disponible et ne ferait dans tous les cas pas sens. Au total, 5 529 contrats sont concernés. Etant donné qu'il n'y avait aucune

[5] Article L132-5-1 du code des assurances

différence significative au niveau des arbitrages entre les personnes morales, les hommes et les femmes, nous avons simplement retiré ces données.

- La PM au 31/12/n-1. Pour les contrats ayant moins d'un an d'ancienneté, cette donnée n'existe pas. Cela concerne 38 727 lignes, dans ce cas on remplace la PM par le versement initial.
- Dernier taux de PB versé. De même que pour la PM, cette donnée n'existe pas pour les contrats ayant moins d'un an d'ancienneté. On remplace alors le dernier taux de PB par la moyenne des derniers taux de PB versés cette année-là.

3.4. Base finale

Notre base finale se compose de 894 104 lignes et 39 colonnes. Une ligne correspond à un contrat et une année. Pour pouvoir rajouter certaines variables que l'on jugeait indispensables[6], nous observons les mouvements d'arbitrage des contrats de notre périmètre uniquement sur la période 2015-2021.

Voici un extrait de notre base où les numéros de police ont été modifiés par soucis de confidentialité.

Tableau 9 : Extrait de la base de modélisation

	NO_POLICE	annee	type_personne	Age	Ancienneté	lb_partenaire	versement_ini	pm	proportion_uc	taux_pb	arbitrage	montant_arbitre
129	A0200	2017	Femme	74	14	BANQUE PALATINE	34000	52132.18983	0	1.916	Aucun	0.00
130	A0200	2018	Femme	75	15	BANQUE PALATINE	34000	52854.951557	0	1.936	Aucun	0.00
131	A0200	2019	Femme	76	16	BANQUE PALATINE	34000	53730.230588	0	2.000	Aucun	0.00
132	A0200	2020	Femme	77	17	BANQUE PALATINE	34000	54620.0052	0	2.000	Aucun	0.00
133	A0200	2021	Femme	78	18	BANQUE PALATINE	34000	55524.515304	0	2.000	Aucun	0.00
134	A0200	2015	Homme	72	12	BANQUE PALATINE	1000	140832.47761	0.3139630853	2.850	Aucun	0.00
135	A0200	2016	Homme	73	13	BANQUE PALATINE	1000	143705.50347	0.2878535552	2.550	EU->UC	49950.00
136	A0200	2017	Homme	74	14	BANQUE PALATINE	1000	145272.25894	0.6447759656	1.950	Aucun	0.00
137	A0200	2018	Homme	75	15	BANQUE PALATINE	1000	152461.00057	0.6681519212	1.950	UC->EU	-69654.85
138	A0200	2019	Homme	76	16	BANQUE PALATINE	1000	144705.01179	0.4952135175	2.000	Aucun	0.00
139	A0200	2020	Homme	77	17	BANQUE PALATINE	1000	168124.77481	0.538557512	2.000	Aucun	0.00
140	A0200	2021	Homme	78	18	BANQUE PALATINE	1000	157091.9482	0.5599120392	1.250	UC->EU	-38491.05
141	A0200	2015	Femme	40	12	BANQUE PALATINE	1000	40997.934052	0	2.850	Aucun	0.00
142	A0200	2016	Femme	41	13	BANQUE PALATINE	1000	41881.33137	0	2.550	Aucun	0.00
143	A0200	2017	Femme	42	14	BANQUE PALATINE	1000	42265.227248	0	1.950	Aucun	0.00
144	A0200	2018	Femme	43	15	BANQUE PALATINE	1000	42853.186294	0	1.950	Aucun	0.00
145	A0200	2019	Femme	44	16	BANQUE PALATINE	1000	43916.904493	0	2.000	Aucun	0.00
146	A0200	2020	Femme	45	17	BANQUE PALATINE	1000	44203.76066	0	2.000	Aucun	0.00
147	A0200	2021	Femme	46	18	BANQUE PALATINE	1000	44969.273117	0	2.000	Aucun	0.00

Variable cible

3.5. Analyse des corrélations entre les variables explicatives

Regarder la corrélation entre deux variables (quantitatives) permet de quantifier la relation de dépendance qu'il y a entre elles, ainsi que le sens de la dépendance.

Néanmoins, il existe plusieurs types de relation de dépendance et donc plusieurs métriques de corrélation. La plus utilisée est de loin le coefficient de corrélation de Pearson. Il mesure la corrélation linéaire entre deux variables X et Y .

Soit $\{ x_1, y_1 , \dots , x_n, y_n \}$ un échantillon de réalisations indépendantes de X et Y. Le coefficient de corrélation de Pearson entre ces deux variables, noté r_{xy} se calcule de la façon suivante :

[6] C'est par exemple le cas de la PM, le dernier taux de pb servis. Ces informations sont recensées dans des bases de données uniquement à partir de 2015 mais pas avant.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Malheureusement, en réalité les relations linéaires entre variables sont plutôt rares. Nous allons donc utiliser une métrique plus large, qui permet de mieux capter les relations monotoniques et non linéaires, le coefficient de corrélation de Spearman. Il s'agit en fait de la même formule que le coefficient de corrélation de Pearson mais appliquée non pas aux valeurs $\{ x_1, y_1 \ , \dots , \ x_n, y_n \}$ mais à leurs rangs dans l'échantillon.

L'idéal pour la construction de nos modèles serait d'avoir des variables explicatives décorrélées. En effet, une des hypothèses théoriques nécessaire à l'application de modèles statistiques classiques tels que les GLM est l'indépendance des variables explicatives. Pour les modèles de machine learning, bien qu'il n'y ait aucunes hypothèses théoriques à vérifier, avoir des variables fortement corrélées peut introduire un biais et une non-convergence du modèle.

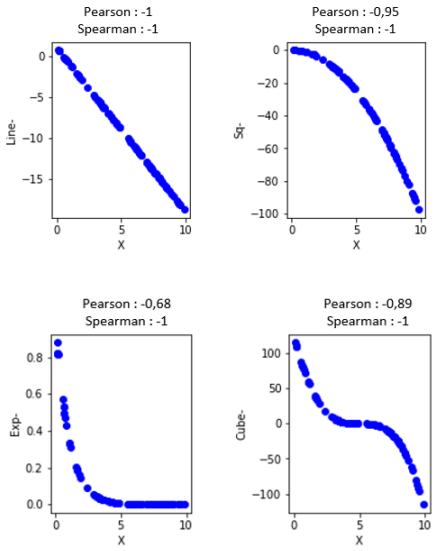


Figure 31 : Différence corrélation Pearson / Spearman

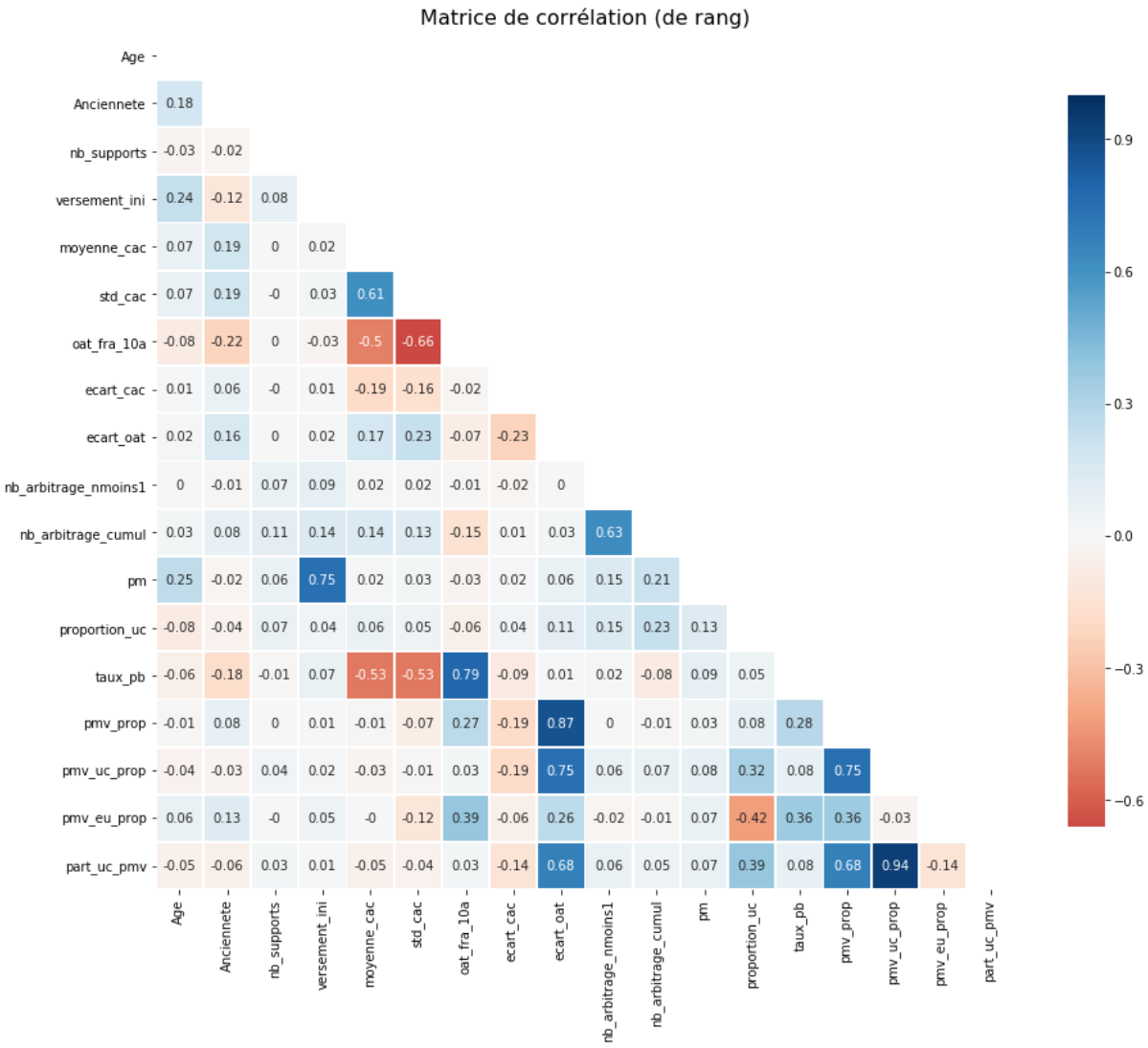


Figure 32 : Matrice de corrélation (calculée avec le coefficient de Spearman)

Globalement nos variables explicatives ne sont pas très corrélées. Les corrélations importantes qui sont observées ne sont pas surprenantes.

Par exemple entre le taux de plus ou moins-value sur l'UC (pmv_uc_prop) avec le pourcentage de plus ou moins-value du contrat lié à l'UC (part_uc_pmv). En effet, l'un fait mécaniquement augmenter l'autre. Autre exemple ; le taux de PB versé avec le taux de l'OAT France sur 10ans. Ici aussi c'est compréhensible étant donné que plus de 80% de la PM EURO est placé par l'assureur sur des obligations, dont une bonne partie d'obligation d'Etat (un tier environ) .

Au total, sur les 153 coefficients de corrélation de notre matrice, 11 ont un coefficient (en valeur absolue) supérieur à 0.6. C'est un élément à garder en tête. Dans l'optique où nos modèles sont mauvais, c'est un des leviers que l'on pourrait actionner pour les améliorer.

PARTIE 2 : MODELISATION

1. Mise en place de la modélisation

1.1. Méthodes de modélisation

1.1.1. Modélisation en une étape

Pour modéliser notre variable cible Y (le taux d'arbitrage), la première méthode à laquelle nous pensons est de prendre toutes nos données d'entrainement et de construire un modèle de régression en utilisant nos variables explicatives $X_1, \dots, X_p$. C'est la méthode classique, où l'on va modéliser simplement Y comme une fonction, aussi complexe soit-elle, de $X_1, \dots, X_p$.

Dans notre cas, la structure de notre variable cible ne s'adapte pas bien à cette façon de procéder. En effet, comme nous l'avons dit plus haut, sur notre période d'observation on remarque que pour une année donnée, seulement 5.30% des contrats effectuent un arbitrage. La conséquence est que sur notre base de données (qui est à la maille contrat x année) d'environ 900 000 lignes, il y a près de 840 000 lignes où le taux d'arbitrage est nul. La modélisation est alors rendue extrêmement complexe dans la mesure où tous les modèles auront tendance à prédire un taux d'arbitrage nul. On parle alors de problème de parcimonie des données (*data sparsity* en anglais).

Il s'est avéré que, peu importe le modèle implémenté, les résultats en utilisant cette méthode ont toujours été médiocres. Ils ne seront donc pas détaillés dans la suite.

1.1.2. Modélisation par fréquence sévérité

Pour les raisons qui ont été explicitées plus haut, il s'est avéré nécessaire de procéder en deux étapes avec une modélisation du type Fréquence/Sévérité.

L'idée est de créer deux modèles distincts :

- Un modèle de classification modélisant la variable binaire *arbitrage_dum* qui indique uniquement si oui ou non l'assuré a arbitré.
- Un modèle de régression modélisant le taux d'arbitrage construit uniquement à partir des lignes où il y a eu arbitrage. Autrement dit, on modélise ici le taux d'arbitrage conditionnellement au fait qu'il y a effectivement eu un arbitrage.

On effectue ensuite le produit des deux prédictions.

Cette façon de procéder s'est avérée légèrement plus efficace que la première mais reste cependant très limitée car le modèle de classification est construit en utilisant toutes les données d'entrainements, on se retrouve donc avec le même problème de parcimonie des données. Il est néanmoins plus facile à gérer pour une classification binaire, étant donné qu'il se résume donc à un problème de classification déséquilibrée.

Sans traitement de notre part, le modèle de classification aura tendance à prédire « non » dans la quasi-totalité du temps. En effet, en prédisant systématiquement, de manière bête : « non », le modèle s'assure un taux de bonnes réponses de plus de 93%.

Fort heureusement, c'est un problème assez connu pour lequel des solutions existent et permettent d'améliorer les résultats (sans toutefois faire des miracles). Les solutions les plus connues sont :

- Le rééchantillonnage (*undersampling* et *oversampling*)
- Les modèles pénalisés
- Les algorithmes de génération d'échantillons synthétiques (SMOTE, ClusterCentroids, …)

Cette dernière solution est nettement plus complexe que les deux premières et suppose également une structure des données particulière. Les deux premières offrent en général des résultats très proches, nous avons donc fait le choix d'utiliser la méthode de rééchantillonnage qui se trouve être la plus simple et la plus transparente des deux méthodes.

Le sous-échantillonnage (*undersampling* en anglais) consiste simplement à retirer aléatoirement un certain nombre d'observations de la classe majoritaire dans le but de rendre les classes plus équilibrées. Elle permet dans un même temps de réduire les temps de calcul. Néanmoins elle cause une perte d'information sur la classe majoritaire et augmente la variance de nos prédictions.

Le sur-échantillonnage (*oversampling* en anglais) consiste à augmenter artificiellement le nombre de données de la classe minoritaire en tirant aléatoirement et avec remise (boostrap) un certain nombre de fois dans les données de cette classe. Aucune information n'est alors perdue. Cette technique s'avère être robuste mais augmente les temps de calculs et le risque de sur-apprentissage (c.f partie 1.2).

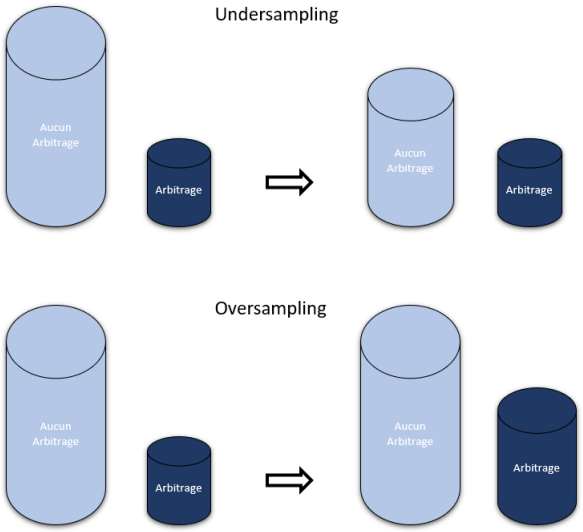


Figure 33 : Illustration du réechantillonnage

En pratique, on utilise les deux méthodes simultanément ce qui permet d'obtenir un bon compromis. On effectue ces méthodes jusqu'à atteindre une répartition classe majoritaire/classe minoritaire fixée au préalable (50/50 ; 70/30 ; …).

On entraîne ensuite le modèle sur les données rééchantillonnées puis on le test sur les données initiales.

1.1.3. Modélisation par fréquence sévérité probabiliste

L'idée derrière est de modéliser l'espérance du montant arbitré et non pas le montant lui-même. Les prédictions individuelles n'ont donc pas vraiment de sens, mais en agrégant les prédictions de l'année on espère observer une convergence (loi des grands nombres). L'intérêt de cette méthode est qu'elle permet également de décomposer le problème.

Soit A la variable aléatoire discrète qui prend la valeur 1 quand il y a arbitrage et 0 sinon, on a d'après le théorème de l'espérance totale :

$$\begin{aligned}\mathbb{E}(Y_i) &= \mathbb{E}(\mathbb{E}(Y_i|A_i)\\ &= \sum_a \mathbb{E}(Y_i|A_i = a)\mathbb{P}(A_i = a)\\ &= \mathbb{E}(Y_i|A_i = 0)\mathbb{P}(A_i = 0) + \mathbb{E}(Y_i|A_i = 1)\mathbb{P}(A_i = 1)\\ &= \mathbb{E}(Y_i|A_i = 1)\mathbb{P}(A_i = 1)\end{aligned}$$

En effet, l'espérance du taux d'arbitrage de l'individu i conditionnelle au fait qu'il n'ait pas arbitré est nulle.

En se basant sur ce résultat, nous avons donc tenté de modéliser l'espérance du taux arbitré en faisant le produit du modèle de régression de la section 1.1.2 et d'une probabilité d'arbitrage estimée.

1.2. Sur apprentissage

En modélisation, on parle de surapprentissage (en anglais « overfitting ») lorsque l'analyse et les prédictions produites correspondent trop précisément à un jeu de données. Il s'interprète comme un apprentissage « par cœur » des données.

Par conséquent, elles correspondent mal, voire pas du tout, à des données supplémentaires. Les prédictions souffriront d'une grande variance.

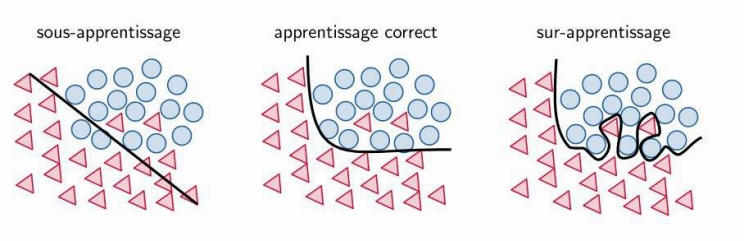


Figure 34 : Illustration du sur/sous apprentissage

A noter qu'il ne faut pas non plus tomber dans le problème inverse du sous-apprentissage (en anglais « underfitting ») avec un modèle trop généraliste, incapable de fournir une analyse et des prédictions précises. On dit également que le modèle souffre d'un grand biais.

Ce sont des problèmes extrêmement fréquents en modélisation, heureusement, il existe des méthodes permettant de les limiter grandement, afin de trouver le juste milieu.

1.3. Base d'entrainement et base de test

Avant de poursuivre, il est nécessaire de diviser notre base de données initiale en deux bases de données :

- Une base d'entrainement.
- Une base de test.

Procéder ainsi a pour objectif d'éviter le surapprentissage de nos modèles (*overfitting*), et de mesurer leurs performances sur des données qu'ils n'ont jamais rencontrées.

En effet, si nous entrainons notre modèle sur certaines données, ce dernier sera naturellement plus performant sur ces données-là. Or, ce qui nous intéresse, c'est de mesurer sa performance sur des données

jamais rencontrées auparavant. On appelle cette performance la généralisation du modèle : il s'agit de sa capacité à effectuer des prédictions de qualité sur des nouvelles situations.

Il faut ainsi trouver un équilibre entre la base d'apprentissage, et la base de test.

Généralement, les proportions choisies sont 80% de la base initiale (tirés aléatoirement sans remise) pour la base d'apprentissage, et 20% pour la base test.

Pour notre étude nous avons procédé différemment. Etant donné que notre objectif est de pouvoir prédire le montant d'arbitrage net l'année suivante, nous avons choisi d'entraîner notre modèle sur les données de 2015 à 2019 (soit 72% de nos données), et de le tester sur les données de 2020 et 2021 (soit 28% de nos données).

1.4. Détermination des hyperparamètres par validation croisée

Tous les modèles que nous allons essayer d'implémenter possèdent des paramètres (hyperparamètres), parfois nombreux, dont la valeur est à optimiser. Pour réaliser ces choix, nous allons comparer les performances d'un modèle donné pour chaque paramètre (ou combinaison de paramètres possible).

Une nouvelle fois on risque de sur-ajuster si on évalue la performance sur la base qui nous a servi pour l'apprentissage. Il nous faut donc une fois encore séparer la base.

Effectuer une séparation simple comme ce qu'on a fait précédemment est néanmoins ici dangereux car le(s) hyperparamètre(s) optimal(aux) seront très dépendants de cette séparation. Pour contrer ça, on procède par validation croisée (ou « *cross validation* » en anglais).

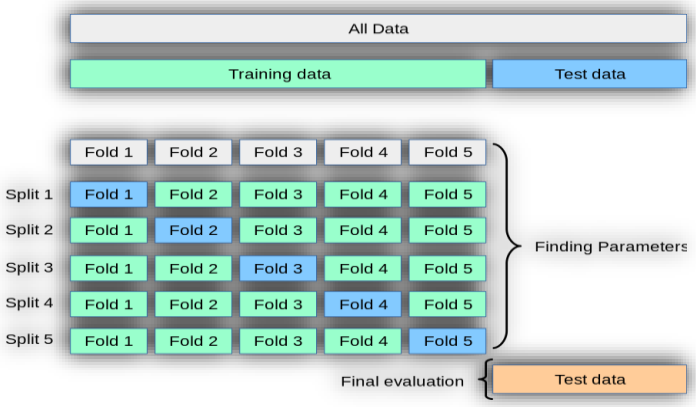


Figure 35 : Illustration de la validation croisée

Cette méthode consiste à découper notre **base d'apprentissage** en K parts égales (folds). L'ajustement du modèle est réalisé avec les folds K-1 (K moins 1), puis le modèle est validé en utilisant le K-fold restant. Tous les scores ainsi que les erreurs doivent être notés. On répète le même processus jusqu'à ce que tous les K-fold servent dans l'ensemble d'entraînement. La moyenne et l'écart type des k scores de performances peuvent être calculés pour estimer le biais et la variance de la performance de validation. On gardera donc les hyperparamètres offrant le meilleur compromis. Mais alors quels scores de performances choisir ?

2. Scores de performances

Le choix des scores de performances est crucial car ils déterminent non seulement le choix des hyperparamètres, mais également le choix du modèle à retenir.

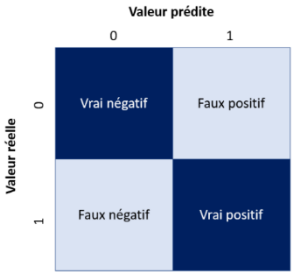


Figure 36 : Matrice de confusion

2.1. Modèles de classification

Pour évaluer un modèle de classification, on peut utiliser des métriques se basant sur les prédictions « brutes », ou bien des métriques se basant sur les probabilités.

- Métriques basées sur les prédictions
  - *L'accuracy* : Elle mesure le taux de prédictions correctes sur l'ensemble des données.

$$Accuracy = \frac{Vrai\ positif + Vrai\ négatif}{Total}$$

Cet indicateur est de loin le plus utilisé pour évaluer des modèles de classification. Néanmoins il n'est pas très pertinent lorsque les classes sont déséquilibrées, ce qui sera notre cas. Il nous faut donc utiliser d'autres scores d'évaluation.

  - La précision : Elle mesure l'exactitude des prédictions de la classe positive soit, dans notre cas, à quel point nous somme précis lorsque l'on prédit qu'il y aura un arbitrage.

$$Précision = \frac{Vrai\ positif}{Vrai\ positif + Faux\ positif}$$

  - Le rappel (*recall* en anglais) : Elle mesure la capacité du modèle à détecter les données de la classe positive, qui sont dans notre cas, les assurés ayant réalisé un ou plusieurs arbitrages.

$$Rappel = \frac{Vrai\ positif}{Vrai\ positif + Faux\ négatif}$$

Ces deux mesures s'opposent, dans le sens où, chercher à augmenter l'une fera mécaniquement baisser l'autre. Ensemble elles permettent d'apprécier la validité d'une classification, mais séparément, elles ne veulent rien dire. Or, pour fixer nos hyperparamètres par validation croisée comme expliqué en *1.4*, nous ne pouvons utiliser qu'une seule métrique. Par chance, il existe un indicateur qui fournit la synthèse de ces deux mesures.

- L'aire sous la courbe Précision/Rappel (PR_auc) : Cette courbe nous indique simplement le taux de précision pour un rappel donné. Elle permet de voir à quel point l'augmentation de l'un dégrade l'autre. Plus un modèle est performant, plus il sera capable de maintenir un niveau de précision élevé à mesure que le rappel augmente. Ainsi, calculer l'aire sous cette courbe nous donne une mesure robuste de la qualité d'une classification. Elle est comprise entre 0 (le score correspondant à des prédictions au hasard) et 1 (classifieur parfait). C'est une des mesures les plus utilisées pour de la classification très déséquilibrée.

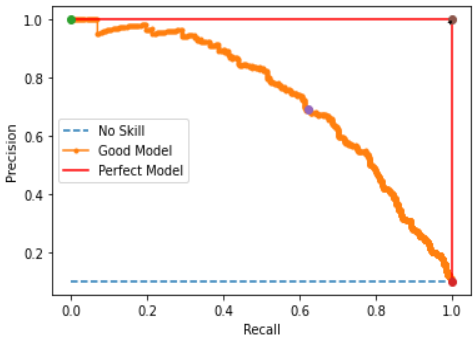
  

  Figure 37 : Aire sous la courbe précision/rappel

  C'est cet indicateur que nous utiliserons pour déterminer les hyperparamètres optimaux. [7]

- Métrique basée sur les probabilités
  - Le log loss : Cette mesure va comparer chaque probabilité prédite à la valeur réelle de sortie. L'avantage de cette mesure est que pour 2 prédictions ratées, le coût de l'erreur ne sera pas forcément le même. En effet, l'erreur sera plus importante si notre modèle prédit « arbitrage » avec une probabilité de 98% et qu'il se trompe, que s'il prédit « arbitrage » avec une probabilité de 51% et qu'il se trompe.
    Cette façon de faire est donc parfaitement adaptée pour un modèle avec des classes déséquilibrées.
    La nature de la pénalité est logarithmique. C'est-à-dire qu'une grande différence contiendra un score énorme tel que 0.9 ou 10. Cependant, les différences plus faibles contiendront de petits scores tels que 0.1 ou 0.2. Plus la valeur est proche de 0, meilleur est le modèle.

$$LogLoss = -\frac{1}{n}\sum_{i=0}^{n}[y_i log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

2.2. Modèles de régression

Comme pour la classification, nous avons besoin de mesures pertinentes pour pouvoir évaluer et comparer les modèles de régression.

Dans notre modèle de régression, la variable à prédire Y correspond au taux d'arbitrage (la proportion de la PM arbitrée). Pour chaque individu i nous avons donc le taux prédit y_i et le taux observé y_i.

On définit ainsi le résidu comme étant l'écart entre ces deux valeurs.

[7] Néanmoins elle ne doit pas être utilisé comme fonction de coût optimisé directement par l'algorithme car elle ne possède pas les propriétés de régularité et de convexité nécessaire.

$$\varepsilon_i = y_i - y_i$$

Plus le résidu est faible, meilleure est la prédiction.

- L'erreur quadratique moyenne

  Un indicateur naturel de la qualité de notre régression est de regarder l'erreur moyenne (au carré) de nos prédictions.

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

  Il est important de prendre les erreurs au carré pour pas qu'elles se compensent, étant donné que nos erreurs peuvent être positives et négatives.

  Notons qu'il est également possible de prendre la moyenne des erreurs en valeur absolue. Mais cette façon de faire, bien que réputée plus stable, est beaucoup plus gourmande en calcul.

- Le coefficient de détermination R^2

  Ce coefficient nous donne une estimation du pourcentage de la variance qui est expliquée par notre modèle. Pour le calculer on se base sur l'égalité suivante, directement donnée par le théorème de Pythagore :

$$\begin{aligned} \|Y - \bar{y}\mathbb{1}\|^2 &= \|\hat{Y} - \bar{y}\mathbb{1}\|^2 + \|\hat{\varepsilon}\|^2 \\ \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 \\ SCT &= SCE + SCR, \end{aligned}$$

  Avec :

  - SCT : La Somme des Carrés Totale, qui représente la variabilité totale.
  - SCE : La Somme des Carrés Expliquée, qui représente la variabilité expliqué par le modèle.
  - SCR : La Somme des Carrés Résiduelle, qui représente la variabilité que notre modèle ne parvient pas à expliquer.

  Le calcul du coefficient de détermination est alors simplement le rapport entre la variabilité expliqué par notre modèle sur la variabilité totale.

$$R^2 = \frac{SCE}{SCT} = \frac{\|\hat{Y} - \bar{y}\mathbb{1}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = 1 - \frac{SCR}{SCT}$$

  La valeur de ce coefficient est comprise entre 0 et 1. Plus il est proche de 1, meilleur est le modèle.

3. Les différents modèles utilisés

3.1. Modèle linéaire généralisé

3.1.1. Généralités sur le modèle linéaire généralisé

Présenté pour la première fois sous ce nom (*Generalized linear model* (GLM) en anglais) en 1972 par John Nelder et Robert Wedderburn, il fût néanmoins exposé de façon complète qu'en 1989 grâce à l'aide de Peter McGullagh. C'est une extension du modèle bien connu de régression linéaire multiple.

Il permet d'étudier la liaison entre une variable réponse Y et des variables explicatives $X_1, \dots, X_p$.

Ils sont formés de trois composantes :

- La variable réponse Y qui est une composante aléatoire à laquelle est associée une loi de probabilité qui appartient à la famille exponentielle.
- Les variables explicatives $X_1, \dots, X_p$ qui sont des composantes déterministes utilisées pour prédire notre variable réponse.
- La fonction de lien g qui décrit la relation entre la combinaison linéaire des variables explicatives et l'espérance mathématiques de la variable réponse. Elle est supposée monotone et différentiable.

Autrement dit on suppose la relation suivante :

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

Avec $\beta_0, \dots, \beta_p$ les paramètres à estimer.

Par exemple, en prenant g la fonction identité, on se retrouve alors dans le cas de la régression linéaire multiple.

A travers les différentes lois de probabilités et fonctions de liens, le GLM peut modéliser à la fois des variables quantitatives et catégorielles.

Tableau 10 : GLM fonction de lien canonique

Loi	Lien canonique $g(\mu)$
Normale	μ
Poisson	$ln(\mu)$
Gamma	$\frac{1}{\mu}$
Binomiale	$ln(\frac{\mu}{(1-\mu)})$

Pour rappel, Y de densité $f_{\theta,\phi}$ appartient à la famille exponentielle si $f_{\theta,\phi}$ s'écrit :

$$f(y|\theta,\phi) = exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right), \qquad y \in S$$

Où :

- θ est un réel, appelé paramètre naturel ou encore paramètre de la moyenne.
- ϕ est un réel strictement positif, appelé paramètre de dispersion (lié à la variance de la loi).
- S le support de la loi, sous-ensemble de $\mathbb{R}$ ou $\mathbb{N}$.
- $b(.)$ et $c(.)$ des fonctions réelles, $b(.)$ devant être deux fois dérivable.

Les densités des lois classiques telles que la loi normale, binomiale, Poisson, Gamma, … font partie de la famille exponentielle et peuvent donc se mettre sous cette forme.

Tableau 11 : valeurs des paramètres pour mettre les lois usuelles sous la forme canonique exponentielle

Loi	**S**	$\boldsymbol{\theta}$	$\boldsymbol{\phi}$	$\boldsymbol{b(\theta)}$	$\boldsymbol{c(y,\theta)}$
$N(\mu,\, \sigma^2)$	$\mathbf{R}$	μ	σ^2	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+ln(2\pi\sigma^2)\right)$
$Gamma(\nu,\, \mu)$	$\mathbf{R}^+$	$-\frac{1}{\mu}$	$\frac{1}{\nu}$	$-ln(-\theta)$	$\nu ln(\nu y)-ln(y)-ln(\Gamma(\nu))$
$Bin(n,\, p)$	$\mathbf{N}$	$ln(\frac{p}{1-p})$	1	$nln(1+exp(\theta))$	$ln\binom{n}{y}$
$Pois(\lambda)$	$\mathbf{N}$	$ln(y)$	1	$exp(\theta)$	$-ln(y!)$

On sait que la log vraisemblance d'un GLM s'écrit :

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

La condition du premier ordre donne alors :

$$\frac{\partial l}{\partial\theta}=\frac{y-b'(\theta)}{\phi}=0$$
$$\Rightarrow \mathbb{E}(\frac{\partial l}{\partial\theta})=\frac{\mathbb{E}(y)-b'(\theta)}{\phi}=0$$

D'où il vient :

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

Pour simplifier les notations, on pose :

$$\mathbb{X}=(\mathbf{1},X_1,...,X_p)$$

$$\beta=(\beta_0,...,\beta_p)$$

Et on a alors que :

$$\begin{aligned}\theta&=(b')^{-1}(y)\\&=(b')^{-1}(g^{-1}(\mathbb{X}^T\beta))\\&=h(\mathbb{X}^T\beta)\end{aligned}$$

Donc finalement, pour un échantillon x_1,y_1 ,…, x_n,y_n , x_j un vecteur de dimension $p+1$ quel que soit j, nous avons la log vraisemblance suivante :

$$l_n(\beta, y, X) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi}$$
$$= \sum_{i=1}^{n} \frac{y_i h(x_i^T\beta) - b(h(x_i^T\beta))}{\phi}$$

On estime ensuite les paramètres de notre modèle en maximisant cette vraisemblance.

3.1.2. Pénalisations

Lorsque l'on a beaucoup de variables explicatives et/ou beaucoup de modalités, on fait face à plusieurs risques. Le sur-apprentissage ainsi que la non-convergence du modèle font partis de ces risques. Or, toutes les modalités et variables explicatives ne sont pas forcément toutes pertinentes pour expliquer notre variable cible. C'est pour ces raisons qu'il est parfois souhaitable de pénaliser, voire d'exclure certains coefficients/variables. Il existe plusieurs façons de faire.

La pénalisation Ridge va pour ce faire intégrer à la quantité qui doit être maximisée, une fonction de coût se basant sur la norme L2 (qui correspond à la distance euclidienne).

$$l_n(\beta, y, X) - \lambda||\beta||_2^2 = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} - \lambda\sum_{j=1}^{p}\beta_j^2$$

Avec $\lambda \geq 0$ un paramètre qui sera déterminer par cross validation. Plus la valeur de λ est proche de 0, plus on s'approche de la solution classique, non pénalisée. Plus λ est élevée, plus la pénalisation est importante.

L'avantage de cette méthode par rapport aux méthodes classiques de selection (*forward stepwise* et *backward stepwise*) est qu'elle conserve toutes les variables du modèle. Les coefficients β des modalités peu pertinentes seront pénalisés de sorte que le coefficient soit proche de 0, sans jamais être exactement 0. Cette méthode permet donc de contourner les problèmes de colinéarité dans un contexte où il y a beaucoup de variables explicatives (ce qui est notre cas).

L'avantage de cette méthode peut aussi s'avérer être un désavantage. En effet, on peut penser qu'il peut parfois être souhaitable d'annuler complétement certains coefficients (c'est-à-dire de retirer une ou plusieurs variables explicatives).

La pénalisation de Lasso permet de faire ça. Elle consiste à maximiser la log vraisemblance auquel on retire une fonction de coût, se basant cette fois sur la norme L1 (qui correspond à la distance de Manhattan).

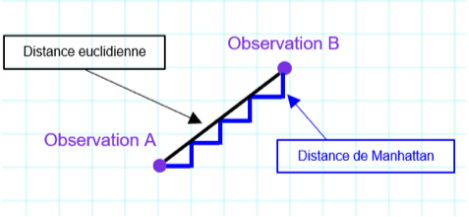


Figure 38 : Norme L1 et L2

$$l_n(\beta, y, X) - \lambda||\beta||_2^2 = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} - \lambda\sum_{j=1}^{p}|\beta_j|$$

Avec toujours $\lambda \geq 0$ un paramètre qui sera déterminé par cross validation.

3.1.3. Discrétisation

Il est souvent conseillé de discrétiser les variables continues lors de la mise en place d'un GLM. Cette pratique a notamment pour but d'essayer d'éviter la linéarité de leur effet sur notre variable cible. En effet, dans un GLM, l'effet d'une variable explicative continue sur la variable continue est par construction monotone, empêchant ainsi le modèle de traduire un changement de comportement à partir d'un certain seuil ou encore un effet non linéaire.

Par exemple : l'impact de la proportion d'UC dans son contrat sur la probabilité de faire un arbitrage (cf section 1.3.2.1). Si l'on ne discrétise pas cette variable, la probabilité d'arbitrer sera croissante avec la proportion d'UC dans le contrat. Or on observe en réalité que les contrats qui arbitrent le plus sont les contrats les plus équilibrés entre fonds Euro fonds UC. Ceux ayant très peu d'UC (entre 0 et 10% de leur épargne) ou à l'inverse ceux en ayant beaucoup (plus de 80%) arbitrent moins. En gardant la variable continue, le modèle sera incapable d'identifier et de prédire ce type de comportement.

Néanmoins, en faisant ça on synthétise l'information et on peut donc en perdre une partie. Une mauvaise discrétisation des variables peut masquer ou sous-estimer l'effet d'une variable explicative sur la variable cible et créer des problèmes de convergence des paramètres.

Une « bonne » discrétisation créer des classes homogènes et séparées, ce qui correspond aux notions statistiques de faibles variances intra classes et fortes variances interclasses. Il faut également que les classes soient les plus équilibrées possible. Plusieurs méthodes existent et se proposent de trouver la meilleure discrétisation d'une variable explicative en fonction de la variable cible. Il y a notamment la discrétisation se basant sur la maximisation de l'AIC de Adrien Ehrhardt ou encore celle se basant sur la pénalisation Lasso de Sander Devriendt. Or comme nous agrégeons le résultat de deux modèles avec deux variables cibles différentes, ces méthodes seraient trop complexes à mettre en place car il y aurait pour chaque modèle une discrétisation.

Nous avons donc fait le choix de discrétiser les données par la méthode classique des quantiles. Soit N le nombre de données, n le nombre de classes et F le nombre d'individus par classe. On trie les données dans un ordre croissant et on met les F premières données dans la classe 1, les F suivantes dans la classe 2 et ainsi de suite. On obtient ainsi des classes équiréparties. Le nombre de classes a été fixé au cas par cas, pour avoir des classes les plus homogènes et séparées possible. Cette façon de faire n'est pas optimale mais est assez robuste et nous permet d'avoir une unique discrétisation pour les deux modèles (classification et régression).

3.1.4. Régression Logistique

Pour implémenter notre modèle de classification dans lequel on souhaite prédire si oui ou non la personne arbitrera dans l'année, on choisit naturellement d'utiliser la régression logistique.

Y est donc une variable binaire et la régression logistique permet de modéliser la probabilité conditionnelle $\mathbb{P}(Y = 1|X = x)$.

La valeur d'une probabilité devant être comprise entre 0 et 1, on utilise la fonction logit comme fonction de lien.

$$
\begin{aligned}
g(\mathbb{P}(Y=1|X)) &= logit(\mathbb{P}(Y=1|X)) \\
&= ln(\frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)}) \\
&= \beta_0 + \sum_{i=1}^{p} \beta_i X_i
\end{aligned}
$$

Comme nous l'avons vu dans la section précédente, on suppose ici que la loi de Y conditionnellement à X suit une loi Binomiale de paramètres $n = 1$ et $p(x) = \mathbb{P}(Y = 1|X = x)$.

Ainsi, en reprenant l'expression de la log vraisemblance dans la section précédente et en remplaçant les paramètres théoriques par ceux à appliquer dans le cadre d'une loi Binomiale, la vraisemblance s'écrit :

$$
\begin{aligned}
l_N(\beta, y, X) &= \sum_{i=1}^{N} \frac{y_i\theta_i - b(\theta_i)}{\phi} \\
&= \sum_{i=1}^{N} \frac{y_i ln(\frac{p(x_i)}{1-p(x_i)}) - ln(\frac{1}{1-p(x_i)})}{1} \\
&= \sum_{i=1}^{N} y_i ln(\frac{p(x_i)}{1-p(x_i)}) + \sum_{i=1}^{N} ln(1-p(x_i)) \\
&= \sum_{i=1}^{N} y_i(\beta_0 + \sum_{j=1}^{p} \beta_j x_{j,i}) + \sum_{i=1}^{N} ln(1-p(x_i)) \\
&= \sum_{i=1}^{N} y_i(\beta_0 + \sum_{j=1}^{p} \beta_j x_{j,i}) - \sum_{i=1}^{N} ln(1 + e^{(\beta_0 + \sum_{j=1}^{p} \beta_j x_{j,i})})
\end{aligned}
$$

Pour maximiser cette vraisemblance et déterminer le paramètre optimal β, le moyen naturel est d'annuler son gradient. En effet, s'il existe il est solution de l'équation :

$$
\nabla l_N(\beta, y, X) = (\frac{\partial l_N}{\partial \beta_0}(\beta, y, X), ..., \frac{\partial l_N}{\partial \beta_p}(\beta, y, X)) = 0
$$

Résoudre cette équation revient à résoudre les $p + 1$ équations à $p + 1$ inconnues suivantes :

$$
x_{j,1}y_1 + ... + x_{j,N}y_N = x_{j,1}\frac{\exp(x_1'\beta)}{1+\exp(x_1'\beta)} + ... + x_{j,N}\frac{\exp(x_N'\beta)}{1+\exp(x_N'\beta)} \quad , j = 0, ..., p
$$

Ce système n'étant pas linéaire en β, il n'admet pas de solution explicite. Dans ce cas on a recours à des algorithmes itératifs pour estimer le paramètre optimal, le plus souvent l'algorithme de Newton-Ramphson.

Une fois les paramètres estimés, nous pouvons donc estimer notre variable Y, c'est-à-dire la probabilité conditionnelle $\mathbb{P}(Y = 1|X = x)$. Au moment de l'agrégation du modèle de classification et du modèle de régression, on utilisera directement cette probabilité lors de la modélisation dite « par espérance » (section 1.1.1.3). Pour la modélisation dite par « fréquence/sévérité » (section 1.1.1.2) où on n'utilise pas de probabilité mais une prédiction binaire, on prédira « arbitrage » si $\mathbb{P}(Y = 1|X = x) \geq 0.5$ et « aucun arbitrage » sinon.

3.1.5. Régression Linéaire

Le taux d'arbitrage est une variable centrée en 0, symétrique et comprise entre -1 et 1. Enoncé de cette manière, on pourrait penser que sa distribution appartient à la famille de la loi normale. Néanmoins, on omet ici la particularité que le taux d'arbitrage à une distribution à queues lourdes. Il n'y a donc aucun modèle GLM classique permettant de modéliser une telle distribution. Nous allons tout de même utiliser la loi normale à défaut de mieux.

On se retrouve alors dans le cadre classique de la régression linéaire. On sait alors estimer le paramètre optimal en utilisant le maximum de vraisemblance ou bien la méthode des moindres carrés ordinaires. Les deux méthodes sont ici équivalentes et la solution s'écrit :

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

3.1.6. Importances des variables

Les modèles GLM ont l'avantage d'être facilement interprétables, notamment sur l'importance des variables dans le modèle. On détermine l'importance d'une variable par rapport à la valeur du coefficient estimé.

Le cadre théorique derrière les GLM permet d'affirmer à un certain niveau de confiance que le coefficient est significativement différent de 0, autrement dit que la variable en question a un impact sur la variable cible.

L'interprétation du coefficient estimé diffère en fonction du modèle utilisé. Pour la régression linéaire, la valeur du coefficient correspond à l'augmentation de la variable cible si la variable explicative en question augmente d'une unité. Pour la régression logistique, la valeur du coefficient correspond à la multiplication des chances que l'évènement se produise si la variable explicative en question augmente d'une unité.

3.1.7. Limites

Les modèles linéaires généralisés ont de nombreux avantages. Parmi eux, l'interprétabilités des résultats ainsi que le cadre théorique qu'il y a derrière, permettant d'effectuer des tests statistiques afin d'en évaluer la qualité. C'est pour toutes ces raisons que les GLM sont encore largement utilisés aujourd'hui.

Néanmoins, plusieurs limites subsistent. La plus contraignante pour nous est que les GLM ne sont pas capables (en pratique[8]) de détecter les interactions entre les variables. De plus, malgré la discrétisation de nos variables quantitatives, la prise en compte des effets non linéaires reste très compliquée. Malheureusement ces deux points sont cruciaux dans la modélisation des arbitrages où les relations ne

[8] C'est en théorie possible en intégrant une variable construite en croisant toutes les modalités des variables. Néanmoins, pour un modèle comme le nôtre avec plus de 20 variables explicatives, même en prenant seulement 2 modalités par variables, cela représente 2^{20} = 1 048 576 interactions, soit autant de paramètres à estimer, ce qui n'est pas pensable.

sont pas forcément linéaires (cf la section 1.3.2) et où c'est très souvent un ensemble de conditions/d'évènements qui conduit à arbitrer.

C'est pour toutes ces raisons qu'il s'avère intéressant de se tourner vers d'autres modèles, notamment des modèles d'apprentissage automatique, ne présentant pas ces limites.

3.2. Modèle CART

3.2.1. Généralités sur le modèle CART

Un arbre de décision commence par la racine, d'où découlent plusieurs résultats possibles. Chacun de ces résultats conduit à d'autres nœuds par des branches, ainsi de suite jusqu'à la fin de l'arbre où se trouvent les résultats sur les nœuds finaux, également appelés feuilles.

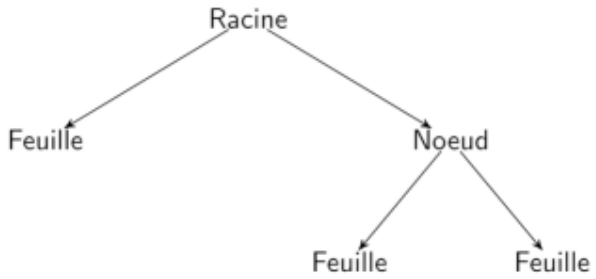


Figure 39 : Arbre de décision

Les arbres de décision CART (Classification And Regression Trees) font partie de la famille des modèles supervisés non paramétriques. Comme le nom l'indique, ils peuvent être utilisés pour de la classification et pour de la régression. Cette méthode fut introduite pour la première fois en 1984, dans la publication portant son nom, *Classification And Regression Trees*, de Leo Breiman.
Ils sont très populaires pour de nombreuses raisons. Premièrement, contrairement aux méthodes classiques, les algorithmes CART ne requièrent aucune hypothèse théorique au préalable (normalité des données, indépendance des variables, …) qui sont rarement vérifiées en pratique. Ce sont également des modèles non linéaires qui permettent de capter des relations autres que linéaires (quadratiques, logarithmiques, …). Ils sont aussi très polyvalents car ils sont utilisables à la fois pour des problèmes de classification et de régression. Pour finir, les résultats peuvent être représentés facilement sous forme d'arbre, ce qui rend simple leur compréhension et leurs utilisations, même pour des non-initiés.
L'objectif consiste donc à expliquer/prédire Y, notre variable cible, à l'aide de p variables explicatives X_1 , … , X_p

Y peut-être catégorielle (binaire ou non), ou continue, de même que pour les variables X_1 , … , X_p

Le principe est alors de trouver une partition A_1 , A_2 ,… de notre espace en formant des groupes homogènes, par rapport à notre variable cible Y, à l'aide d'une série de questions fermées appelées critères de segmentation, basée uniquement sur les valeurs prises par les variables explicatives X_1 , … , X_p

Pour ce faire, on commence par diviser la base de données en deux classes. Cette partition est faite en sélectionnant parmi toutes les variables explicatives, celle dont la division maximise l'homogénéité des classes.

On réitère l'opération sur les deux nouvelles classes obtenues.

L'opération s'arrête soit quand :

- Le nœud est parfaitement homogène, c'est-à-dire qu'au sein des classes, les individus ont tous la même modalité de la variable Y
- Il n'y a plus de division admissible ou le nombre de division déjà effectué est supérieur à la valeur du seuil fixé.
- Le nombre d'observations présentes dans ce nœud est inférieur à la valeur du seuil fixé.

C'est pendant la phase d'entraînement que l'on recherche les questions offrant les partitions les moins hétérogènes possible.

3.2.2. Fonctions d'hétérogénéité

Pour mesurer l'hétérogénéité, il est nécessaire de définir une fonction. Cette fonction doit être :

- Positive
- Egale à 0 si, et seulement si, les individus au sein du sous ensemble ont tous la même modalité de Y
- Maximale si, et seulement si, les modalités de Y sont présentes de manière équiprobable au sein du sous ensemble

La division du nœud K crée deux fils, gauche et droit, que l'on note K_G et K_D. L'algorithme retient la division qui minimise la somme $D_{K_G} + D_{K_D}$ des hétérogénéités des nœuds fils.

Autrement dit, on cherche à résoudre à chaque étape de construction de l'arbre :

$$\max_{\{divisions\ de\ X_j\ :\ j=1,\dots,p\}} D_K - (D_{K_G} + D_{K_D})$$

Pour une régression, la fonction d'hétérogénéité d'un nœud K est définie par :

$$D_K = \frac{1}{|K|}\sum_{i \in K}(y_i - \overline{y}_K)^2$$

Où $|K|$ est le nombre d'observations dans le nœud K.

Le but est de chercher pour chaque nœud la règle de division qui permettra de rendre les deux nœuds fils le moins hétérogène possible. Cela revient à minimiser la variance intraclasse, c'est-à-dire :

$$\frac{|K_G|}{n}\sum_{i \in K_G}(y_i - \overline{y}_{K_G})^2 + \frac{|K_D|}{n}\sum_{i \in K_D}(y_i - \overline{y}_{K_D})^2$$

Pour une classification, les fonctions d'hétérogénéité les plus courantes sont :

- L'entropie :

$$D_K = -2\sum_{l=1}^{m} |K| p_K^l \log(p_K^l)$$

Où p_K^l représente la proportion de la classe T_l de Y dans le nœud K.

- L'indice de Gini :

$$D_K = \sum_{l=1}^{m} p_K^l (1 - p_K^l)$$

Il en existe d'autres, mais nous nous limiterons à ces deux-là qui sont les seules à être présentes dans *sklearn*.

De même que pour la régression, le but est de chercher la règle de division qui permettra de minimiser la fonction choisie, qui sera alors simplement l'erreur quadratique (ou absolue) moyenne.

3.2.3. Prédictions

Une fois l'algorithme arrêté, la valeur prédite pour Y catégorielle sera soit :

- La classe majoritaire sur la feuille
- La classe la moins coûteuse si une fonction de coût de mauvais classement a été définie au préalable.

Dans la modélisation dite « par l'espérance » (section 1.1.1.3), on utilise une estimation de la probabilité d'appartenance à la classe positive (« arbitrage ») et non pas une estimation catégorielle de Y. Dans ce cas, on estimera cette probabilité par la fréquence observée, c'est-à-dire en divisant le nombre de personne sur la feuille de la classe positive par le nombre total de personne sur la feuille. Ci-dessous un exemple illustratif.

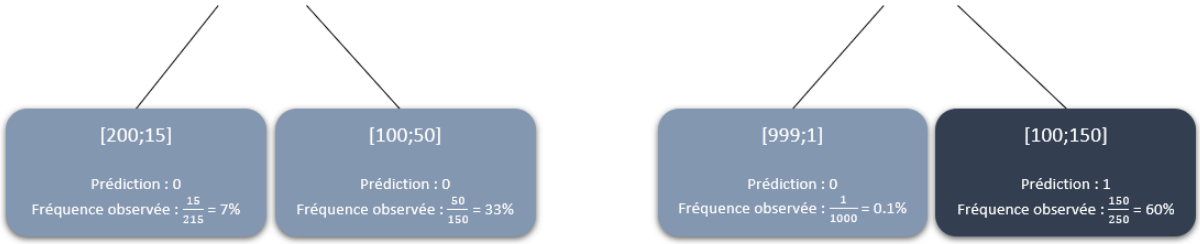


Figure 40 : Calcul de la probabilité estimée avec un arbre de décision

Pour Y quantitative, la valeur prédite sera simplement la moyenne de Y sur chaque feuille.

3.2.4. Construction du modèle CART

La construction de l'arbre CART se fait en deux étapes. On commence par déterminer l'arbre maximal (ou saturé) sur la base d'apprentissage, puis on élague pour obtenir l'arbre optimal.

- Arbre maximal

L'arbre maximal A_{max} que nous souhaitons créer ici, est obtenu en répétant les opérations vues précédemment, jusqu'à ce que chaque feuille soit composée uniquement d'individus ayant la même modalité de Y. L'apprentissage de cet arbre est parfait, mais il est inutilisable tel quel. En effet, il est bien trop sensible au sur-apprentissage.

Il faut donc déterminer une ou plusieurs conditions d'arrêt, pour que la construction de l'arbre s'arrête avant cette étape.

- Elagage

L'élagage consiste à fixer des critères d'arrêts. Le but est de garder un sous arbre de l'arbre maximal A_{max} en supprimant des branches qui ne sont pas représentatives pour ne garder que celle ayant de bonnes performances prédictives et qui permettent la généralisation du modèle.

Ils existent plusieurs critères d'arrêts, par exemple :

- La profondeur maximale de l'arbre
- Le nombre minimum d'observations sur un nœud pour le séparer

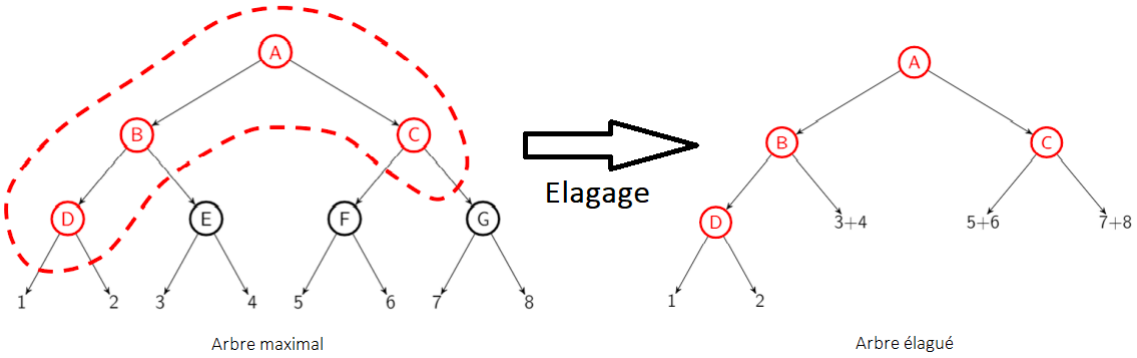


Figure 41 : Illustration élagage d'un arbre de décision

3.2.5. Importances des variables

Pour pouvoir interpréter nos résultats, il est intéressant d'avoir une mesure de l'importance de nos variables. La mesure utilisée par *sklearn* dans la sortie « .feature_importances_ » est la suivante.

On commence par calculer la qualité globale de l'arbre, notée Δ_{arbre} , par la différence entre le critère d'hétérogénéité choisi du nœud racine et la moyenne pondérée de ce critère sur les L feuilles, chacunes d'effectif n_l. Soit H la fonction d'hétérogénéité choisie, on a :

$$\Delta_{arbre} = H(racine) - \sum_{l=1}^{L} \frac{n_l}{n} H(l)$$

Puis on calcul la contribution de la segmentation d'une variable i qui s'effectue à un nœud k

$$\Delta_{segmentation_i} = \frac{n_k}{n}[H(racine) - (\frac{n_{k_G}}{n_k} \times H(n_{k_G}) + \frac{n_{k_D}}{n_k} \times H(n_{k_D})]$$

Avec k_G et k_D les deux nœuds fils gauche et droit du nœud k.

On effectue ensuite le rapport entre ces deux quantités pour que la somme des contributions fasse 1.

$$Contribution_i = \frac{\Delta_{segmentation_i}}{\Delta_{arbre}}$$

A noter que si une variable apparaît dans plusieurs segmentations, on calcul alors la somme des contributions de ses segmentations.

3.2.6. Limites

Les algorithmes CART ne sont néanmoins pas parfaits. Ils possèdent plusieurs limites :

- Le sur-apprentissage : Bien que l'élagage permette de réduire ce risque, ça ne suffit pas. Ces algorithmes gardent une fâcheuse tendance à sur apprendre les données d'entraînements, ce qui est typique des algorithmes qui parcourt de nombreuses fois les mêmes données.

- L'instabilité : Les critères de segmentation sont fixés à partir des données d'entraînements. Si ces données d'entraînements changent légèrement, cela peut changer certains critères de segmentation et mener à une répartition potentiellement très différentes.

- Importances des variables explicatives : Les arbres CART ont tendance à surestimer l'importance des variables les plus discriminantes ce qui a pour effet de cacher l'effet des variables explicatives légèrement moins discriminantes.

Les méthodes d'agrégation d'arbres permettent d'atténuer ces risques.

3.3. Agrégation de modèles

Les algorithmes qui utilisent les agrégations de modèles sont basés sur des stratégies adaptatives (boosting, gradient boosting) ou aléatoires (bagging, random forest). Ces agrégations de modèles simples et/ou instables permettent d'améliorer l'ajustement tout en réduisant le sur apprentissage. Ces méthodes sont donc parfaitement adaptées pour les arbres de décisions.

3.3.1. Bagging

À la suite du fort succès du modèle CART introduit en 1984, Leo Brieman introduit le Bagging en 1996 (qui fut l'une des premières méthodes d'ensemble[9]).

[9] Méthodes ayant pour but de créer plusieurs modèles légèrement différents à partir d'une seule base initiale pour ensuite les regrouper afin de diminuer la variance et d'améliorer les prédictions.

Le bagging (contraction de boostrap aggregating), consiste à construire B échantillons de notre base d'entraînement de taille q en utilisant le boostrap (échantillonnage uniforme avec remise), puis construire un modèle (arbre de décisions dans notre cas) sur chaque échantillon boostrap. On agrège ensuite les résultats. SI Y est catégorielle, on prend la classe majoritaire. Si Y est quantitative, on prend alors la moyenne des prédictions.

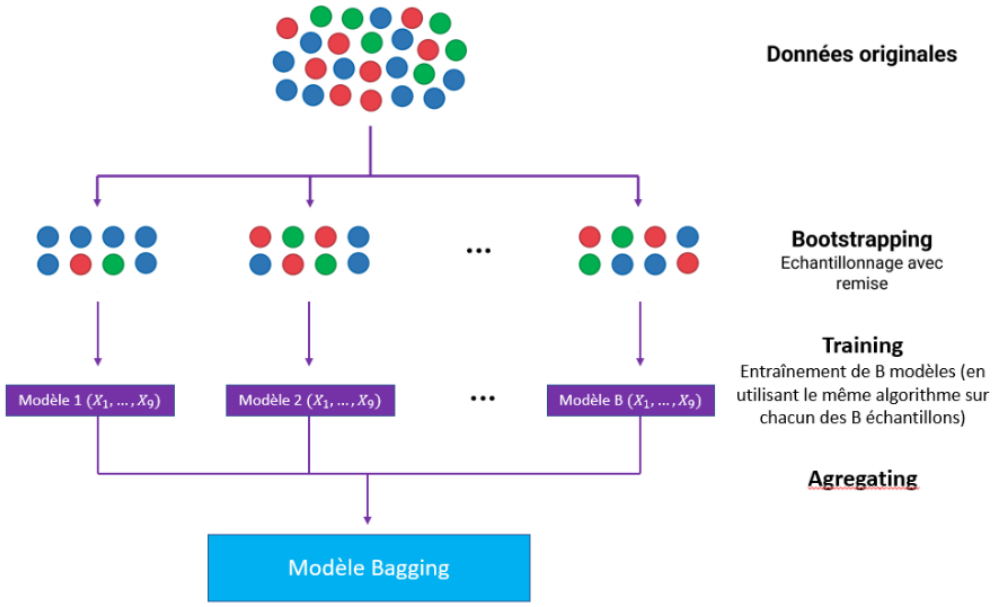


Figure 42 : Illustration des différentes étapes dans un modèle de Bagging

En prenant les modèles CART pour construire le bagging, la prédiction de Y pour l'individu i est donc :

- Pour Y quantitative :

$$\hat{Y}_{bag}(X_{1,i}, \dots, X_{p,i}) = \frac{1}{B}\sum_{b=1}^{B} A_b(X_{1,i}, \dots, X_{p,i})$$

- Pour Y qualitative :

$$\hat{Y}_{bag}(X_{1,i}, \dots, X_{p,i}) = \underset{\mathbf{k}\subset\{Modalités\ de\ Y\}}{\operatorname{argmax}} \sum_{b=1}^{B} A_b(X_{1,i}, \dots, X_{p,i}) = \mathbf{k}$$

Avec $A_b\ X_{1,i}, \dots, X_{p,i}$ la prédiction de l'arbre construit sur l'échantillon boostrap b pour l'individu i. Pour alléger les notations, on pose pour la suite :

$$\mathbb{X}_i = X_{1,i}, \dots, X_{p,i}$$

Le vecteur des variables explicatives pour l'individu i.

Comme on l'a dit, le bagging permet de réduire la variance des prédictions (instabilité) et d'être plus précis. Ce résultat se retrouve mathématiquement. Par exemple dans le cas de la régression et de l'erreur quadratique, l'erreur commise par le premier arbre $A_{1,i}$ pour l'individu i est :

$$\begin{aligned}\mathbb{E}[(A_1(\mathbb{X}_i) - Y_i)^2] &= \mathbb{E}[A_1(\mathbb{X}_i)^2 - 2Y_i A_1(\mathbb{X}_i) + Y_i^2] \\ &= \mathbb{E}[A_1(\mathbb{X}_i)^2] - 2Y_i\mathbb{E}[A_1(\mathbb{X}_i)] + Y_i^2 \\ &= \mathbb{E}[A_1(\mathbb{X}_i)^2] - \mathbb{E}[A_1(\mathbb{X}_i)]^2 + \mathbb{E}[A_1(\mathbb{X}_i)]^2 - 2Y_i\mathbb{E}[A_1(\mathbb{X}_i)] + Y_i^2 \\ &= \mathbb{E}[A_1(\mathbb{X}_i)^2] - \mathbb{E}[A_1(\mathbb{X}_i)]^2 + \mathbb{E}[A_1(\mathbb{X}_i) - Y_i]^2 \\ &= \mathrm{Var}(A_1(\mathbb{X}_i)) + \mathrm{Biais}(A_1(\mathbb{X}_i))^2\end{aligned}$$

De même, l'erreur commise par le prédicteur agrégé est :

$$\mathbb{E}[(\hat{Y}_{bag}(\mathbb{X}_i) - Y_i)^2] = \mathrm{Var}(\hat{Y}_{bag}(\mathbb{X}_i)) + \mathrm{Biais}(\hat{Y}_{bag}(\mathbb{X}_i))^2$$

Or, en supposant les modèles CART de mêmes lois et tels que la corrélation entre deux modèles CART soit égale à ρ, on a :

$$\begin{aligned}\mathrm{Var}(\hat{Y}_{bag}(\mathbb{X}_i)) &= \mathrm{Var}(\frac{1}{B}\sum_{b=1}^{B} A_b(\mathbb{X}_i)) \\ &= \frac{1}{B^2}[\sum_{b=1}^{B}\mathrm{Var}(A_b(\mathbb{X}_i)) + \sum_{l\neq k}^{B}\mathrm{Cov}(A_b(\mathbb{X}_l), A_b(\mathbb{X}_k))] \\ &= \frac{1}{B^2}[B\mathrm{Var}(A_b(\mathbb{X}_1)) + B(B-1)\mathrm{Var}(A_b(\mathbb{X}_1))\rho] \\ &= \frac{1}{B}[\mathrm{Var}(A_b(\mathbb{X}_i) + B\mathrm{Var}(A_b(\mathbb{X}_1))\rho - \mathrm{Var}(A_b(\mathbb{X}_1))\rho] \\ &= \mathrm{Var}(A_b(\mathbb{X}_1))\rho + \frac{(1-\rho)\mathrm{Var}(A_b(\mathbb{X}_1))}{B}\end{aligned}$$

Et :

$$\begin{aligned}\mathrm{Biais}(\hat{Y}_{bag}(\mathbb{X}_i))^2 &= (\mathbb{E}[\hat{Y}_{bag}(\mathbb{X}_i)] - Y_i)^2 \\ &= (\frac{1}{B}\mathbb{E}[\sum_{b=1}^{B} A_b(\mathbb{X}_i))] - Y_i)^2 \\ &= (\mathbb{E}[A_b(\mathbb{X}_1))] - Y_i)^2 \\ &= \mathrm{Biais}(A_1(\mathbb{X}_i))^2\end{aligned}$$

Ce qui nous donne finalement :

$$\mathbb{E}[(\hat{Y}_{bag}(\mathbb{X}_i) - Y_i)^2] = \mathrm{Var}(A_b(\mathbb{X}_1))\rho + \frac{(1-\rho)\mathrm{Var}(A_b(\mathbb{X}_1))}{B} + \mathrm{Biais}(A_1(\mathbb{X}_i))^2$$

Cette égalité nous permet de conclure plusieurs choses. Premièrement, plus le nombre de modèles agrégés B est important, plus la variance (et donc l'erreur) du modèle Bagging est faible. Deuxièmement, l'erreur dépend positivement de ρ, autrement dit, plus les modèles agrégés sont corrélés entre eux, plus la variance (et donc l'erreur) est grande.

En effet :

$$\frac{\partial \mathbb{E}[(\hat{Y}_{bag}(\mathbb{X}_i) - Y_i)^2]}{\partial \rho} = \mathrm{Var}(\Lambda_b(\mathbb{X}_1)) - \frac{\mathrm{Var}(\Lambda_b(\mathbb{X}_1))}{B} = \mathrm{Var}(\Lambda_b(\mathbb{X}_1))(1 - \frac{1}{B}) > 0$$

A noter que si $\rho = 1$ ou bien si $B = 1$, on se retrouve dans le cas de l'arbre simple, ce qui paraît cohérent.

Pour finir, on peut également montrer que l'erreur de ce modèle est strictement plus faible que l'erreur de l'arbre simple dès lors que le nombre de modèles agrégés B est strictement plus grand qu'un et qu'ils ne sont pas entièrement corrélés ($\rho < 1$).

$$\begin{aligned}
\rho < 1 &\Leftrightarrow \rho(B-1) < B - 1 \\
&\Leftrightarrow \rho B - \rho + 1 < B \\
&\Leftrightarrow \rho + \frac{1-\rho}{B} < 1 \\
&\Leftrightarrow \mathrm{Var}(\Lambda_b(\mathbb{X}_1))\rho + \frac{\mathrm{Var}(\Lambda_b(\mathbb{X}_1))(1-\rho)}{B} < \mathrm{Var}(\Lambda_b(\mathbb{X}_1)) \\
&\Leftrightarrow \mathrm{Var}(\Lambda_b(\mathbb{X}_1))\rho + \frac{\mathrm{Var}(\Lambda_b(\mathbb{X}_1))(1-\rho)}{B} + \mathrm{Biais}(\Lambda_1(\mathbb{X}_i))^2 < \mathrm{Var}(\Lambda_b(\mathbb{X}_1)) + \mathrm{Biais}(\Lambda_1(\mathbb{X}_i))^2 \\
&\Leftrightarrow \mathbb{E}[(\hat{Y}_{bag}(\mathbb{X}_i) - Y_i)^2] < \mathbb{E}[(\Lambda_1(\mathbb{X}_i) - Y_i)^2]
\end{aligned}$$

C'est donc pour ces raisons que le modèle Bagging est plus intéressant qu'un arbre simple.

Néanmoins, comme nous l'avons vu, si les arbres sont trop corrélés entre eux, les effets positifs du bagging sont atténués.

Or, comme nous avons vu, un des défauts du modèle CART est qu'il a tendance à sur estimer l'importance de certaines variables explicatives et à ne pas capter l'effet de variables moins discriminantes. Cet effet couplé au fait que pour chaque modèle, toutes les variables explicatives sont utilisées, conduit très souvent à construire des arbres trop corrélés.

3.3.2. Forêt aléatoire

Les forêts aléatoires, introduite en 2001, également par Leo Breiman, font aussi parties de la famille des méthodes d'ensemble. Elles sont très similaires au modèle de bagging, à la seule différence que les variables utilisées dans la construction des modèles à agréger diffèrent.

En effet, on procède au début de la même manière. On construit B échantillons de notre base d'entraînement de taille q en utilisant le boostrap. On va ensuite construire l'arbre maximal pour chaque échantillon, mais en utilisant que m variables, $m \leq p$, que l'on aura choisie via un tirage sans remise parmi les p variables. On agrège ensuite les résultats de la même manière que pour le bagging. Si Y est catégorielle, on prend la classe majoritaire. Si Y est quantitative, on prend alors la moyenne des prédictions.

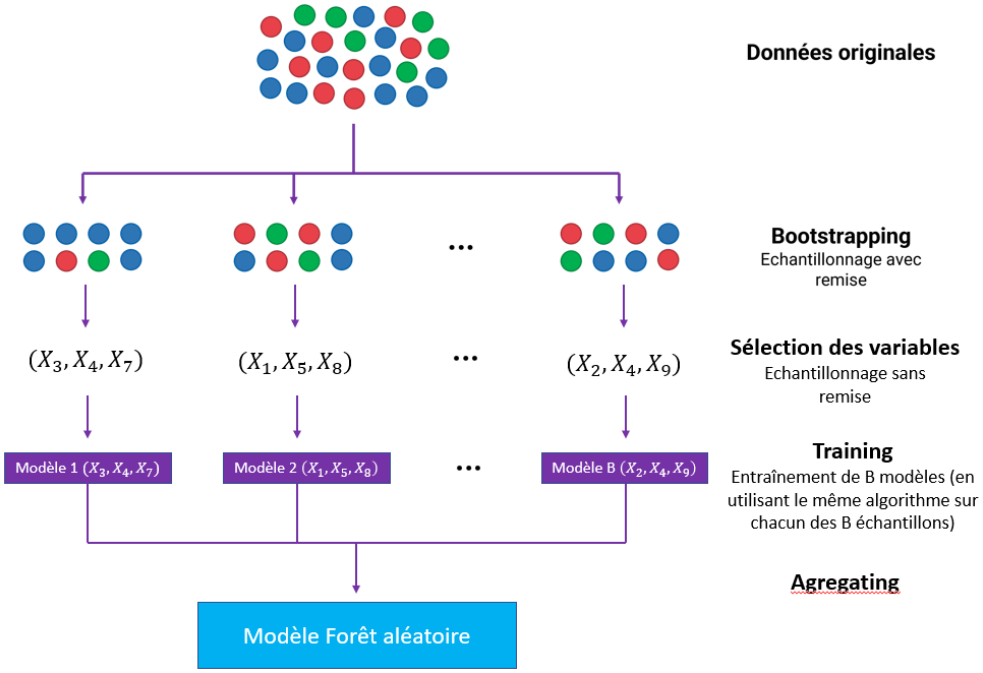


Figure 43 : Illustration des différentes étapes dans un modèle de forêt aléatoire

En procédant ainsi, nous obtenons des arbres moins corrélés entre eux que les arbres provenant du bagging. On fait ainsi baisser la variance de notre prédicteur et on améliore donc nos prédictions.

PARTIE 3 : RESULTATS

1. Implémentation des modèles

Maintenant que nous avons présenté différentes façons de modéliser ainsi plusieurs modèles, regardons les résultats obtenus sur nos données.

1.1. Modèles de classification

Pour rappel, nous avons entrainé les modèles sur la période 2015-2019 que nous testons ensuite sur 2020 et 2021. Ci-contre quelques chiffres clés pour mieux apprécier la qualité de prédictions des modèles.

Tableau 12 : Information sur les données test (1)

Données de test	
Séparation	**2015-2019 / 2020-2021**
Taille de l'échantillon test	248 239
Aucun arbitrage	233 217
proportion	93,9%
Arbitrage	15 022
proportion	6,1%

1.1.1. Modèle GLM de classification

Pour optimiser au maximum les performances de notre régression logistique, nous avons ajusté par cross validation les deux paramètres suivants :

- La pénalité appliquée (Lasso, Ridge ou aucune)
- La force de pénalisation λ

Le modèle retenu utilise la pénalité de Ridge avec $\lambda = 0.1$.

Ci-dessous se trouve la sortie du modèle :

Tableau 13 : Importances des variables dans le GLM fréquence (simple)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.62402	0.04090	-88.604	< 2e-16***
nb_arbitrage_cumul 1.[1 : ...[	0.65062	0.01931	33.698	< 2e-16***
nb_arbitrage_nmoins1 1.[1 : ...[	0.82560	0.01267	65.151	< 2e-16***
classe_age 2.[70 : ...[	-0.17252	0.01373	-12.565	< 2e-16***
classe_anciennete 2.[5 : 11[	-0.24628	0.01402	-17.571	< 2e-16***
classe_anciennete 3.[11 : ...[	-0.40421	0.01541	-26.232	< 2e-16***
classe_ecart_cac 2.[-5% : ...[	0.09626	0.02036	4.728	2.27e-06***
classe_ecart_oat 2.[1.3% : ...[	-0.08815	0.02178	-4.047	5.20e-05***
classe_ecart_pmv_uc_eu 2.[-2% : 1%[	-0.17243	0.01917	-8.995	< 2e-16***
classe_ecart_pmv_uc_eu 3.[1% : ...[	-0.12163	0.03415	-3.561	0.000369***
classe_moyenne_cac 2.[4500 : 5200[	-0.15143	0.02316	-6.539	6.20e-11***
classe_moyenne_cac 3.[5200 : ...[	-0.15662	0.03334	-4.698	2.62e-06***
classe_nb_supports 2.[400 : 1200[	0.33094	0.01552	21.328	< 2e-16***
classe_nb_supports 3.[1200 : ...[	0.29858	0.01632	18.294	< 2e-16***
classe_part_uc_pmv 2.[10% : ...[	0.02731	0.03264	0.837	0.402819
classe_pm 2.[50K : 300K[	0.42002	0.01390	30.214	< 2e-16***
classe_pm 3.[300K : ...[	0.70117	0.01746	40.164	< 2e-16***
classe_pmv_eu_prop 2.]1.5% : 2.5%[	0.13417	0.01656	8.103	5.36e-16***
classe_pmv_eu_prop 3.[2.5% : ...[	0.06725	0.03199	2.102	0.035520*
classe_pmv_prop 2.]1% : 3%[	-0.51824	0.02115	-24.507	< 2e-16***
classe_pmv_prop 3.[3% : ...[	-0.69596	0.03192	-21.803	< 2e-16***
classe_pmv_uc_prop 2.]0.5% : 2%[	0.30735	0.03683	8.346	< 2e-16***
classe_pmv_uc_prop 3.[2% : ...[	0.65828	0.04258	15.460	< 2e-16***
classe_prop_uc 2.[10% : 50%[	0.69361	0.02011	34.494	< 2e-16***
classe_prop_uc 3.[50% : ...[	0.45413	0.02182	20.809	< 2e-16***
classe_taux_pb 2.[1.5% : 2.5%[	-0.05998	0.02384	-2.516	0.011880*
classe_taux_pb 3.[2.5% : ...[	0.10930	0.02909	3.758	0.000171***
classe_tx_rdmt_cac 2.[0% : 10%[	0.00644	0.02192	0.294	0.768954
classe_tx_rdmt_cac 3.[10% : ...[	NA	NA	NA	NA

Comme nous pouvons le constater, exceptés le taux de rendement du cac40 ainsi que la proportion de la plus-value générée par l'UC (l'année précédente), toutes les variables sont significatives. Un assuré ayant toutes les caractéristiques de référence a une probabilité d'effectuer un arbitrage dans l'année qui vient de exp(-3.62402), soit environ 2.67%, selon notre modèle.

Aux vues des résultats, la variable qui augmente le plus la probabilité d'effectuer un arbitrage dans l'année à venir est la variable *nb_arbitrage_nmoins1*. Toutes choses égales par ailleurs, un individu ayant effectué au moins un arbitrage l'année passée (classe [1, …[) a une probabilité d'arbitrer multipliée par exp(0.82560), soit 2.28 par rapport à la classe de référence, c'est-à-dire un individu n'ayant pas arbitré l'année passée. L'interprétation se fait de la même manière pour les autres variables. Ainsi, on note que les variables ayant le plus gros impact sur la probabilité d'arbitrer sont les variables liées aux comportements d'arbitrages passés : la provision mathématique, le taux de plus ou moins-value l'année précédente et enfin la proportion d'UC le dernier jour de l'année passée.

Pour le modèle rééchantillonné, les interprétations sont les mêmes, voici les résultats des prédictions des deux modèles selon les métriques énoncées précédemment.

Tableau 14 : Résultats des GLM fréquence

Méthode	GLM	GLM 80/20
Hyperpamètres optimaux	Pénalité : Ridge Lambda : 0.1	Pénalité : Ridge Lambda : 1
Nombre de prédictions : "Aucun" correctes	231166 99,12%	227426 97,52%
Nombre de prédictions : "Arbitrage" correctes (Rappel)	1633 10,87%	4057 27,01%
Nombre de prédictions : "Arbitrage" (Précision)	3677 44,41%	9861 41,14%
Log Loss	**0,221**	**0,304**
AUC Precision-Recall	**0,26**	**0,29**

L'aire sous la courbe *precision-recall* est plus élevée sur le modèle rééchantillonné. C'est donc le modèle à privilégier si on utilise les prédictions brutes. En revanche, la *Log Loss* étant plus faible sur le modèle classique, c'est le modèle à privilégier si on utilise la probabilité d'effectuer un arbitrage.

1.1.2. Modèle CART de classification

Le modèle CART du module *scikit-learn* de Python possède une dizaine d'hyperparamètres à ajuster. Pour des raisons de temps de calcul, nous avons ajusté uniquement les 3 principaux qui sont :

- La fonction d'hétérogénéité (limité à l'entropie et à l'indice de Gini sur *scikit-learn*)
- La profondeur maximale de l'arbre
- Le nombre de données minimum sur une feuille

Après avoir procédé par cross validation comme décrit dans la partie 1, l'arbre retenu a une profondeur maximale de 8 et un minimum de 190 données sur chaque feuille. La fonction d'hétérogénéité retenue est l'indice de Gini.

Ci-dessous le modèle écrit en pseudo-code.

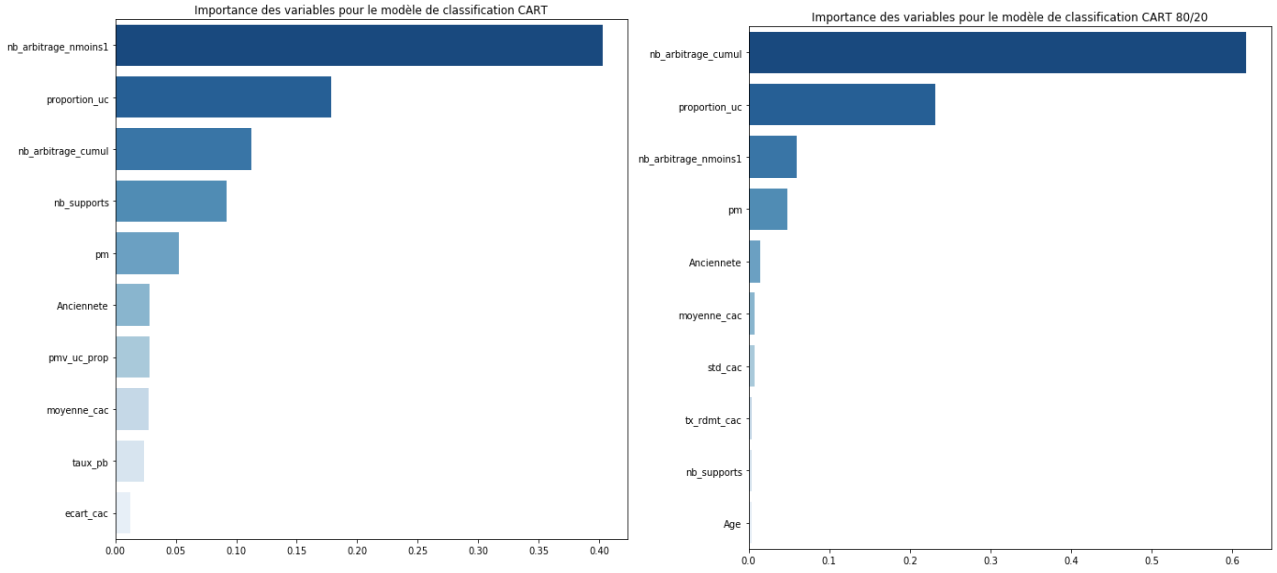
Algorithme : Classification And Regression Tree (CART)

Initialisation : Nœud = base d'apprentissage
Pour tout Noeud
    Si Profondeur est inférieure à 8 et que division homogène possible **faire**
        Déterminer toutes les partitions admissibles pour les p variables prédictives
        Calculer l'indice de Gini sur les différentes divisions
        S'il existe une division qui assure un nombre d'observations par nœud fils supérieur à 190
            Sélectionner la division avec l'indice le plus élevé
            Diviser le nœud en deux nœud fils
        Sinon
            Nœud = Nœud terminal (feuille)
            Attribuer au nœud la classe majoritaire et déterminer les probabilités associées à la population de cette classe
        Fin si
    Sinon
        Nœud = Nœud terminal (feuille)
        Attribuer au nœud la classe majoritaire et déterminer les probabilités associées à la population de cette classe

Fin si

Pour le modèle rééchantillonné, l'arbre retenu est légèrement plus simple (cf tableau des résultats). Dans tous les cas, il est clair que les variables les plus discriminantes sont de loin celles liées aux comportements d'arbitrages antérieurs. Cela concorde avec les graphiques de la section 1.3.2. Toutes choses égales par ailleurs, plus une personne a arbitré, plus elle a de chances d'arbitrer de nouveau.
En moindre mesure, il semble que la PM ainsi que la proportion d'UC dans le contrat ont un impact sur le fait qu'un assuré arbitre ou non. On retrouve également ces résultats graphiquement.

Tableau 15 : Importance des variables dans les modèles de classification CART



Ci-dessous l'évaluation des performances prédictives de l'arbre CART.

Tableau 16 : Résultats des modèles de classification CART

Méthode	**CART**	**CART 80/20**
Hyperpamètres optimaux	Profondeure maximale : 8 Fonction d'hétérogénéité : Indice de Gini Nombre d'observation minimale sur une feuille : 190	Profondeure maximale : 5 Fonction d'hétérogénéité : Indice de Gini Nombre d'observation minimale sur une feuille : 330
Nombre de prédictions : "Aucun" correctes	231 590 99,30%	220 418 94,51%
Nombre de prédictions : "Arbitrage" correctes (Rappel)	2 392 15,92%	6 390 42,54%
Nombre de prédictions : "Arbitrage" (Précision)	4 041 59,19%	19 189 33,30%
Log Loss	**0,183**	**0,265**
AUC Precision-Recall	**0,32**	**0,35**

Sans surprise le modèle rééchantillonné permet d'augmenter le nombre de prédictions d'arbitrages correctes, mais ce au détriment de la précision. L'aire sous la courbe PR étant plus élevée, on privilégiera ce modèle si on utilise les prédictions brutes. Cependant on privilégiera une nouvelle fois le modèle simple si on utilise les probabilités estimées d'appartenance à une classe donnée car la *Log Loss* de ce modèle est plus faible.

On note néanmoins une nette amélioration des prédictions par rapport à la régression logistique classique.

1.1.3. Modèle Bagging de classification

Comme présenté dans la section 2.2.3.2, le modèle Bagging est un modèle d'agrégation d'arbres. En théorie, il est censé offrir de meilleures performances que celles d'un arbre simple.

Le module *scikit-learn* de Python possède également une dizaine d'hyperparamètres à ajuster mais nous nous sommes une nouvelle fois restreint à 5.

- Le nombre d'arbres à agréger.
- La proportion de données à boostraper pour construire les arbres.
- La fonction d'hétérogénéité (limitée à l'entropie et à l'indice de Gini sur *scikit-learn*)
- La profondeur maximale des arbres
- Le nombre de données minimum sur une feuille

Le modèle Bagging retenu est un modèle où 50 arbres sont agrégés en utilisant à chaque fois 75% des données (déterminés par bootstrap). Les arbres sont construits en utilisant l'entropie comme fonction d'hétérogénéité et ne doivent pas avoir une profondeur supérieure à 8. Le nombre de données minimum sur une feuille a été fixé à 400.

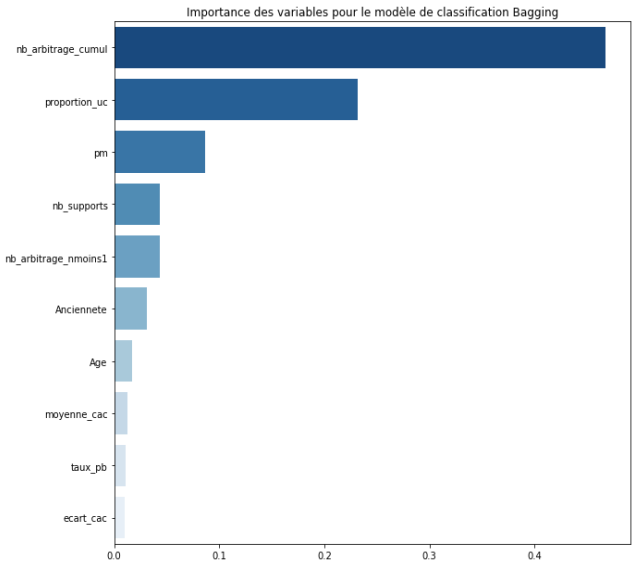
Ci-dessous le modèle, écrit en pseudo-code.

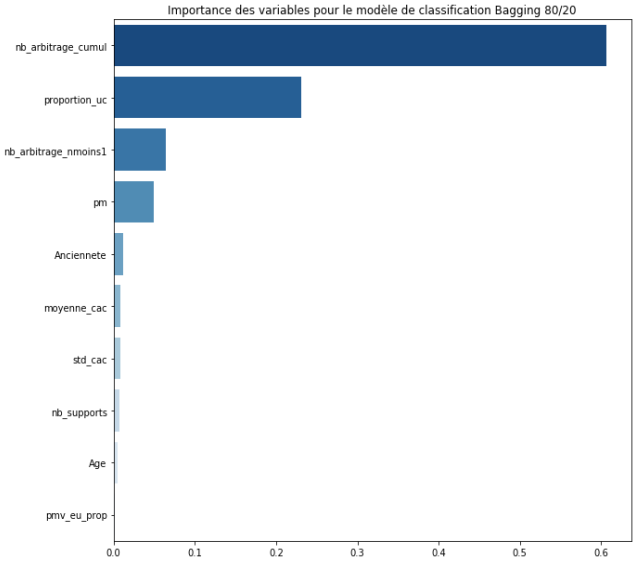
Algorithme : Bagging

Pour *i* allant de 1 à 50 faire
    Tirer aléatoirement avec remise un 75% de la base initiale
    Construire sur cet échantillon un arbre A_b en utilisant **Algorithme** CART avec profondeur maximale de 8, 400 données minimum par feuille et l'entropie
    $i = i+1$
Fin pour
Restituer le modèle agrégé

Pour le modèle rééchantillonné, il est une nouvelle fois plus simple (cf tableau des résultats).
Ce sont logiquement les mêmes variables qui ressortent comme étant discriminantes. Aucune différence notable avec les arbres à ce niveau-là.

Tableau 17 : Importance des variables dans les modèles de classification Bagging





Ci-dessous l'évaluation des performances prédictives du modèle bagging.

Tableau 18 : : Résultats des modèles de classification Bagging

Méthode	**Bagging**	**Bagging 80/20**
Hyperpamètres optimaux	Nombre d'arbre : 50 Taille des échantillions boostrap : 75% Profondeure maximale : 8 Fonction d'hétérogénéité : Entropie Nombre d'observation minimale sur une feuille : 400	Nombre d'arbre : 50 Taille des échantillions boostrap : 75% Profondeure maximale : 6 Fonction d'hétérogénéité : Entropie Nombre d'observation minimale sur une feuille : 300
Nombre de prédictions : "Aucun" correctes	231 470	221 284
	99,25%	94,88%
Nombre de prédictions : "Arbitrage" correctes	2 582	6 176
(Rappel)	17,19%	41,11%
Nombre de prédictions : "Arbitrage"	4 329	17 723
(Précision)	59,64%	34,85%
Log Loss	**0,178**	**0,261**
AUC Precision-Recall	**0,33**	**0,35**

Les performances des prédictions obtenues par le modèle Bagging sont légèrement décevantes. Elles sont certes meilleures que celles de l'arbre simple, mais l'amélioration est minime.

1.1.4. Modèle Forêt aléatoires de classification

Notre dernier modèle de classification est obtenu en utilisant les forêts aléatoires. Comme vu précédemment, elles sont censées théoriquement apporter les meilleurs résultats parmi les modèles de machine learning que nous avons présentés.

Une nouvelle fois nous nous sommes restreints à l'optimisation de 5 hyperparamètres parmi la dizaine proposée.

- Le nombre d'arbres à agréger.
- Le nombre de variables à boostraper pour construire les arbres.
- La fonction d'hétérogénéité (limitée à l'entropie et à l'indice de Gini sur *scikit-learn*)
- La profondeur maximale des arbres

- Le nombre de données minimum sur une feuille

Algorithme : Forêt aléatoire

Pour i allant de 1 à 75 faire

Tirer aléatoirement avec remise un 75% de la base initiale.

Tirer aléatoirement avec remise 50% des variables explicatives.

Construire avec ces variables et sur cet échantillon un arbre A_b en utilisant **Algorithme** CART avec profondeur maximale de 8, 100 données minimum par feuille et l'indice de Gini.

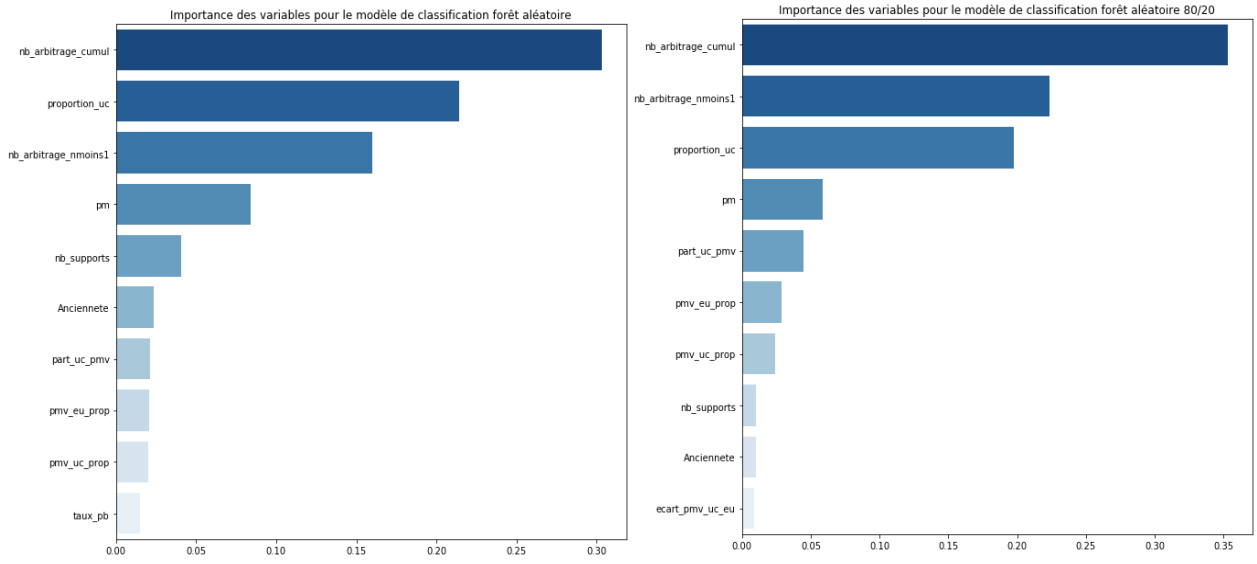
$i = i+1$

Fin pour

Restituer le modèle agrégé

C'est encore une fois les mêmes variables qui ressortent comme étant les plus importantes. On remarque néanmoins que le pourcentage de contribution des variables principales est plus équilibré. Ce résultat est souvent observé lors de l'implémentation de forêts aléatoires. Il est inhérent à la construction des forêts qui implémente plusieurs arbres construits sur un sous ensemble de variables. Les résultats obtenus sont donc en théorie plus robustes car moins sensibles à la valeur d'une seule variable très discriminante par exemple.

Tableau 19 : Importance des variables dans les modèles de classification Random Forest



Ci-dessous l'évaluation des performances prédictives du modèle de forêt aléatoire.

Tableau 20 : Résultats des modèles de classification Random Forest

Méthode	**Random Forest**	**Random Forest 80/20**
Hyperpamètres optimaux	Nombre d'arbre : 75 Nombre de variable : p/2 Profondeure maximale : 8 Fonction d'hétérogénéité : Indice de Gini Nombre d'observation minimale sur une feuille : 100	Nombre d'arbre : 50 Nombre de variable : p/2 Profondeure maximale : 5 Fonction d'hétérogénéité : Entropie Nombre d'observation minimale sur une feuille : 300
Nombre de prédictions : "Aucun" correctes	232 992	225 038
	99,90%	96,49%
Nombre de prédictions : "Arbitrage" correctes	1 142	5 299
(Rappel)	7,60%	35,27%
Nombre de prédictions : "Arbitrage"	1 367	13 581
(Précision)	83,54%	39,02%
Log Loss	**0,171**	**0,272**
AUC Precision-Recall	**0,36**	**0,38**

Les résultats obtenus par les prédictions émises du modèle de forêt aléatoire sont en effet les meilleures parmi tous les modèles essayés jusqu'à présent.

1.1.5. Conclusion et interprétations

Tous les modèles de classification implémentés ce sont avérés être très proche dans leurs interprétations. Les comportements d'arbitrage passés, la proportion d'UC dans le contrat, la PM ainsi que le nombre de supports disponibles pour investir sur l'UC sont les variables qui sont les plus importantes, tous modèles confondus. On peut noter dans ces variables l'absence de variables exogènes, liées au contexte économique.

Nous pensons que la principale raison à cela est la lourde procédure nécessaire pour pouvoir effectuer un arbitrage. En effet, comme expliqué dans la section 1.1.3, pour pouvoir effectuer un arbitrage l'assuré a l'obligation de prendre rendez-vous avec son conseiller, pour ensuite remplir un document papier qui sera soumis à validation (cf. formulaire et procédure d'arbitrage en annexe). Il peut parfois s'écouler plusieurs semaines entre ces deux moments. La procédure est donc très dissuasive aux comportements réactionnaires des marchés financiers. Il existe néanmoins diverses options (stop loss, sécurisation plus-value, …) auxquelles l'assuré peut souscrire et qui effectue des arbitrages automatiquement en fonction des comportements des marchés. Mais comme nous l'avons expliqué dans la section 1.3.3, nous avons retiré ces arbitrages qui biaisaient énormément nos modèles.

Au niveau des performances des prédictions, on note que les modèles à base d'arbres de décision sont significativement meilleurs que la régression logistique. Ce dernier étant un modèle additif, il ne prend pas en compte les interactions entre variables explicatives. Il ne semble donc pas adapté à la modélisation des comportements d'arbitrage. En effet, tout porte à croire que ces comportements sont la conséquence de plusieurs phénomènes, il est donc logique que les arbres performent mieux car ils sont plus adaptés à ce type de problématique.

Globalement, ce qui ressort des résultats est que nous ne sommes pas parvenus à trouver un modèle de classification qui émet des prédictions correctes. En revanche les scores de *Log Loss* obtenus par nos modèles sont très satisfaisants. Ce qui nous laisse penser qu'ils sont plutôt performants pour estimer la probabilité d'un individu à arbitrer dans l'année.

Pour la suite, dans la modélisation dite par « fréquence sévérité » nous garderons la foret aléatoire rééchantillonnée étant donné que c'est le modèle qui permet d'obtenir les meilleures prédictions (l'aire sous la courbe PR est la plus élevée).

En revanche, dans la modélisation dite par « l'espérance » nous garderons la forêt aléatoire simple. En effet c'est le modèle ayant la *Log loss* la plus faible. C'est donc en théorie le modèle le plus approprié pour estimer la probabilité d'arbitrer.

Ci-dessous la synthèse des résultats obtenus.

Tableau 21 : Synthèse des performances obtenues par les modèles de classification

Méthode	Hyperpamètres optimaux	Taux de précision	Log Loss	AUC Precision-Recall
GLM	Pénalité : Ridge Lambda : 0.1	93,78%	**0,221**	**0,26**
GLM 80/20	Pénalité : Ridge Lambda : 1	93,25%	**0,304**	**0,29**
CART	Profondeure maximale : 8 Fonction d'hétérogénéité : Indice de Gini Nombre d'observation minimale sur une feuille : 190	94,26%	**0,183**	**0,32**
CART 80/20	Profondeure maximale : 5 Fonction d'hétérogénéité : Indice de Gini Nombre d'observation minimale sur une feuille : 330	91,37%	**0,265**	**0,35**
Bagging	Nombre d'arbre : 50 Taille des échantillions boostrap : 75% Profondeure maximale : 8 Fonction d'hétérogénéité : Entropie Nombre d'observation minimale sur une feuille : 400	94,28%	**0,178**	**0,33**
Bagging 80/20	Nombre d'arbre : 50 Taille des échantillions boostrap : 75% Profondeure maximale : 6 Fonction d'hétérogénéité : Entropie Nombre d'observation minimale sur une feuille : 300	91,63%	**0,261**	**0,35**
Random Forest	Nombre d'arbre : 75 Nombre de variable : p/2 Profondeure maximale : 8 Fonction d'hétérogénéité : Indice de Gini Nombre d'observation minimale sur une feuille : 100	94,32%	**0,171**	**0,36**
Random Forest 80/20	Nombre d'arbre : 50 Nombre de variable : p/2 Profondeure maximale : 5 Fonction d'hétérogénéité : Entropie Nombre d'observation minimale sur une feuille : 300	92,79%	**0,272**	**0,38**

1.2. Modèles de régression

Tableau 22 : Information sur les données test (2)

Données de test	
Séparation	**2015-2019 / 2020-2021**
Taille de l'échantillon test	15 022
Taux arbitré moyen	2,51%
Taux arbitré médian	-0,79%

Dans cette section nous tentons de modéliser la variable *taux_arbitrage*. Comme nous l'avons vu, c'est une variable qui prend ces valeurs entre -1 et 1 et qui est quasiment centrée en 0. Pour évaluer la qualité de nos modèles nous utiliserons les scores de performance énoncés dans la section 2.1.5.2. Nous mettrons ces scores en perspective avec ceux obtenus par le modèle simpliste qui prédit à chaque fois la moyenne (ou la médiane) dont les scores obtenus sont les suivants :

Tableau 23 : Performances des modèles triviaux de régression

Méthode	Taux médian	Taux moyen
Erreur absolue moyenne test	0,284	0,287
R^2	0	0

1.2.1. Modèle GLM de régression

Comme pour la régression logistique, on optimise par *cross validation* le type de pénalité à appliquer ainsi que λ l'intensité de la pénalité.

Le modèle retenu après *cross validation* est le modèle classique, sans pénalisation.

D'après le modèle, les variables ayant le plus gros impact sur le taux arbitré sont l'ancienneté, le taux de pb, la moyenne du CAC40 sur l'année précédente, la proportion d'unités de compte et la PM du contrat.

Tableau 24 : Importances des variables GLM taux d'arbitrage

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0191442	0.0008540	22.417	< 2e-16***
nb_arbitrage_nmoins1	0.0048651	0.0006876	7.075	1.49e-12***
nb_arbitrage_cumul	-0.0031722	0.0004107	-7.723	1.13e-14***
classe_age2.[70 : ...[	-0.0007181	0.0002815	-2.551	0.01075*
classe_anciennete2.[5 : 11[	-0.0017832	0.0003096	-5.759	8.45e-09***
classe_anciennete3.[11 : ...[	-0.0038247	0.0003207	-11.928	< 2e-16***
classe_ecart_cac2.[-5% : ...[	-0.0032449	0.0004511	-7.193	6.32e-13***
classe_ecart_oat2.[1.3% : ...[	0.0015241	0.0004624	3.296	0.00098***
classe_ecart_pmv_uc_eu2.[-2% : 1%[	0.0027095	0.0003762	7.203	5.91e-13***
classe_ecart_pmv_uc_eu3.[1% : ...[	0.0018823	0.0008279	2.274	0.02299*
classe_moyenne_cac2.[4500 : 5200[	-0.0051100	0.0005013	-10.194	< 2e-16***
classe_moyenne_cac3.[5200 : ...[	-0.0141814	0.0006843	-20.724	< 2e-16***
classe_nb_supports2.[400 : 1200[	0.0002926	0.0003015	0.971	0.33178
classe_nb_supports3.[1200 : ...[	0.0012853	0.0003275	3.924	8.70e-05***
classe_part_uc_pmv2.[10% : ...[	0.0037776	0.0008134	4.644	3.41e-06***
classe_pm2.[50K : 300K[	0.0015874	0.0002677	5.930	3.04e-09***
classe_pm3.[300K : ...[	0.0052942	0.0004164	12.715	< 2e-16***
classe_pmv_eu_prop2.]1.5% : 2.5%[	-0.0019903	0.0003825	-5.203	1.96e-07***
classe_pmv_eu_prop3.[2.5% : ...[	-0.0012953	0.0007063	-1.834	0.06665.
classe_pmv_prop2.]1% : 3%[	0.0005864	0.0004433	1.323	0.18586
classe_pmv_prop3.[3% : ...[	-0.0021577	0.0006837	-3.156	0.00160**
classe_pmv_uc_prop2.]0.5% : 2%[	0.0021230	0.0009048	2.346	0.01895*
classe_pmv_uc_prop3.[2% : ...[	0.0014450	0.0010476	1.379	0.16777
classe_prop_uc2.[10% : 50%[	-0.0039981	0.0004549	-8.789	< 2e-16***
classe_prop_uc3.[50% : ...[	-0.0225051	0.0004809	-46.799	< 2e-16***
classe_taux_pb2.[1.5% : 2.5%[	-0.0064463	0.0005059	-12.742	< 2e-16***
classe_taux_pb3.[2.5% : ...[	-0.0056079	0.0006286	-8.922	< 2e-16***
classe_tx_rdmt_cac2.[0% : 10%[	0.0002185	0.0004429	0.493	0.62170
classe_tx_rdmt_cac3.[10% : ...[	NA	NA	NA	NA

Malgré certaines p-value sans équivoques, les coefficients estimés sont extrêmement faibles. L'effet marginal de chaque variable est quasiment nul. Si un individu possède toutes les caractéristiques de base, le modèle estime que le taux net arbitré sera de 1.9%. Or pour un individu possédant toutes les caractéristiques propices (selon le modèle) à arbitrer une grosse partie de sa PM vers l'UC, le modèle estime qu'il arbitrera 4.2% de sa PM vers l'UC sur l'année. Autrement, le taux net arbitré maximal que peut prédire notre modèle est 4.2%.

On aperçoit ici clairement la limite de la modélisation par modèle linéaire qui n'est pas capable de modéliser la distribution du taux d'arbitrage qui a des queues lourdes.

C'est donc sans grande surprise que l'on constate des prédictions très mauvaises. L'erreur absolue moyenne sur la base de test diminue de seulement 2% par rapport au modèle simpliste qui prédit le taux médian à chaque fois.

Tableau 25 : Résultats du modèle GLM taux d'arbitrage

Méthode	**GLM**
Hyperpamètres optimaux	Pénalité : aucune Lambda : 0
Erreur absolue moyenne train	0,272
Erreur absolue moyenne test	**0,277**
R^2	**0,052**

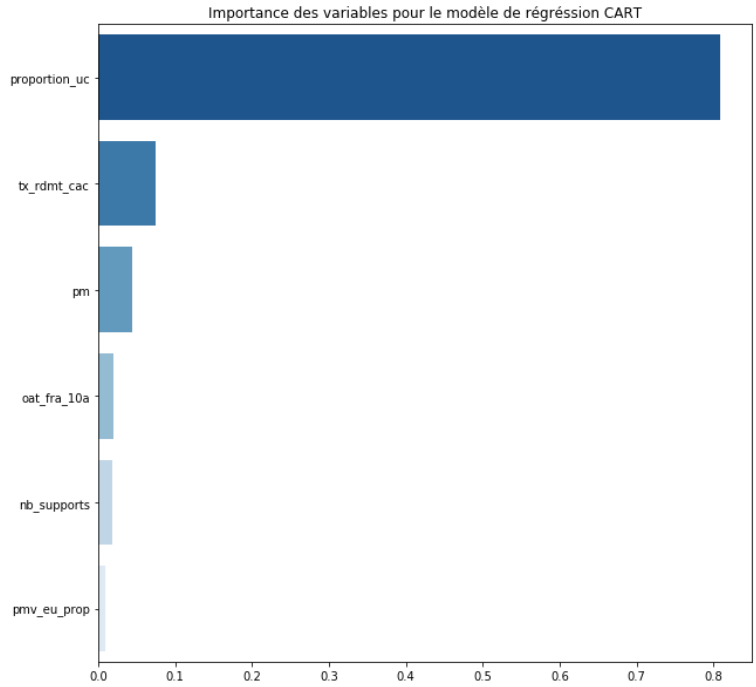
1.2.2. Modèle CART de régression

Pour la régression avec le modèle CART, nous nous sommes restreints au même hyperparamètres que pour la classification, à l'exception de la fonction d'hétérogénéité. Dans le cadre de la régression, nous avons le choix entre l'erreur quadratique moyenne (MSE) et l'erreur absolue moyenne (MAE). La seconde, bien que réputée plus stable, nécessite un temps de calcul bien plus important, raison pour laquelle nous avons choisi d'utiliser la fonction MSE comme fonction d'hétérogénéité.

Le modèle retenu après cross validation est un arbre d'une profondeur maximale de 7 avec un minimum de 220 données par feuille.

La variable qui est de loin la plus importante est la proportion d'UC dans le contrat en n-1 avec une contribution (cf section 2.2.2.5) de 0.8. Autrement dit, la prédiction émise par notre modèle dépend à 80% de la proportion d'UC en n-1.

Tableau 26 : Importance des variables modèle CART taux d'arbitrage



Sans surprise, le fait qu'une seule variable est une contribution de 80% à elle seule conduit à des résultats mauvais. L'erreur absolue moyenne décroit de seulement 5% comparé à celle des modèles triviaux qui prédisent la valeur médiane à chaque fois. Le coefficient de détermination R^2 s'élève lui à 0.13, ce qui est très faible.

Tableau 27 : Résultats du modèle CART taux d'arbitrage

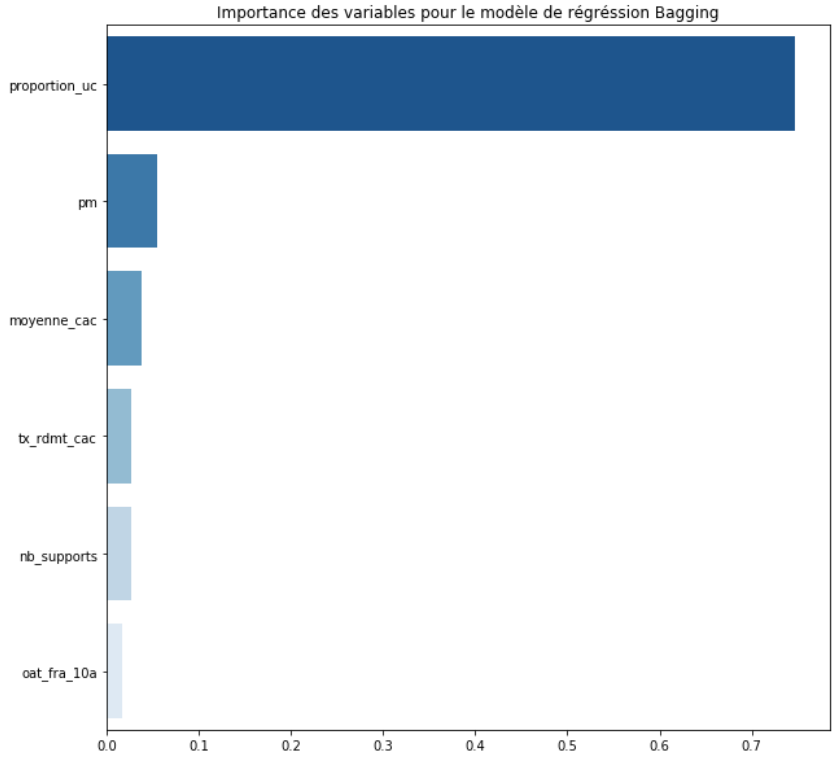
Méthode	**CART**
Hyperpamètres optimaux	Profondeure maximale : 7 Fonction d'hétérogénéité (fixée) : MSE Nombre d'observation minimale sur une feuille : 220
Erreur absolue moyenne train	0,230
Erreur absolue moyenne test	**0,267**
R^2	**0,132**

1.2.3. Modèle Bagging de régression

Le modèle retenu après cross validation est une agrégation de 150 arbres d'une profondeur maximale de 9 avec un minimum de 50 données par feuille, construits en utilisant à chaque fois 80% de la base d'entrainement.

En agrégant les résultats de plusieurs arbres construits sur différents échantillons, le bagging permet d'équilibrer très légèrement l'importance des variables en réduisant celle de *proportion_uc*.

Tableau 28 : Importance des variables modèle Bagging taux d'arbitrage



Néanmoins, cette dernière reste bien trop élevée, conduisant inévitablement à des résultats médiocres, bien que légèrement supérieurs à ceux de l'arbre simple.

Tableau 29 : Résultats du modèle Bagging taux d'arbitrage

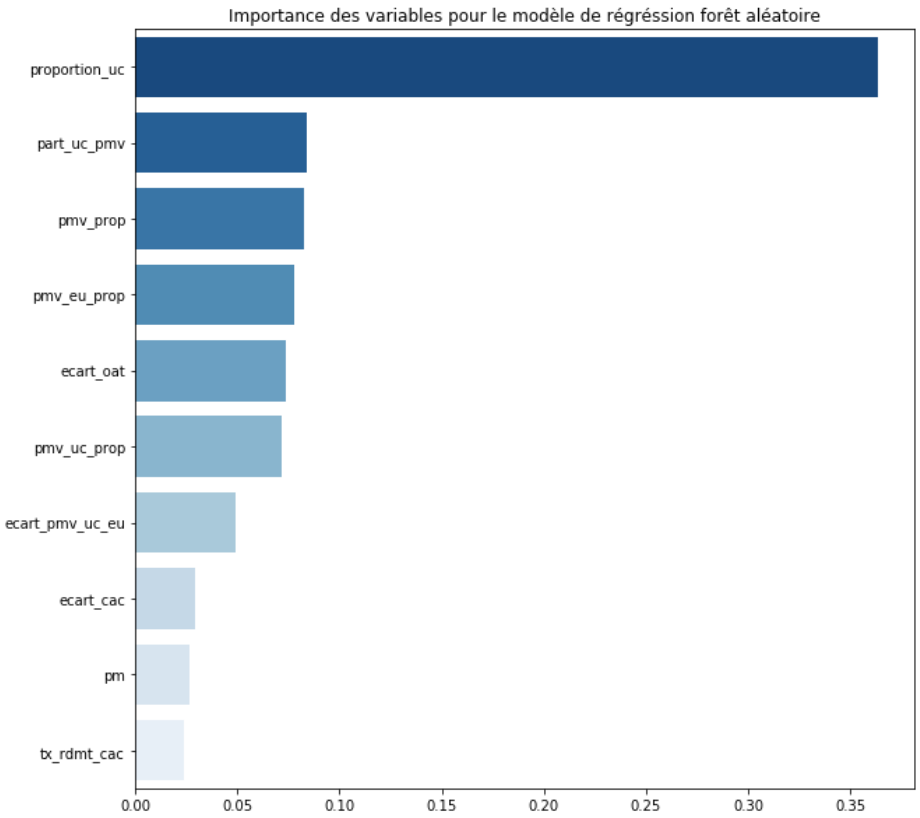
Méthode	**Bagging**
Hyperpamètres optimaux	Nombre d'arbre : 150 Taille des échantillions boostrap : 80% Profondeure maximale : 9 Fonction d'hétérogénéité (fixée) : MSE Nombre d'observation minimale sur une feuille : 50
Erreur absolue moyenne train	0,222
Erreur absolue moyenne test	**0,259**
R^2	**0,165**

1.2.4. Modèle Forêt aléatoires de régression

Aux vues des résultats précédents, on remarque bien que la variable *proportion_uc* est problématique. Elle est indéniablement très significative, mais elle semble masquer l'effet des autres variables sur le taux arbitré.

Une forêt aléatoire est donc parfaitement adaptée à notre problématique car elle va agréger les résultats de plusieurs arbres, chacun construit avec un sous-ensemble de nos variables explicatives, tirées aléatoirement et sans remise. Autrement dit, certains arbres seront construits avec la variable *proportion_uc*, et d'autres non, ce qui va en théorie, permettre au modèle d'apprendre à faire sans cette variable.

Tableau 30 : Importance des variables modèle Random Forest taux d'arbitrage



Effectivement, on constate que l'importance des variables est bien plus équilibrée. La proportion d'UC dans le contrat reste la variable la plus importante mais avec une contribution de seulement 35%, ce qui est bien plus raisonnable. Grâce à cette façon de procéder, plusieurs variables se révèlent être importantes alors qu'elles n'apparaissaient pas avec les modèles CART et de Bagging. C'est par exemple le cas des variables liées à la plus ou moins-value réalisée en n-1.

Parmi les variables conjoncturelles on remarque que *ecart_oat*, qui est la différence entre le taux de plus ou moins-value réalisée en n-1 et le taux de l'OAT France 10ans, a une contribution de près de 10%. Graphiquement nous avions remarqué que plus cette différence est faible, plus les assurés arbitrent vers l'UC.

Les performances du modèle forêt aléatoire sont de loin meilleures que les résultats obtenus avec les autres modèles à base d'arbre de décision. L'erreur absolue moyenne obtenue sur la période de test est de 0.184 et le coefficient de détermination R^2 s'élève à près de 0.3, soit près du double des autres modèles à base d'arbre.

Tableau 31 : Résultats du modèle Random Forest taux d'arbitrage

Méthode	**Random Forest**
Hyperpamètres optimaux	Nombre d'arbre : 150 Nombre de variable : √p Profondeure maximale : 9 Fonction d'hétérogénéité (fixée) : MSE Nombre d'observation minimale sur une feuille : 50
Erreur absolue moyenne train	0,167
Erreur absolue moyenne test	**0,184**
R^2	**0,290**

1.2.5. Conclusion et interprétations

Contrairement aux modèles de classifications, ici un modèle se distingue clairement des autres. En effet, les performances obtenues avec le modèle de forêts aléatoires sont nettement supérieures. C'est donc ce modèle que nous avons retenu pour l'agrégation par année. Ceci est dû à la méthode de construction des forêts aléatoires qui utilise seulement une partie des variables pour construire les arbres qui composent la forêt. De cette manière, le modèle est capable de capter l'effet de certaines variables jusqu'ici caché par la variable *proportion_uc*.

La proportion d'UC dans le contrat en $n-1$ est de loin la plus discriminante pour notre modèle de régression. Le fait que cette variable soit la plus importante pour notre modèle de régression ne nous surprend pas. En effet, comme nous l'avons vu dans la section 1.3.2.1, les différences de taux arbitrés en fonction de la proportion d'UC sont très marquées graphiquement. Nous avions émis l'hypothèse que la valeur de cette variable va souvent guider le sens de l'arbitrage. Plus la proportion d'UC dans le contrat est faible, plus l'assuré aura tendance à arbitrer vers de l'UC. Dans la plupart du temps, l'arbitrage est en direction du fond minoritaire[10], dans une logique d'équilibre des fonds (cela étant particulièrement vrai pour les contrats très déséquilibrés). Pourtant, dans les récentes recherches portant sur la modélisation des arbitrages, cette variable ne s'est jamais avérée être une variable aussi importante. Nous pensons que ceci est dû à la procédure d'arbitrage si particulière chez AG2R LAMONDIALE. En effet, comme il est impossible pour nos assurés d'arbitrer sans consulter leur conseillé, on peut légitimement penser que ces derniers encouragent les assurés à diversifier leurs placements.

Le fait que cette variable est à elle seule une contribution de plus de 75% dans les modèles CART et Bagging s'est avéré être très problématique pour les prédictions.

Grâce à la forêt aléatoire, on obtient une amélioration significative des performances comparé aux autres modèles. L'erreur absolue moyenne obtenue sur la période de test est de 0.184, soit plus de 35% inférieur à celle obtenue avec le modèle qui prédit le taux médian.

Les variables les importantes de ce modèle sont bien-sûr, *proportion_uc*, mais aussi les variables liées à la plus ou moins-value réalisée l'année précédente. L'importance de ces variables est effectivement visible graphiquement (cf graphiques en annexe) mais elles ne sont pourtant pas utilisées par les modèles CART et de Bagging. Pour finir, la différence entre le taux de plus ou moins-value réalisée en n-1 et le taux de l'OAT France 10ans joue également un rôle significatif. Plus cette différence est faible, plus les assurés ont tendance à se diriger vers les fonds UC pour espérer obtenir des rendements plus conséquents et

[10] Pour rappel, pour les assurés ayant moins de 15% d'UC, 83% des arbitrages sont en direction de l'UC.

significativement différents du taux considéré comme étant le taux « sans risque » des placements à long terme.

Bien que les prédictions faites par ce modèle soient moyennes, il possède un pouvoir explicatif très limité. En effet, que le coefficient de détermination R^2 s'élève à seulement 0.3 signifiant donc que notre modèle est capable d'expliquer environ 30% des variations du taux arbitré, le reste n'étant pas capté par notre modèle. Les raisons peuvent être multiple. Premièrement, les variables liées à la plus-value qui apparaissent comme étant significatives dans notre forêt, sont les variables qui sont les plus corrélées à *proportion_uc*. Même si ces corrélations ne sont pas extrêmement élevées (aux alentours de 0.35 en valeur absolue), cela limite le gain d'information. Deuxièmement, l'impact des variables exogènes est assez limité. En cumulé, ces variables ont une contribution de près de 20%. Pourtant il est assez naturel de penser que les arbitrages sont extrêmement dépendant du contexte économique et financier et donc que ces variables devraient avoir une contribution plus importante. Ce que semble nous dire notre modèle est que c'est d'avantage l'impact que va avoir ce contexte économique sur la rentabilité des différents fonds qui est important, plutôt que le contexte lui-même, qui impact les comportements d'arbitrage. Cette hypothèse n'est pas absurde, d'autant plus sur notre portefeuille de contrat, où la procédure d'arbitrage y est extrêmement lourde et dissuade les comportements réactionnaires aux marchés.

2. Agrégation des deux modèles et prédictions

Maintenant que nous avons déterminé les modèles les plus pertinents à utiliser selon le type de modélisation choisie, nous allons agréger les prédictions par année.

2.1. Modélisation par fréquence sévérité

Pour la modélisation dite par « fréquence sévérité » le montant net arbitré durant l'année n, M_n est estimé de la façon suivante :

$$\hat{M}_n = \sum_{i=1}^{c_n} \hat{A}(\mathbb{X}_{n_i})\hat{Y}(\mathbb{X}_{n_i}|A_{n_i} = 1)PM_i$$

Avec :

- c_n le nombre de contrat en cours d'année n.
- $\mathbb{X}_{n_i}$ les caractéristiques de l'individu i à l'année n.
- $A(.)$ le classifieur qui prend la valeur 1 s'il prédit au moins un arbitrage, 0 sinon.
- $Y\ .|A_{n_i} = 1$ la régression qui prédit le taux net arbitré dans le cas où l'individu i effectue un arbitrage.
- PM_i la provision mathématique au 31/12/n-1 de l'individu i.

Comme énoncé dans la section précédente, le classifieur $A(.)$ retenu est la forêt aléatoire rééchantillonnée à 80/20 et la régression $Y\ .|A_{n_i} = 1$ retenue est également la forêt aléatoire.

Voici les résultats obtenus :

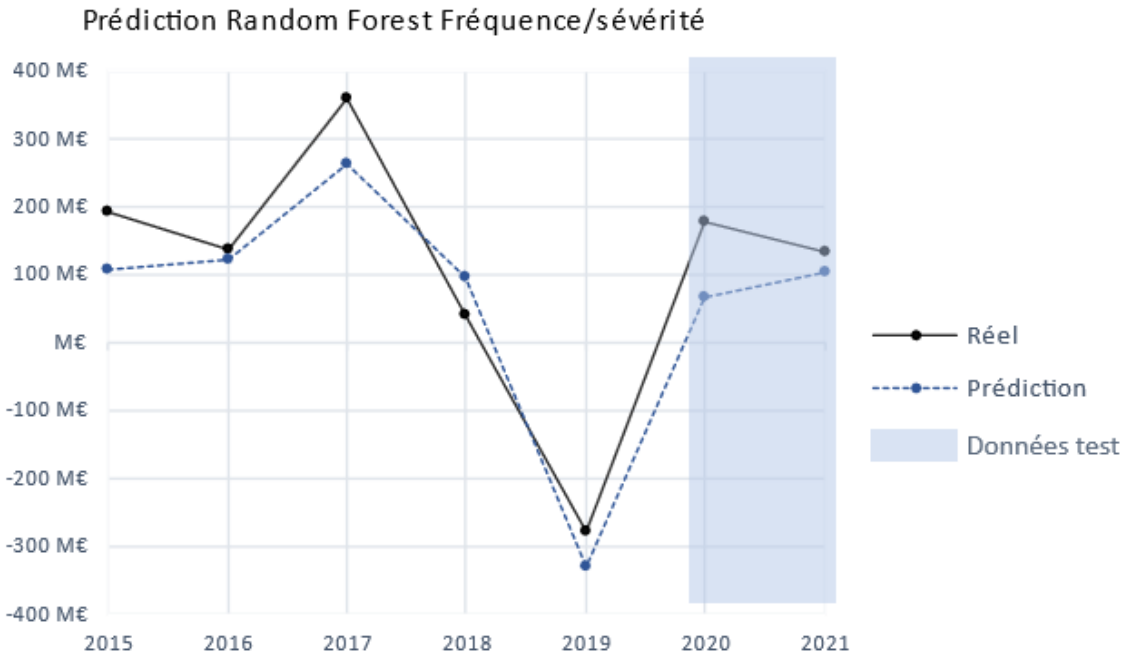


Figure 44 : Prédiction du montant net annuel arbitré par la modélisation Fréquence/Sévérité

Comme nous pouvons le constater, le modèle n'est pas parfait mais il arrive à prévoir la tendance avec plus ou moins de précision, et ce même sur la période de test (2020-2021). En effet, on constate par exemple qu'entre 2019 et 2020, le montant net arbitré à drastiquement changé. On est passé d'un montant arbitré net en 2019 de -280 millions d'euros à 180 millions d'euros en 2020, soit une différence de 460 millions d'euros. Notre modèle, bien que n'ayant jamais rencontré de données de 2020, et que l'année en tant que telle ne fasse pas partie des variables explicatives, avait anticipé un tel mouvement.

Malgré ça, il demeure une différence de près de 100 millions d'euros entre le montant prédit et le montant réel en 2020. En 2021 la différence est d'environ 30 millions d'euros. Aux vues de l'étendue des valeurs possibles, 30 millions d'euros est une différence qui peut être considérée comme acceptable. En revanche une différence de 100 millions est quand même plus problématique. Ce qui nous empêche de penser que le modèle n'est pas si mauvais est que l'année 2020 fut une année très particulière au niveau des arbitrages. En effet, entre mars et mai 2020, juste après la chute du CAC40 liée à la crise sanitaire, on a observé des mouvements d'arbitrages massifs vers l'UC. D'une part, ces mouvements étaient historiquement élevés, d'autre part notre modèle semble ne pas bien capter les effets conjoncturels des arbitrages, ceci faisant que la prédiction finale de 2020 est moins précise que celle de 2021.

La dernière chose attirant notre attention est que, excepté en 2018, tous les montants prédits sont inférieurs aux montants réels. Idéalement on aimerait pouvoir relever légèrement la courbe des prédictions. La cause est probablement que le classifieur ne prédit pas assez souvent « arbitrage », autrement dit que le nombre d'arbitrages est sous-estimé. On peut corriger ça en accentuant le rééchantillonnage. Malheureusement, le résultat n'est pas celui escompté. Certes la courbe des prédictions est légèrement plus haute, mais on observe une grosse déviance aux années de « pique » 2017 et 2019 (cf. résultats avec un rééchantillonnage 70/30 en annexe).

2.2. Modélisation par fréquence sévérité probabiliste

Pour la modélisation dite par « fréquence sévérité probabiliste » le montant net arbitré durant l'année n, M_n est estimé de la façon suivante :

$$\hat{M}_n = \sum_{i=1}^{c_n} \hat{P}(\mathbb{X}_{n_i})\hat{Y}(\mathbb{X}_{n_i}|A_{n_i} = 1)PM_i$$

En effet, on a par la loi forte des grands nombres le résultat suivant :

$$\sum_{i=1}^{c_n} \mathbb{E}(M_{n_i}) \xrightarrow{lfgn} M_n$$

Or ;

$$M_{n_i} = y_{n_i} PM_{n_i}$$

La PM étant déterministe, nous avons :

$$\mathbb{E}(M_{n_i}) = PM_{n_i}\mathbb{E}(y_{n_i})$$

En notant $\mu_{n_i} = \mathbb{E}\ y_{n_i}$ pour simplifier les notations, on a d'après le résultat détaillé dans la section 2.1.1.3 :

$$\mu_{n_i} = \mathbb{E}(y_{n_i}|A_{n_i} = 1)P(A_{n_i} = 1)$$

Que l'on estime de la façon suivante :

$$\hat{\mu}_{n_i} = \hat{Y}(\mathbb{X}_{n_i}|A_{n_i} = 1)\hat{P}(\mathbb{X}_{n_i})$$

Avec

- $Y\ .|A_{n_i} = 1$ la régression qui prédit le taux net arbitré dans le cas où l'individu i effectue un arbitrage.
- $P(.)$ l'estimation de la probabilité d'arbitrer.

Finalement on a donc :

$$\hat{M}_n = \sum_{i=1}^{c_n} \hat{P}(\mathbb{X}_{n_i})\hat{Y}(\mathbb{X}_{n_i}|A_{n_i} = 1)PM_i$$

Comme énoncé dans la section précédente, $P(.)$ est calculé par la forêt aléatoire classique (sans rééchantillonnage) et la régression $Y\ .|A_{n_i} = 1$ retenue est également la forêt aléatoire.

Voici les résultats obtenus :

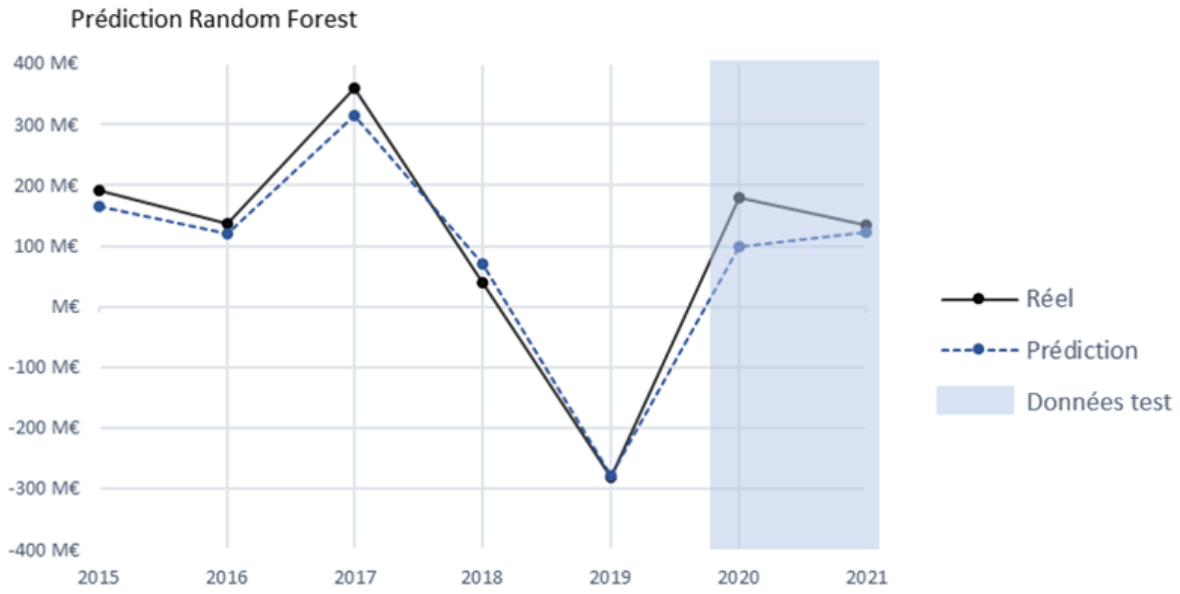


Figure 45 : Prédiction du montant net annuel arbitré par la modélisation fréquence/sévérité probabiliste

Nous pouvons noter que cette méthode de modélisation conserve les défauts de la première méthode, à savoir la sous-estimation des montants arbitrés et l'écart important entre le montant prédit et le montant réel en 2020. Néanmoins elle permet quand même de les atténuer très largement.

On obtient avec cette méthode des prédictions très satisfaisantes, plus stables qu'avec la première méthode. Chaque année, les prédictions suivent parfaitement la tendance réelle alors que l'année en tant que telle ne fait pas partie des variables explicatives. En essayant de prédire pour chaque assuré, l'espérance du taux net arbitré, et non pas le taux lui-même, il semble que l'on observe bien la convergence espérée. C'est donc cette modélisation que nous retenons pour la suite.

3. Interprétation et limites

Aucune des deux méthodes n'a permis d'obtenir de bonnes prédictions individuelles, et ceux peu importe les modèles utilisés. Pourtant on observe bien une convergence vers les montants réels au moment de l'agrégation par année. Plusieurs raisons peuvent expliquer ceci.

La raison principale est certainement dû à la loi des grands nombres.

Tableau 32 : Illustration effet de la convergence

No_police	Annee	Probabilité d'arbitrer	Montant arbitrage prédit s'il arbitre	Montant arbitrage prédit	Montant réel	Erreur
0001	20XX	0,25	50 000 €	12 500 €	- €	- 12 500 €
0002	20XX	0,25	50 000 €	12 500 €	- €	- 12 500 €
0003	20XX	0,25	50 000 €	12 500 €	50 000 €	37 500 €
0004	20XX	0,25	50 000 €	12 500 €	- €	- 12 500 €

Prédiction annuelle	Montant réel annuel
50 000 €	50 000 €

L'exemple trivial illustre bien le problème auquel nous faisons face et cette notion de convergence. Chaque prédiction est mauvaise mais au total sur l'année nous avons le bon montant. Pour pouvoir observer un tel phénomène, il faut bien sûr que le nombre de données que l'on agrège soit suffisamment

grand. Dans notre cas, cette agrégation est faite sur plus de 100 000 prédictions ce qui semble être suffisant pour que la loi des grands nombres s'applique.

D'autres raisons peuvent venir s'ajouter à ça. Il se peut que la partie des modèles qui prédit bien soit de poids majoritaire sur le résultat final, c'est-à-dire que la plupart des gros arbitrages sont bien prédits, et que les petits arbitrages le soient moins. Il se peut également que les erreurs des deux modèles se compensent. Par exemple on surestime la probabilité d'arbitrage et on sous-estime le montant, au moment de faire le produit des deux prédictions les deux erreurs se compensent.

Quoi qu'il en soit, les résultats semble robuste et nous aurions aimé pouvoir appuyer et légitimer cette robustesse avec un intervalle de confiance permettant d'encadre nos prédictions assez précisément et avec un niveau de confiance élevé. Malheureusement cela n'a pas été possible. Il existe bien des méthodes (courbes quantiles) permettant d'obtenir un intervalle de confiance pour une prédiction émise par une forêt aléatoire, mais l'amplitude de l'intervalle finale était beaucoup trop large. La raison à cela est que nous faisons d'abord le produit de deux prédictions (fréquence x sévérité), puis nous sommons toutes les prédictions de l'année.

Tableau 33 : illustration problème intervalle de confiance

No_police	Annee	Probabilité d'arbitrer	IC proba	Montant arbitrage prédit s'il arbitre	IC sévérité	Montant arbitrage prédit	IC prédiction
0001	20XX	0,25	[0.15 ; 0.35]	50 000 €	[45 000 € ; 55 000 €]	12 500 €	[6 750 € ; 19 250 €]
0002	20XX	0,25	[0.15 ; 0.35]	50 000 €	[45 000 € ; 55 000 €]	12 500 €	[6 750 € ; 19 250 €]
0003	20XX	0,25	[0.15 ; 0.35]	50 000 €	[45 000 € ; 55 000 €]	12 500 €	[6 750 € ; 19 250 €]
0004	20XX	0,25	[0.15 ; 0.35]	50 000 €	[45 000 € ; 55 000 €]	12 500 €	[6 750 € ; 19 250 €]

Prédiction annuelle	IC annuel
50 000,00 €	[27 000 € ; 77 000 €]

[11]

Ci-dessous un exemple illustratif qui reprend la même situation que le tableau précédent. On constate que malgré des intervalles de confiance relativement étroit pour chaque modèle (amplitude de +/- 10 points de pourcentage pour la probabilité d'arbitrer et amplitude de +/- 5 000 € pour le montant), on se retrouve avec une prédiction annuelle de 50 000 € et un intervalle de confiance d'une amplitude de 50 000 € également. Cet effet s'accentuant à mesure que l'on somme les prédictions.

[11] 6 750 € correspond à 0.15*45 000 € et 19 250 € correspond à 0.35*55 000 €

PARTIE 4 : PROJECTION

1. Objectif

La dernière partie de notre étude consiste à implémenter notre modélisation des arbitrages dans notre outil qui projette la valeur de notre portefeuille sur 60ans. On regardera ensuite l'impact sur différents indicateurs clés.

Idéalement, nous aurions aimé implémenter notre modèle dans l'outil de modélisation prospective d'AG2R LA MONDIALE mais cela n'a pas été possible dans le laps de temps donné. En effet, cet outil est très gourmand en temps de calcul car une modélisation stochastique (1000 GSE) est utilisée pour prédire l'évolution de l'actif. De plus, les inputs passifs sont légèrement agrégés contrairement à notre modèle qui est construit à la maille contrat, ce qui nous aurait obligé à construire un modèle agrégé. La solution subsidiaire retenue a été d'implémenter notre modèle dans un outil Excel qui approxime avec une modélisation déterministe les valeurs officielles calculées. Cet outil est très souvent utilisé en interne pour faire des calculs de sensibilités car il réplique très convenablement les calculs officiels et les résultats sont instantanés.

Cependant, l'outil Excel projette un unique Model Point représentant l'individu moyen de notre portefeuille. De plus, il est impossible d'implémenter des forêts aléatoires dans la version classique d'Excel. De ce fait, pour implémenter la modélisation des arbitrages, deux possibilités sont envisageables :

- Prédire sur 60 ans le taux net arbitré de l'individu moyen sur Python puis intégrer dans l'outil Excel ces prédictions.
- Prédire sur 60 ans les taux nets arbitrés pour tous les individus encore présents dans le portefeuille le 31/12/2021. On calcule ensuite le taux moyen annuel puis on intègre dans l'outil Excel ces prédictions.

La première méthode présente l'avantage de respecter une certaine cohérence avec l'outil Excel mais la deuxième méthode est plus réaliste. De plus, nous avons vu que les performances individuelles de notre modélisation des arbitrages ne sont pas très précises. La principale force de notre modèle est de converger vers le taux réel quand on agrège les prédictions de l'année. De ce fait c'est la deuxième méthode qui a été retenue.

Nous allons donc faire tourner notre modèle sur un horizon de 60 ans sur les contrats encore ouverts le 31/12/2021, indépendamment des autres lois comportementales (rachats, décès, …) puis d'intégrer la chronique de taux d'arbitrages annuels prédits dans l'outil Excel. Ces taux seront ensuite multipliés par la PM globale pour avoir les montants annuels nets arbitrés puis pouvoir calculer les indicateurs qui nous intéressent. Cette PM évolue dans l'outil Excel en fonction de la modélisation de l'actif mais aussi en fonction des autres lois comportementales, qui sont supposées être indépendantes les unes des autres. Néanmoins, pour pouvoir faire évoluer notre loi d'arbitrage sur 60 ans, quelques modifications sont nécessaires.

2. Modification du modèle

En voulant faire ceci, nous faisons face à un problème de taille. Pour prédire le montant net arbitré en 2022, on utilise les modèles que nous avons retenus que nous agrégeons ensuite.

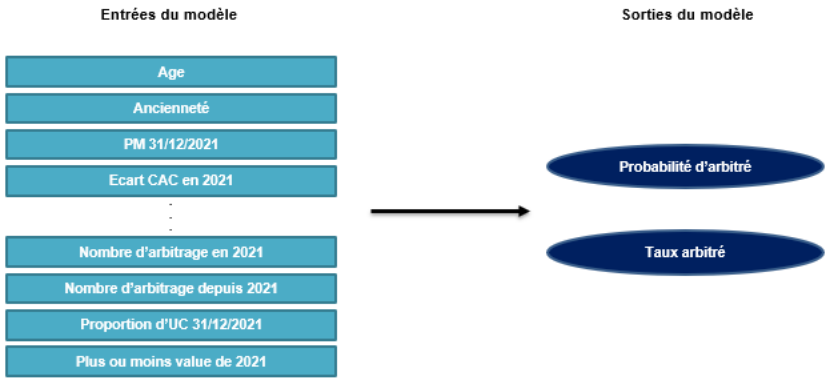


Figure 46 : Illustration du problème pour implémenter le modèle sur plusieurs années (1)

En revanche, si on désire prédire le montant net arbitré en 2023, on se retrouve bloqué car plusieurs variables sont indisponibles.

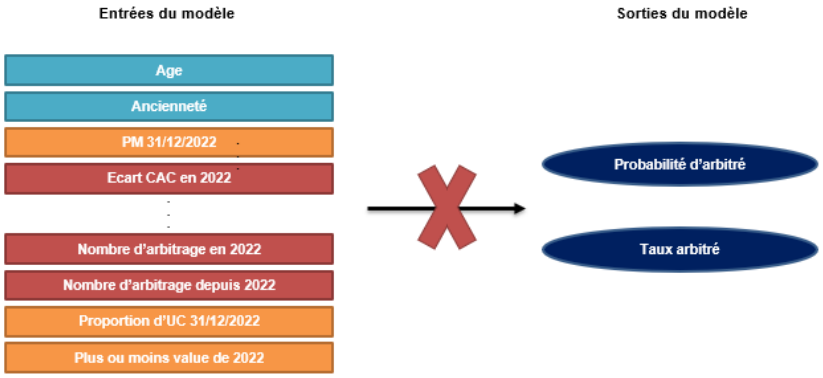


Figure 47 : Illustration du problème pour implémenter le modèle sur plusieurs années (2)

Premièrement, toutes les variables conjoncturelles ne sont pas encore disponibles (rendement du CAC 40, taux de l'OAT France 10 ans, …), ces variables ne peuvent donc pas être utilisée.

Deuxièmement, certaines variables endogènes doivent être modifiées pour permettre l'implémentation sur plusieurs années.

- Le nombre d'arbitrages en n-1 : Pour prédire 2022, on pouvait se servir du nombre d'arbitrages effectués en 2021 qui est actuellement connu. Néanmoins, si on désire au 31/12/2022 avoir des prédictions pour l'année 2042, on ne connaît pas le nombre d'arbitrages effectués en 2041. Or, pour rappel notre modèle a pour sorties une estimation de la probabilité d'arbitrer ou non dans

l'année et une estimation du taux d'arbitrages dans le cas où il arbitre. Donc, en aucun cas une prédiction est faite sur le nombre d'arbitrage qui sera effectué.
La solution qui s'est avérée être la plus pertinente a été de remplacer cette variable par une variable binaire qui indique seulement si oui ou non il y a eu arbitrage en n-1.

- Le nombre d'arbitrages depuis n-1 : Pour cette variable, nous faisons face au même problème étant donné que nous ne connaissons pas le nombre d'arbitrage effectué en 2041. Nous avons donc choisi de remplacer cette variable par le nombre d'années où l'assuré a effectué un (ou plusieurs) arbitrage(s). De cette manière on peut estimer cette valeur en utilisant les prédictions du modèle de classification :

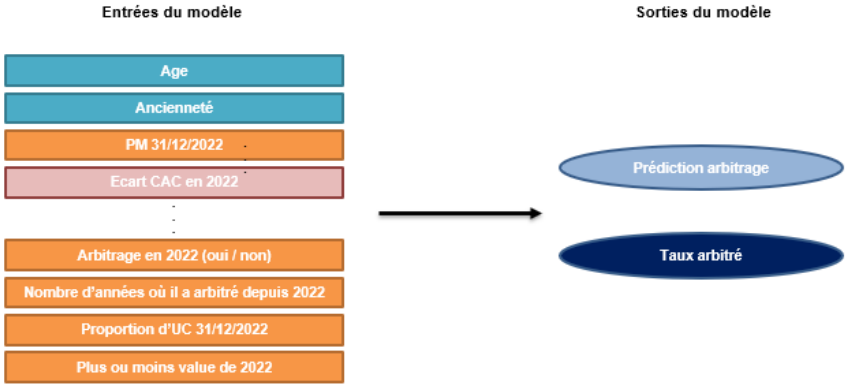


Figure 48 : Illustration du problème pour implémenter le modèle sur plusieurs années (3)

3. Résultat du modèle simplifié

Avec ces simplifications, on s'attend forcément à observer une dégradation de notre modèle. En réalité, cette dégradation s'avère assez limitée.

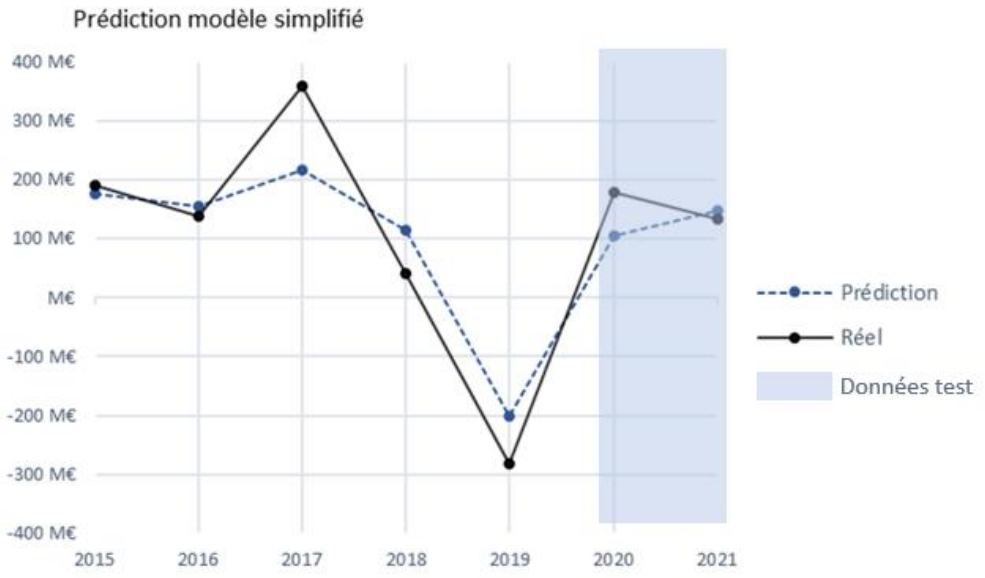


Figure 49 : Prédiction du modèle simplifié

Nous expliquons ce phénomène par deux raisons.

La première est que l'impact des variables exogènes dans le modèle complet était faible. Ce n'est donc pas choquant que leur retrait n'affecte pas énormément le modèle.

La seconde raison est que les modifications retenues pour les variables *nb_arbitrage_nmoins1* et *nb_arbitrage_cumul* permettent de garder l'essentiel de l'information, à savoir si l'individu est actif sur son contrat au niveau des arbitrages.

4. Implémentation des variables

La dernière étape pour pouvoir faire vivre notre modèle sur plusieurs années est d'estimer la future valeur des variables explicatives. Par exemple, pour l'âge de l'assuré ou l'ancienneté d'un contrat en 2042, il n'y a pas de soucis à avoir ces valeurs. En revanche, la tâche s'avère moins directe pour les autres variables. Voici les estimations retenues pour une année donnée n qui seront donc utilisées pour faire les prédictions de l'année $n+1$:

- La PM au 31/12/n : La projection de la PM devrait prendre en compte une estimation des flux entrants (versements, revalorisation) et sortants (rachats).
  Ces flux étant trop couteux à modéliser par contrat, et la PM étant dans notre modèle un indicateur de richesse du client, nous nous contenterons d'estimer la projection par la formule suivante :

$$PM_{31/12/\mathrm{n}} = PM_{31/12/\mathrm{n\text{-}1}} + PMV_{\mathrm{n}}$$

- La proportion d'UC au 31/12/n : Cette valeur dépend également des rachats et versements faits pendant l'année. Comme énoncé précédemment, ces flux sont couteux à modéliser par contrat et nous ne tirerons à priori pas d'avantage à les intégrer. En effet ces flux ne sont pas modélisés par fonds mais en proportion du contrat. Savoir que l'individu i a racheté x% de son contrat et

a versé $y\%$ ne nous aidera pas à mieux projeter sa proportion d'UC. On estimera donc cette proportion par la formule suivante :

$$\begin{aligned} Proportion\ UC_{31/12/n} &= \frac{PM\ UC_n}{PM_n} \\ &= \frac{PM\ UC_{n-1} + PM_{n-1} \times probabilité\ d'arbitrer_n \times\ taux\ arbitrage\ net_n}{PM_n} \\ &= \frac{PM_{n-1} \times\ Proportion\ UC_{31/12/n-1} + probabilité\ d'arbitrer_n \times taux\ arbitrage\ net_n}{PM_n} \end{aligned}$$

- Le taux de participation aux bénéfices en n : Ce taux est estimé pour les 60 prochaines années dans l'outil Excel. On reprend donc logiquement cette estimation.

- La plus ou moins-value sur l'EURO en n : Cette valeur dépend du taux de participation aux bénéfices versé cette année-là, du Taux Minimum Garantit, ainsi que de la répartition Euro/UC de l'épargne. On estime donc la plus ou moins-values sur l'Euro en prenant :

$$PMV\ EURO_n = PM_{n-1} \times \left(1 - Proportion\ UC_{31/12/n}\right) \times Max(taux\ PB_n\ , TMG_n)$$

- La plus ou moins-value sur l'UC en n : De même, cette valeur va dépendre de la répartition Euro/UC de l'épargne et du taux de rendement des UC. Une estimation de ce dernier est disponible pour les 60 prochaines années dans l'outil Excel. On estime donc la plus ou moins-values sur l'UC en prenant :

$$PMV\ UC_{31/12/n} = PM_{n-1} \times Proportion\ UC_{31/12/n} \times taux\ rendement\ UC_n$$

- La plus ou moins-value en n : Il suffit de faire la somme des deux estimations précédentes.

- La proportion des plus-values de l'année générée par les UC en n : Etant donné qu'il est possible d'être en moins-value sur l'UC, il faut faire attention à distinguer les deux cas :

$$Part\ UC\ PMV_n = \begin{cases} 0\ si\ PMV\ UC_{31/12/n} < 0 \\ \dfrac{PMV\ UC_{31/12/n}}{PMV_{31/12/n}}\ sinon \end{cases}$$

- Différence entre le taux de rendement UC et celui EU en n :

$$Ecart\ PMV_n = taux\ rendement\ UC_n - \ taux\ PB_n$$

- Arbitrage en n :

$$Arbitrage_n = \begin{cases} 1\ si\ probabilité\ d'arbitrage_n \geq seuil \\ 0\ sinon \end{cases}$$

Le seuil utilisé dans un cas classique est 0.5. Comme nos données sont extrêmement déséquilibrées, ce seuil n'est pas optimal pour nous car il ne va pas assez prédire d'arbitrages.

Pour corriger ça, nous avons abaissé le seuil à 0.35[12]. Ce dernier a été fixé par cross validation et maximise l'air sous la courbe PR.

- Le nombre d'années où l'assuré à effectuer au moins un arbitrage depuis n :

$$Nombre\ années\ arbitrage_n = Nombre\ années\ arbitrage_{n-1} + Arbitrage_n$$

Avec ces estimations et ces deux modifications, le modèle est maintenant capable d'évoluer sur autant d'années que désiré.

5. Projections

5.1. Résultat

Ci-dessous se trouve les prédictions du taux net arbitré pour les 60 prochaines années.

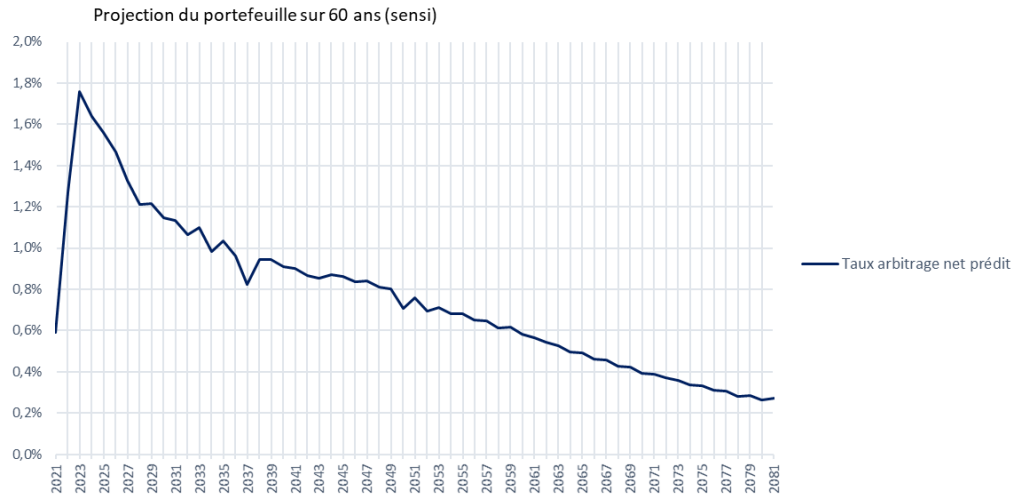


Figure 50 : Prédiction du taux d'arbitrage (net vers UC) sur 60 ans

Le taux d'arbitrage prédit est positif à chaque année et décroit de manière quasiment constante à partir de 2023. Ce phénomène est rassurant car on projette ici notre portefeuille au 31/12/2021. On ne modélise pas de nouvelles souscriptions. Donc mécaniquement, le portefeuille vieillit. Or comme on l'a vu, plus le souscripteur est âgé et plus le contrat est ancien, plus les probabilités d'effectuer un arbitrage diminuent, toutes choses égales par ailleurs.

Nous pensons également que cette chute à partir de 2023 est déclenchée et entretenue par la prédiction d'une hausse du taux de rendement du fond Euro à partir de cette année. Ce faisant, l'intérêt du fond UC est donc moindre. Pour rappel, les variables se basant sur les plus ou moins-values (PMV, PMV_EURO, PMV_UC, PART_UC_PMV, ECART_PMV) sont estimées avec ces prédictions de rendement et figurent parmi les plus importantes du modèle de régression. Dans le but de renforcer cette hypothèse, nous effectuerons un calcul de sensibilité en modifiant volontairement les prédictions des taux de rendements des fonds Euro et UC.

[12] Comme nous observons en moyenne sur notre portefeuille que moins de 7% des assurés arbitrent dans une année. Une probabilité d'arbitrer de 35% est donc plus de 5 fois supérieur à la moyenne.

On intègre maintenant ces prédictions de taux d'arbitrage net dans l'outil Excel. On procède de la même manière que les autres lois comportementales comme la loi de rachat par exemple, c'est-à-dire en multipliant le taux par le montant de PM total en début d'année. De cette manière on obtient une prédiction des montants d'arbitrages nets.

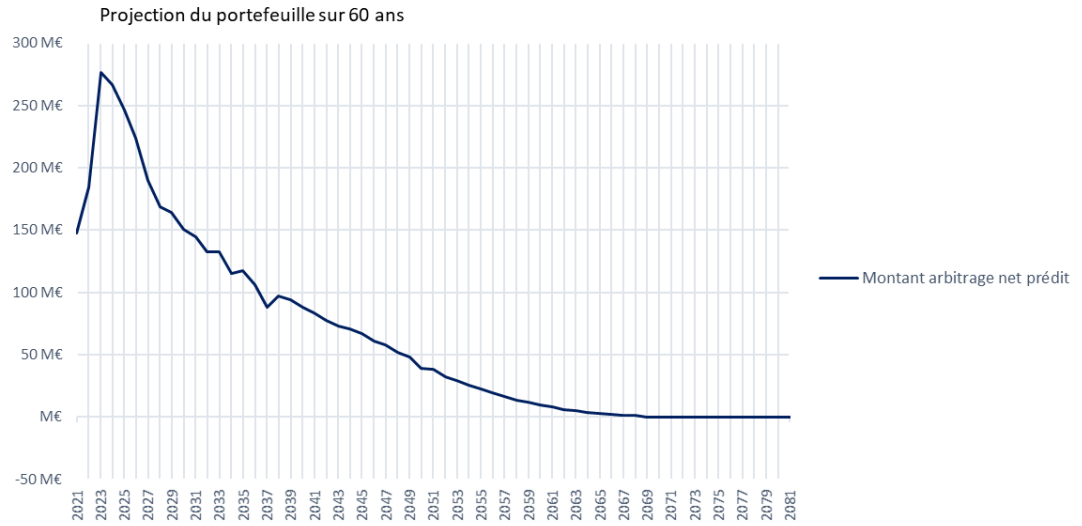


Figure 51 : Prédiction du montant arbitré (net vers UC) sur 60 ans

Comme on peut le constater, la tendance reste la même mais on converge plus rapidement vers des montants nets arbitrés nuls. Cela vient du fait que, dans l'outil Excel, la PM évolue en fonction des rachats et des décès, ce qui vient naturellement faire baisser la PM avec le temps. Ceci combiné au fait que les taux d'arbitrages nets prédits décroissent également avec le temps pour les raisons énoncées précédemment, provoque cette convergence vers 0.

5.2. Calcul de sensibilité

Comme expliqué précédemment, nous allons modifier volontairement les prédictions des taux de rendements des fonds Euro et UC à une année donnée pour voir comment nos prédictions sont impactées. Pour ça nous modifions les rendements de l'Euro et de l'UC de 2021 par un rendement du fond Euro de 4% et un rendement des UC de -8%. Voici les résultats observés avec cette hypothèse :

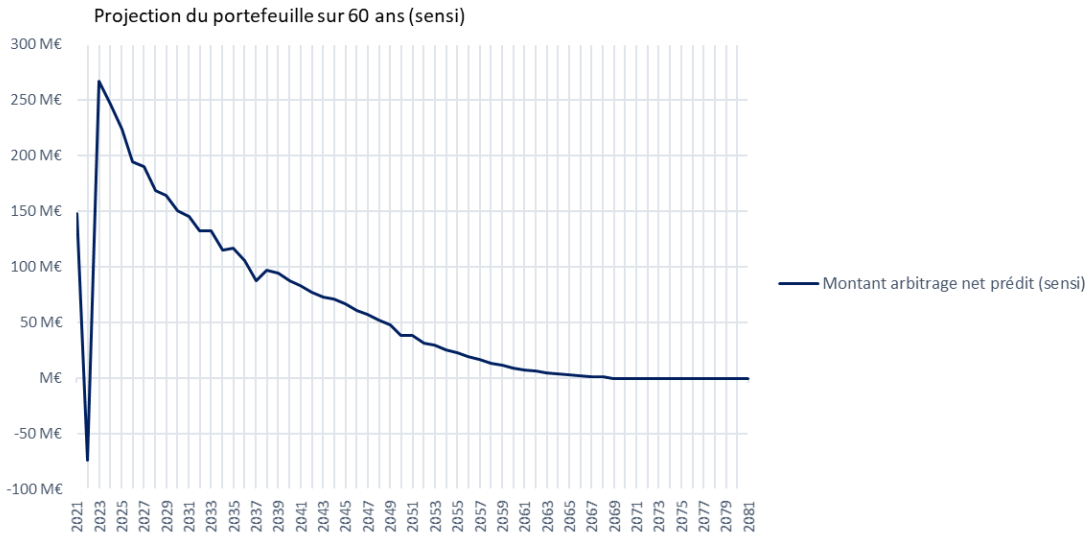


Figure 52 : Analyse de la sensibilité aux performances des fonds

L'année 2021 ne change pas, ce qui est logique étant donné que nous avons choisi un pas de temps annuel, le modèle n'est pas capable de réagir à ce qui se passe durant l'année en cours. Néanmoins, l'année suivante, le montant annuel arbitré (net vers UC) est d'environ -75 millions d'euros, signifiant que sur cette année-là, les arbitrages se sont faits majoritairement en direction du fond Euro. Le modèle réagit donc bien aux performances des différents fonds et semble agir de manière rationnelle en s'orientant vers celui offrant le plus de rendement.

En revanche, l'ampleur de cette réaction est discutable. Il aurait été légitime de penser que la réorientation vers l'Euro aurait été plus massive mais ce n'est pas le cas. De plus, le fait que toutes les autres années les arbitrages soient majoritairement en direction de l'UC nous laisse penser que le modèle est peut-être légèrement biaisé. Comme durant la période d'entraînement, il n'y a eu qu'une seule année où les arbitrages n'ont pas été majoritairement vers l'UC, le modèle considère sans doute que la norme est d'aller dans ce sens.

6. Impacts sur la modélisation prospective

Maintenant que les arbitrages sont intégrés dans le fichier, nous pouvons comparer les résultats des projections sur l'EV du portefeuille, ainsi que le TRI, avec et sans arbitrage.

L'EV (Embedded Value) d'une compagnie d'assurance vie correspond à la valeur intrinsèque d'une compagnie d'assurance. Cet indicateur s'est largement imposé sur le marché. Il est utilisé dans les communications financières internes comme base pour fixer le prix d'une compagnie assurance vie. Il se calcule comme la valeur actualisée des résultats futurs retraités du coût d'immobilisation des fonds propres.

Le TRI (Taux de Rendement Interne) est une métrique mesurant le taux de rendement d'un investissement. Il est défini comme le taux d'actualisation qui annule la somme des Cash-Flows. Dans notre cas ces cash flows sont simplement les résultats annuels réalisés. Comme nous projetons sur 60ans, on calcule le TRI de la façon suivante :

$$\sum_{t=0}^{60} \frac{\text{Résultat}_t}{(1+TRI)^t} = 0$$

Plus le TRI est élevé, plus la rentabilité de l'assureur est importante.

Ci-dessous les valeurs de ces indicateurs avec et sans les arbitrages.

Tableau 34 : Impacts sur la modélisation prospective

	Sans arbitrage	Avec arbitrage
VA produit assurance (1)	1740,46	1737,56
VA impots sur les sociétés (2)	184,21	197,12
VA charges (3)	1027,28	974,42
VA résultats futurs (4)=(1)-(2)-(3)	528,97	566,02
MS initial (5)	501,60	502,96
VA dotation MS (6)	-149,36	-156,65
VA intérêt ms (7)	207,96	203,81
VA IS intérêt ms (8)	53,71	52,64
Cout immo fp (9)=(5)+(6)-(7)+(8)	198,00	195,15
EV (10)=(3)-(9)	384,60	423,76
TRI (11)	8,47%	8,98%

L'intégration des arbitrages dans nos modélisations prospectives impacte de manière positive les résultats. Premièrement la valeur actuelle des résultats futurs augmente. Cette augmentation est causée en grande partie par la diminution des charges. En effet, comme nous l'avons vu plus haut, notre modèle prévoit majoritairement des arbitrages vers l'UC. Ces prédictions font donc diminuer la PM Euro au profit de la PM UC. Or environ 80% de la PM Euro est réassuré. Une diminution de la PM Euro engendre donc une diminution des charges de réassurances. Deuxièmement, on observe une légère baisse des coûts d'immobilisation en fonds propres. Cette baisse provient de la diminution de la Marge de Solvabilité (MS). En effet, la PM Euro qui n'est pas réassurée nécessite une immobilisation du capital plus importante que la PM UC[13]. De ce fait, la valeur actuelle de la dotation MS diminue. Cette baisse est atténuée par la diminution de la valeur actuelle des intérêts sur la Marge de Solvabilité. En effet, les fonds immobilisés génèrent malgré tous des (faibles) intérêts.

Finalement, ces deux effets vont dans le sens d'une augmentation de l'EV. De même, les effets énoncés précédemment participent également à l'augmentation du TRI.

[13] Suivant les années, le Besoin de Marge de Solvabilité (BMS) S2 Euro est d'environ 7%. Autrement dit, pour 100 de PM Euro, l'assureur doit immobiliser 7 pour être considéré comme solvable. Or, le BMS S2 UC est d'environ 2%

CONCLUSION

Avec la naissance des contrats multisupports, est apparue la possibilité d'arbitrer entre le fond Euro et les fonds UC. Ces derniers ont un impact sur la rentabilité et la solvabilité de l'assureur. Dès lors, la question se pose de savoir ce qui motive ces mouvements ? Comment les anticiper ? A quel point ces derniers impactent-ils la solvabilité et la rentabilité de l'assureur ?
C'est pour répondre à ces questions que nous avons tenté de modéliser ces flux.
La démarche qui s'est avérée être la plus pertinente a été de faire une modélisation individuelle avec un pas de temps annuel. Cette démarche a nécessité la construction d'une base de données adaptée. Il a d'abord fallu récupérer tous les arbitrages effectués sur notre période d'observation, puis dans un second temps récupérer toutes informations complémentaires (sens de l'arbitrage, pm au moment de l'arbitrage, frais, …). Ensuite trouver une agrégation pertinente pour synthétiser ces mouvements par année et par contrat.
L'approche dite par « fréquence sévérité probabiliste » a été la plus précise et robuste.
Plusieurs modèles ont été implémentés, en commençant par des modèles linéaires jusqu'aux agrégations de modèles non linéaires d'apprentissage automatique. Pour éviter le sur-apprentissage, les hyperparamètres ont été fixés en utilisant la validation croisée et les modèles seront testés sur les données de 2020 et 2021 qui n'ont pas servi à leur construction.
Pour modéliser la fréquence, le modèle retenu est une forêt aléatoire. Ce choix a été fait sur la base des métriques que nous avons jugés pertinentes et adaptées à une classification déséquilibrée (Log Loss, AUC P/R). Les variables qui semblent le plus jouer dans la fréquence d'arbitrage sont les variables liées aux comportements d'arbitrages passés (le nombre d'arbitrage l'année précédente, le nombre d'arbitrage déjà effectué) mais aussi la proportion d'UC dans le contrat, la PM. Globalement, le profil d'assuré avec une forte probabilité d'effectuer un arbitrage est un assuré avec beaucoup d'épargne sur son contrat, répartie de manière équilibrée entre le fond Euro et UC et qui a déjà beaucoup arbitré par le passé. Notons néanmoins qu'aucun modèle implémenté a permis d'obtenir des prédictions convenables.
Pour modéliser la « sévérité », le modèle retenu est également une forêt aléatoire, ce dernier s'étant démarqué dans les métriques d'évaluation retenues (R^2, MAE). Les variables qui semblent impacter le plus le sens de l'arbitrage ainsi que son ampleur sont la proportion d'UC dans le contrat, les plus ou moins-values réalisées sur chaque fond et l'écart entre ces deux valeurs. Globalement, un assuré ayant une petite partie de son épargne sur l'UC et qui a réalisé sur cette partie un taux de plus-value 2 à 3 fois supérieur que l'Euro a de grandes chances d'arbitrer une partie conséquente de son épargne vers de l'UC. Notons en revanche que même si ce modèle est de loin le meilleur de tous ceux implémenté, les scores obtenus aux métriques d'évaluation demeurent moyens. Notons également qu'aucune variable exogène ressort comme étant réellement importante. Il semble que dans notre modèle, l'impact du contexte économique et financier soit pris en compte à travers les performances du fond Euro et des fonds UC (qui sont directement impactés par ce contexte).
En globalité, du fait de l'approche choisie, les prédictions individuelles sont relativement mauvaises car nous prédisons une espérance d'un montant d'arbitrage et que ce dernier est un évènement rare. Toute la force de notre modélisation est d'arriver à converger vers la bonne valeur lorsqu'on agrège les résultats à l'année.

Nous avons ensuite fait évoluer notre modèle sur 60 ans et nous avons intégré ces prédictions dans un outil de modélisation actif/passif simplifié (déterministe). Le modèle prévoyant d'avantage d'arbitrages

vers l'UC, l'impact sur la solvabilité et la rentabilité de l'assureur est positif avec L'EV et le TRI qui augmente. Cet effet positif est lié à la PM Euro qui diminue au profit de la PM UC, faisant baisser mécaniquement les charges de réassurance et le coût d'immobilisation du capital.

Plusieurs choses auraient pu être faites pour améliorer cette étude mais n'ont pas pu être mises en place par manque de temps et/ou pour des raisons techniques.
Par exemple, il aurait été intéressant de refaire cette étude avec un pas de temps trimestriel ou mensuel, nous aurions sans doute pu améliorer notre compréhension des comportements d'arbitrage par rapport aux variables exogènes.
De plus, nous aurions aimé pouvoir intégrer notre loi d'arbitrage dans un outil de modélisation actif/passif stochastique, comme celui utilisé officiellement par AG2R LA MONDIALE. Ceci nous aurait permis de mesurer l'impact sur les indicateurs S2 comme le SCR ou le BE. Mais pour des raisons de temps de calcul il s'est avéré impossible d'implémenter pour chaque individu, le produit de deux prédictions faites par une forêt aléatoire (une centaine d'arbre chacun), pendant 60 ans et ceux pour des milliers de scénarios d'actifs différents.

BIBLIOGRAPHIE

John Nelder et Robert Wedderburn, « *Generalized Linear Models* », 1972.

Trevor Hasie, Gareth James, Robert Tibshirani et Daniela Witten, « *An Introduction To Statistical Learning : With Applications in R* », 2013.

Cédric Asfa « *Le modèle Logit : Théorie et application* » 2016.

Peter McCullagh et John Nelder, « *Generalized Linear Models* », 1989.

Leo Breiman, Jerome Friedman, Richard Olshen et Charles Stones, « *Classification And Regression Tree* », 1984.

WikiStat « *Arbres binaires de décision* », https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-cart.pdf, Université de Toulouse.

Leo Brieman, « *Bagging Predictors* », 1996.

Mehryar Mohri, Afshin Rostamizadeh et Ameet Talwalker, « Foundations of Machine Learning », 2018.

Marie Chavent, « Bagging et Forêts aléatoires », http://www.math.u-bordeaux.fr/~mchave100p/wordpress/wp-content/uploads/2013/10/Apprentissage_C4_print.pdf, Université de Bordeaux.

Gilles Hunault, « Découpage en classes et discrétisation », https://gilles-hunault.leria-info.univ-angers.fr/wstat/discr.php

AG2R La Mondiale Matmut, nouveau poids lourd de l'assurance [archive], la Tribune, 23 janvier 2019.

Omar Berrada, « *Modélisation des arbitrages dynamiques par approche machine learning* », 2016

Chloé Nicolas, « *Les arbitrages en assurance vie sont-ils soumis à un phénomène de contagion ?* », 2017

Sofiane Feniza, « *Modélisation du comportement d'arbitrage en assurance vie* », 2019

ANNEXES

AG2R LA MONDIALE

PERSONNE PHYSIQUE

DEMANDE D'ARBITRAGES OU DE TRANSFERTS

Contrat d'assurance vie/contrat de capitalisation

Le souscripteur/adhérent souhaite réaliser une opération sur son contrat. Certaines opérations en cours de contrat ont des conséquences notamment sur le plan fiscal dont il faut tenir compte pour en apprécier l'opportunité. Le souscripteur/adhérent connaît-il bien ces conséquences ? Si cela n'est pas le cas ou en cas de doute sur l'opportunité de l'opération demandée, il lui est recommandé de se rapprocher de son conseiller.

Souscripteur/adhérent

○ Madame ○ Monsieur

Nom : _____

Prénom(s) : _____

Tél. portable : _____

E-mail : _____

Titulaire(s) du contrat : _____

Co-souscripteur/co-adhérent

○ Madame ○ Monsieur

Nom : _____

Prénom(s) : _____

Tél. portable : _____

E-mail : _____

Titulaire(s) du contrat : _____

ci-après individuellement ou collectivement « le souscripteur/adhérent », demande à modifier son contrat comme suit, **en respectant les minima et les conditions inscrits dans la Proposition/le Projet de contrat d'assurance vie/de capitalisation qui lui a été remis(e) à la souscription/adhésion.**

INFORMATION SUR L'OPÉRATION D'ARBITRAGES OU DE TRANSFERTS

Le souscripteur/adhérent désinvestit son épargne et le repartit entre les orientations, les profils de gestion, les options de gestion et/ou les supports suivants[1] :

	Désinvestir :		**Réinvestir :**
	Montant en euros	En % de l'épargne du support	En % du désinvest. total
GESTION LIBRE / ORIENTATION LIBRE **Actif en euros**			
_____	_____	_____ %	_____ %
_____	_____	_____ %	_____ %
AG2R LA MONDIALE Croissance[1]	_____	_____ %	_____ %

A renseigner uniquement si premier investissement sur le support AG2R LA MONDIALE Croissance[2] :

• Sélection de l'échéance de la garantie en capital du support Croissance (minimum 8 ans, maximum 40 ans) : _ _ ans

• Sélection du niveau de la garantie en capital à l'échéance choisie ci-dessus du support Croissance : ○ 80% ○ 90% ○ 100%

Sélection d'unités de compte **Code ISIN** à renseigner obligatoirement	**Nom du support**			
_____	_____	_____	_____ %	_____ %
_____	_____	_____	_____ %	_____ %
_____	_____	_____	_____ %	_____ %
_____	_____	_____	_____ %	_____ %
_____	_____	_____	_____ %	_____ %

Page 1/2

Figure 53 : Formulaire à remplir pour effectuer un arbitrage (1)

	Désinvestir :		Réinvestir :
	Montant en euros	En % de l'épargne du support	En % du désinvest. total
PROFILS DE GESTION (uniquement en gestion libre) **Nom du/des profils**			
_____	_____	_____ %	_____ %
_____	_____	_____ %	_____ %
OPTIONS DE GESTION (uniquement en gestion libre)			
Réallocation programmée de l'épargne(3)	_____	_____ %	_____ %
Arbitrages automatiques	_____	_____ %	_____ %
Autres : (nom) : _____	_____	_____ %	_____ %
Le souscripteur/adhérent complète le bulletin de l'option correspondant et le joint au présent document.			
○ **ORIENTATION PERSONNALISÉE OU ○ ORIENTATION CONSEILLÉE**(4)			
Unité de compte monétaire de l'orientation	_____	_____ %	_____ %

(1) Si prévu dans la Proposition/le Projet de contrat d'assurance vie/de capitalisation.

(2) Les choix de l'échéance et du niveau de garantie du support Croissance sont définitifs. L'échéance de la garantie peut toutefois être prorogée par le souscripteur/adhérent à l'arrivée de celle-ci dans les conditions fixées au contrat.

(3) Les versements programmés ne sont pas compatibles avec les options de gestion « investissement progressif » et « sécurisation progressive de l'épargne ».

(4) L'Orientation personnalisée et l'Orientation conseillée ne sont pas compatibles.

Commentaires :

_____

_____

_____

_____

SIGNATURE(S)

Le souscripteur/adhérent déclare être informé que l'épargne constituée sur les supports libellés en unités de compte ne bénéficie d'aucune garantie en capital de la part de l'assureur. L'entreprise d'assurance ne s'engage que sur le nombre d'unités de compte mais pas sur leur valeur. La valeur des unités de compte, qui reflète la valeur des actifs sous-jacents, n'est pas garantie mais est sujette à des fluctuations à la hausse ou à la baisse dépendant en particulier de l'évolution des marchés financiers.

Le souscripteur/adhérent reconnaît que toute opération de gestion sur son contrat ne pourra être demandée par lui qu'au terme du délai de renonciation et sous réserve de réception par l'assureur de la preuve que le souscripteur/adhérent ait été informé que le contrat est conclu.

En cas de souscription/adhésion conjointe, la présente demande doit obligatoirement être signée par l'ensemble des co-souscripteurs/ co-adhérents du contrat, lesquels reconnaissent et acceptent la modification demandée.

Fait à _____ Le _____ en 3 exemplaires*

Code + cachet du conseiller	SIGNATURE(S) (précédée(s) de la mention "lu et approuvé")
	Le souscripteur/adhérent — Le co-souscripteur/co-adhérent

*Exemplaire : Assureur - Conseiller - Souscripteur

Merci de parapher chaque page du présent document

Figure 54 : Formulaire à remplir pour effectuer un arbitrage (2)

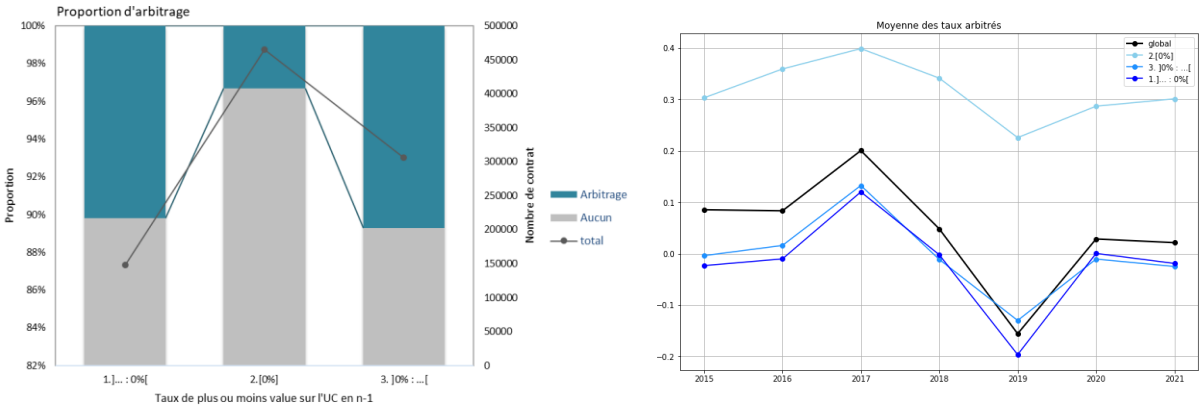


Figure 55 : Variables cible en fonction de la plus ou moins-value sur l'UC en n-1

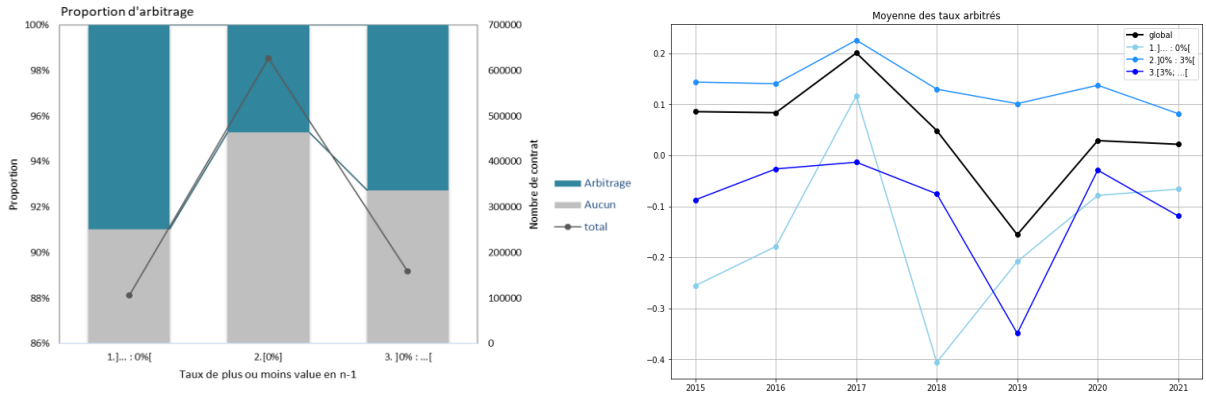


Figure 56 : Variables cible en fonction de la plus ou moins-value en n-1

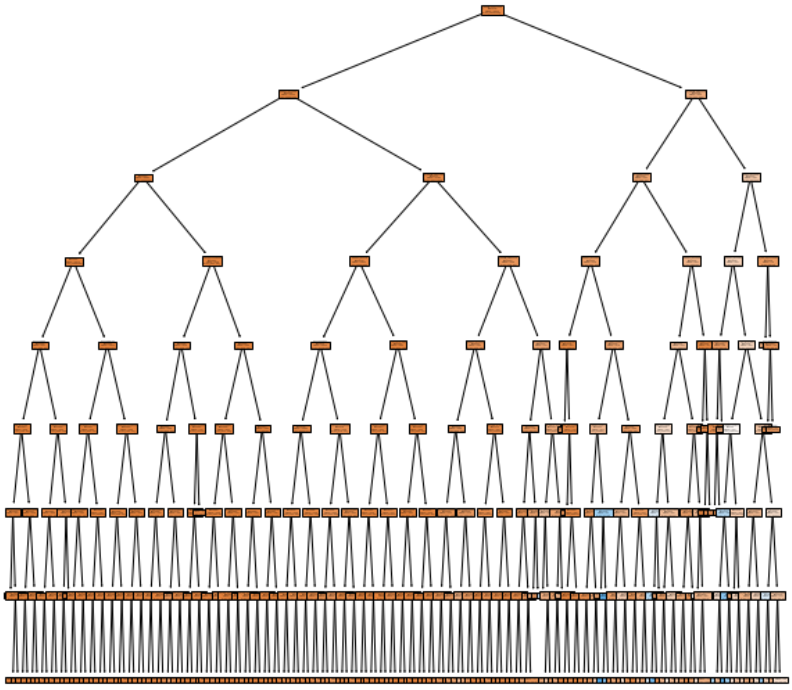


Figure 57 : Allure du modèle fréquence CART

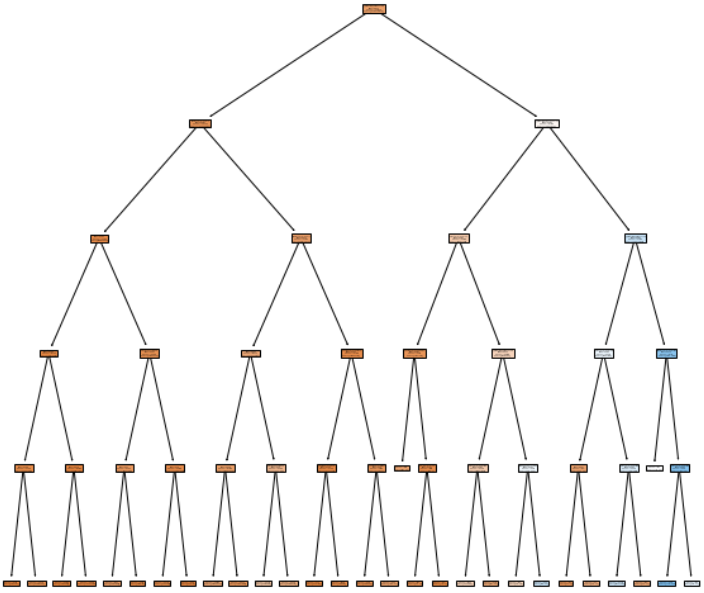


Figure 58 : Allure du modèle fréquence rééchantillonné CART

	annee	montant_arbitre	pred_montant_arbitre
0	2015	1.911201e+08	1.045781e+08
1	2016	1.384965e+08	1.449298e+08
2	2017	3.626239e+08	4.990519e+08
3	2018	4.070175e+07	1.174996e+08
4	2019	-2.797996e+08	-7.460074e+08
5	2020	1.795874e+08	8.535958e+07
6	2021	1.349113e+08	2.411152e+08

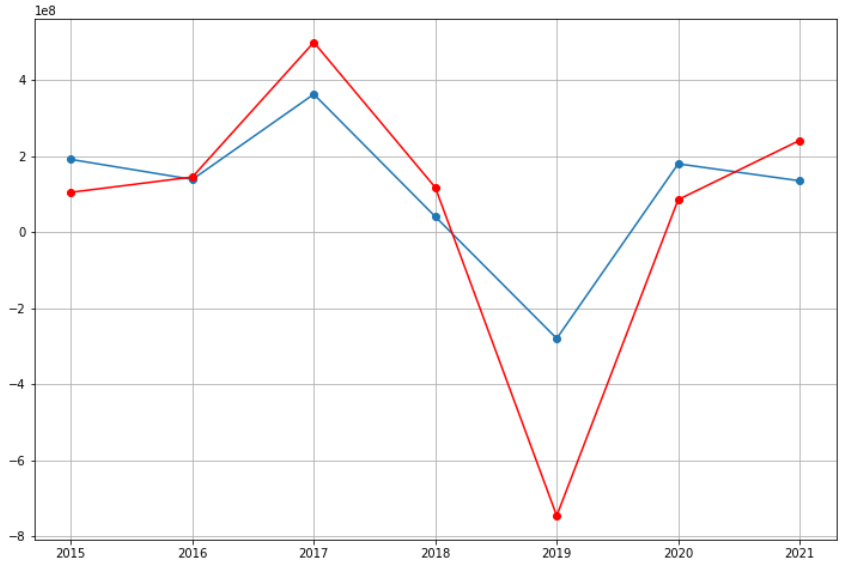


Figure 59 : Résultat des prédictions avec la modélisation fréquence/sévérité rééchantillonné 70/30 (sortie python)

	annee	montant_arbitre	pred_montant_arbitre
0	2015	1.911201e+08	1.101307e+08
1	2016	1.384965e+08	1.157708e+08
2	2017	3.626239e+08	3.041220e+08
3	2018	4.070175e+07	9.232545e+07
4	2019	-2.797996e+08	-3.839134e+08
5	2020	1.795874e+08	5.401422e+07
6	2021	1.349113e+08	1.224821e+08

Figure 60 : Résultat des prédictions avec la modélisation fréquence/sévérité rééchantillonné 75/25 (sortie python)