Institut des ACTUAIRES

ENSAE

IP PARIS

Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 10/11/2021

Par : **Hayam Agerouassi**

Titre : **Impact de la réglementation australienne PYS sur le comportement des assurés**

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

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

Membres présents du jury de la filière

Nom : PICARD Pierre

Membres présents du jury de l'Institut des Actuaires

Entreprise : AXA France

Signature :

Directeur du mémoire en entreprise :

Nom : WU Hao

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

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé :

Ce travail s'inscrit dans le cadre de l'étude de l'impact d'une nouvelle réglementation sur le marché australien des pensions de retraite ; cette réglementation, appelée PYS, concerne les membres qui sont inactifs pendant une certaine période. Ces membres perdent leur couverture d'assurance, sauf s'ils demandent expressément de la conserver.

L'objectif de cette étude est d'abord de vérifier la présence d'un phénomène d'anti-sélection qui a fait l'objet d'un chargement tarifaire spécifique par le réassureur. Pour ce faire, nous examinons les caractéristiques de la population qui s'est manifestée pour rester couverte et nous comparons les taux d'incidence des sinistres avant et après le changement de réglementation. Deuxièmement, nous voulons modéliser et identifier les facteurs qui influencent le choix de rester couvert, afin de pouvoir généraliser les résultats pour le PMIF ; une autre réglementation qui succède au PYS.

Nous avons réussi à montrer dans un premier lieu que la population voulant rester couverte est plus âgée et plus risquée que la population non concernée par ce changement. Nous avons également utilisé une régression logistique pour détecter la présence d'une information cachée non visible à l'assureur sur la sinistralité chez les opt-in.

Pour la modélisation du comportement des assurés, nous avons testé plusieurs méthodes de classification classiques et du machine learning, nous avons également utilisé les techniques de sous-échantillonnage pour pallier le déséquilibre des données. Nous avons choisi un modèle Xg-boost pour la modélisation. Nous avons ensuite testé la robustesse de nos résultats à l'aide de deux autres fonds de retraites, nous avons également appliqué notre modèle dans le cadre du PMIF.

Les principales limites de ce travail sont les hypothèses faites pour le calcul du taux d'incidence avant et après le PYS, et la difficulté d'interpréter les résultats du Xg-boost. Nous avons utilisé le graphe d'importance des variables, le graphique de dépendance partielle ainsi que LIME (Local interpretable Model-agnostic Explanations) pour résoudre la

dernière limite.

Mots clés : *Retraite Australienne, modélisation du comportement des assurés, Asymétrie d'information, Antisélection, Xg boost, régression logistique, LIME, dépendance partielle.*

Abstract

This work is part of a study of the impact of a new regulation on the Australian superannuation market ; this regulation, called PYS, affects members who are inactive for a certain period. These members lose their insurance coverage unless they specifically request to keep it.

The objective of this study is first to verify the presence of an anti-selection phenomenon that has been subject to specific rate loading by the reinsurer. To do so, we examine the characteristics of the population that has come forward to remain covered and compare claim incidence rates before and after the regulatory change. Second, we want to model and identify the factors that influence the choice to remain covered, so that we can generalize the results for the PMIF ; another successor regulation to the PYS.

We were able to show first that the population wanting to stay covered is older and riskier than the population not affected by this change. We also used logistic regression to detect the presence of hidden information not visible to the insurer on the claims experience of the opt-in population.

To model the behavior of the insureds, we tested several classical classification methods and machine learning. We also used subsampling techniques to compensate for the imbalance in the data. We chose an Xg-boost model for the modeling. We then tested the robustness of our results using two other pension funds, we also applied our model in the PMIF framework.

The main limitations of this work are the assumptions made in calculating the incidence rate before and after PYS, and the difficulty of interpreting the Xg-boost results. We used the variable importance plot, the partial dependence plot, as well as LIME (Local interpretable Model-agnostic Explanations) to address the latter limitation.

Key-words : *Australian superannuation, modelling of policyholder behaviour, informa-*

tion asymmetry, anti-selection, Xg boost, logistic regression, LIME, partial dependence.

Note de Synthèse

Problématique et contexte

Les changements réglementaires sont très courants dans le domaine de l'assurance, et l'assurance australienne ne fait pas exception. En 2019, le régulateur australien a modifié les règles d'éligibilité aux garanties de décès, d'invalidité et d'incapacité temporaire liées aux membres des fonds de pension de retraite. À partir de juillet 2019, les membres dont les comptes n'ont pas reçu de cotisation depuis 16 mois perdront leurs couvertures d'assurance, à moins qu'ils ne s'expriment pour les conserver, auquel cas nous les appellerons les opt-ins. Cette réglementation s'appelle le PYS : Protect Your Super.

	chargement	d'expérience	Chargement pure dû au	mathématique PYS	chargement l'antisélection	dû à PYS	Chargement	total
	Death et TPD	IP	Death et TPD	IP	Death et TPD	IP	Death et TPD	IP
Fonds 1	+7,1%		+2,5%		+3,1%		+13,2%	
Fonds 2			-5,9%	+10,1%	+0,0%	+0,0%	-5,9%	+10,1%
Fonds 3			+5,1%	+24,5%	+8,9%	+8,9%	+14,5%	+35,9%
Fonds 4	+33,8%	+41,3%	-7,0%	+21,0%	+22,0%	+1,0%	+51,8%	+72,6%

Table 1 – Taux de chargement appliqué suite au PYS sur les 4 fonds étudiés

En prévision des conséquences de ce changement, ALHIS a appliqué un chargement tarifaire aux cédantes. Ce chargement est composé d'un chargement mathématique, qui prend en compte le changement observable du profil de risque du portefeuille, et d'un chargement d'antisélection, qui prédit la présence d'antisélection.

L'objectif de cette étude est d'abord de vérifier la présence d'antisélection et la justesse du chargement appliqué. Le PYS a été succédé par le PMIF en 2020, il s'agit d'un autre changement réglementaire de même nature. Il serait donc intéressant de modéliser le comportement des assurés en matière de faire un opt-in afin de prédire l'effet du PMIF.

Ceci nous amène à la deuxième partie de l'objectif de cette étude qui est de modéliser le comportement des assurés et de comprendre les facteurs qui influencent leurs choix.

Analyse des taux d'opt-in et antisélection

Pour faciliter le traitement du sujet, nous avons défini la terminologie utilisée. Nous distinguons 3 catégories de membres : Actifs, Inactifs et Opt-ins. Nous avons également présenté les données disponibles. Ces données sont issues de 4 fonds et contiennent des informations sur les assurés, leurs sinistres ainsi que des indicatrices permettant de repérer s'ils sont des opt-ins.

Nous avons montré la croissance des taux d'opt-in en fonction de l'âge, l'impact du sexe sur ce taux, les femmes faisant généralement plus d'opt-ins, ainsi que la distribution par âge de chacune de nos catégories d'intérêt, les opt-ins étant généralement plus âgés que les deux autres catégories.

Par la suite, le phénomène d'anti-sélection que nous voulons mettre en évidence a été défini. Nous avons effectué une régression logistique afin d'identifier la présence d'informations cachées parmi les opt-ins. Nous voulions dissocier l'effet des caractéristiques observables par l'assureur et les caractéristiques inobservables qui, selon nous, sont cachées dans le fait qu'un individu est un opt-in sur sa sinistralité.

Nous avons réussi à trouver l'impact positif, indépendant de ce qui est observable par l'assureur, que la variable indicatrice opt-in exerce sur la sinistralité.

Puis, Nous avons calculé les taux d'incidence des sinistres décès, TPD et IP avant et après la PYS. Nous avons constaté que les personnes ayant choisi de garder leurs couvertures d'assurance avaient un taux plus élevé que les actifs, ce qui a entraîné des changements dans les taux globaux. Nous avons comparé l'impact de PYS sur l'incidence des sinistres avec le taux de chargement appliqué.

Fonds	PYS Impact on GL Frequency	GL Premium Loading for PYS	PYS Impact on IP Frequency	IP Premium Loading for PYS
Fonds 1	+12%	+13%		
Fonds 2	-14%	-6%	+2%	+10%
Fonds 3	+1%	+14%	+4%	+36%
Fonds 4	+34%	+52%	+33%	+22%

TABLE 2 – Récapitulatif et comparaison de l'impact PYS sur les taux d'incidence avec les taux de chargement appliqués

Modélisation du comportement des assurés

Nous sommes passés ensuite à la modélisation du comportement d'opt-in chez les individus étudiés. En pratique, nous avons testé plusieurs modèles de classification en les comparant entre eux à l'aide du f1-score et AUC. Nous avons utilisé les techniques du sous-échantillonnage pour équilibrer les modèles et améliorer leurs scores. Nous avons choisi le modèle Xg-boost pour modéliser ce comportement. En effet, ce modèle, tout en étant performant, manque d'interprétabilité. Nous avons utilisé alors le graphe d'importance des variables, le graphique de dépendance partielle ainsi que LIME pour interpréter ses résultats.

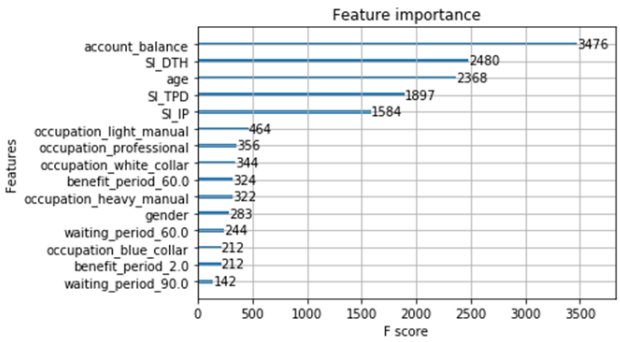

FIGURE 1 – Le graphe de l'importance des variables selon le XG-boost

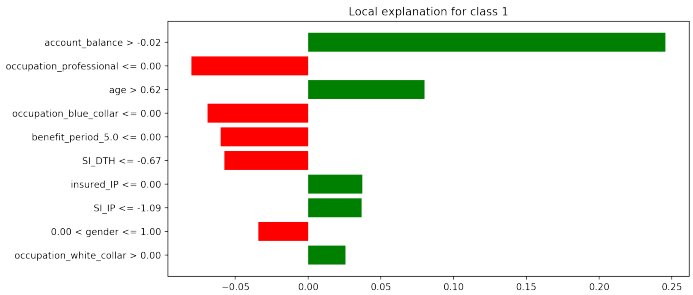

FIGURE 2 – Application de la méthode LIME sur un individu pour le modèle Xg-boost optimisé et entraîné sur la base sous-séchantillonnée.
En rouge : contredit l'opt-in. En vert : soutient l'opt-in.

Application des résultats

Nous avons appliqué notre modèle dans le but de pouvoir prédire l'impact du changement réglementaire PMIF. Ainsi, nous avons analysé les résultats obtenus et jugé le fonds 3 comme le fonds le plus impacté par le PMIF.

Fonds	Taux d'opt-in PMIF prédit	Taux d'opt-in PYS
Fonds 1	11,6%	11%
Fonds 2	2,8%	23%
Fonds 3	41,6%	24%
Fonds 4	27,4%	38%

TABLE 3 – Taux d'opt-in pour le PMIF prédit par le modèle et taux d'opt-in PYS observé par fond

Dans une seconde partie, nous avons testé notre modèle entraîné sur les fonds 1, 2, 3, 4 avec le comportement d'opt-in PYS observé sur deux nouveaux fonds 5 et 6. D'abord, nous avons analysé les taux d'opt-ins réels observés. Puis, nous les avons comparés avec ce que notre modèle prédit. Enfin nous avons proposé quelques pistes d'amélioration pour encore polir le pouvoir généralisateur du modèle.

Fonds	Taux d'opt-in réel	Taux d'opt-in prédit	Score AUC	f1 score	f1 score dummy
Fonds 5	36,34%	44,87%	0,6039	0,5148	0,5018
Fonds 6	18,68%	18,32%	0,6102	0,3634	0,3148

TABLE 4 – Résultats et score du modèle pour la prédiction du PYS opt-ins dans les fonds 5 et 6

Conclusion et perspectives

Les résultats de cette étude ont été utilisés pour justifier la surtarification imposée après le PYS, et feront partie du Framework adopté pour comprendre les circonstances des changements réglementaires en superannuation australienne comme le PMIF.

Cette étude peut être améliorée de plusieurs manières, notamment par le biais de l'utilisation de plus de données pour renforcer notre compréhension du comportement des assurés. Nous pouvons également proposer de nouvelles estimations des taux IBNR qui ne réagissent pas excessivement aux déclarations plus rapides par les assurés.

Executive Summary

Context

Regulatory changes are very common in the insurance industry, and Australian insurance is no exception. In 2019, the Australian regulator changed the eligibility rules for death, terminal and permanent disability and income protection benefits related to superannuation fund members. From July 2019, members whose accounts have not received a contribution for 16 months will lose their insurance coverages, unless they expressingly ask to keep them, in which case we will call them opt-ins. This regulation is called PYS : Protect Your Super.

	chargement	d'expérience	Chargement pure dû au	mathématique PYS	chargement l'antisélection	dû à PYS	Chargement	total
	Death et TPD	IP	Death et TPD	IP	Death et TPD	IP	Death et TPD	IP
Fonds 1	+7,1%		+2,5%		+3,1%		+13,2%	
Fonds 2			-5,9%	+10,1%	+0,0%	+0,0%	-5,9%	+10,1%
Fonds 3			+5,1%	+24,5%	+8,9%	+8,9%	+14,5%	+35,9%
Fonds 4	+33,8%	+41,3%	-7,0%	+21,0%	+22,0%	+1,0%	+51,8%	+72,6%

Table 5 – Loading rate applied following the PYS on the 4 funds studied

In anticipation of the consequences of this change, ALHIS applied a tariff loading to cedants. This loading is composed of a mathematical loading, which takes into account the observable change in the risk profile of the portfolio, and an anti-selection loading, which predicts the presence of anti-selection.

The aim of this study is first to verify the presence of anti-selection and the correctness of the loading applied. The PYS has been succeeded by the PMIF in 2020, another change of the same nature. Therefore, it would be interesting to model the behavior of policyholders with respect to making an opt-in in order to predict the effect of the PMIF. This brings

us to the second part of the objective of this study, which is to model the behavior of policyholders and understand the factors that influence their choice.

Analysis of opt-in and anti-selection rates

To facilitate the treatment of the subject, we have defined the terminology used. We distinguish 3 categories : Active, Inactive and Opt-in members. We have also presented the available data. These data come from 4 funds and contain information on the insureds, their claims as well as indicators to identify if they are opt-ins.

We showed the growth of opt-in rates as a function of age, the impact of gender on this rate, with women generally making more opt-ins, as well as the age distribution of each of our categories of interest, with opt-ins generally being older than the other two categories.

We then defined the phenomenon of anti-selection that we want to exhibit. We performed a logistic regression in order to identify the presence of hidden information among the opt-ins. We wanted to dissociate the effect of the characteristics observable by the insurer and the inobservable characteristics that we believe are hidden in the fact that an individual is an opt-in on their claim experience.

We were able to find the positive impact, independent of what is observable by the insurer, that the opt-in indicator variable has on claims experience.

We then calculated the incidence rates of Death, TPD and IP claims before and after PYS. We found that the opt-ins have a higher rate than the actives which triggered changes in the overall rates. We compared the impact of PYS on the incidence of claims with the loading rate applied.

Fonds	PYS Impact on GL Frequency	GL Premium Loading for PYS	PYS Impact on IP Frequency	IP Premium Loading for PYS
Fonds 1	+12%	+13%		
Fonds 2	-14%	-6%	+2%	+10%
Fonds 3	+1%	+14%	+4%	+36%
Fonds 4	+34%	+52%	+33%	+22%

Table 6 – Summary and comparison of PYS impact on incidence rates with applied loading rates

Modeling of policyholder behavior

We then proceeded to model the opt-in behavior of the individuals studied. We tested several classification models by comparing them with each other using the f1-score and AUC. We used undersampling techniques to balance the models and improve their scores. We chose the Xg-boost model to model this behavior. This model, while performing well, lacks interpretability. We hence used the importance graph of the variables, the partial dependence graph and LIME to interpret its results.

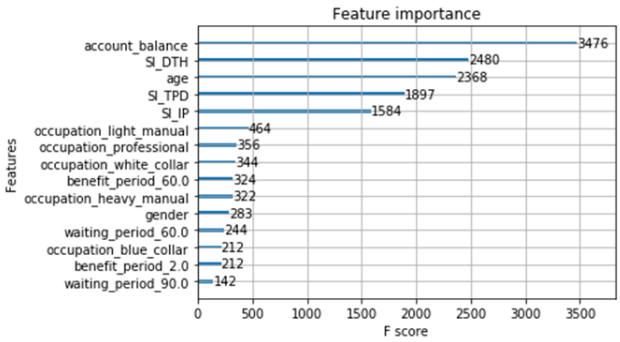

FIGURE 3 – The variable importance graph according to the XG-boost

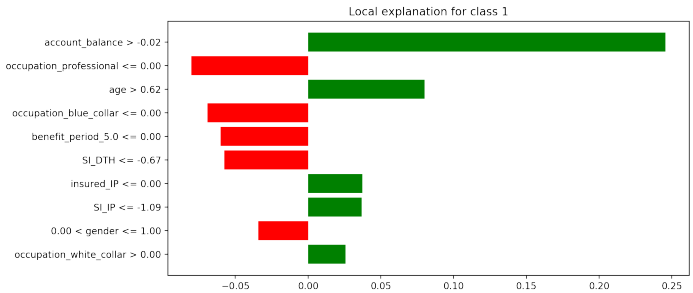

FIGURE 4 – Application of the LIME method on an individual for the optimized Xg-boost model trained on the undersampled database.
In red : contradicts the opt-in. In green : supports the opt-in.

Results' application

We applied our model with the goal to predict the impact of the PMIF regulatory change. We analyzed the results obtained and judged fund 3 as the most impacted fund by PMIF.

Fonds	Taux d'opt-in PMIF prédit	Taux d'opt-in PYS
Fonds 1	11,6%	11%
Fonds 2	2,8%	23%
Fonds 3	41,6%	24%
Fonds 4	27,4%	38%

TABLE 7 – PMIF opt-in rate predicted by the model and PYS opt-in rate observed by fund

In a second part, we compared our model trained on funds 1, 2, 3, 4 with the observed PYS opt-in behavior on two new funds 5 and 6. First, we analyzed the actual observed opt-in rates. Then, we compared them with what our model predicts. Finally, we have proposed a few ways to improve the generalizability of the model.

Fonds	Taux d'opt-in réel	Taux d'opt-in prédit	Score AUC	f1 score	f1 score dummy
Fonds 5	36,34%	44,87%	0,6039	0,5148	0,5018
Fonds 6	18,68%	18,32%	0,6102	0,3634	0,3148

TABLE 8 – Results and model score for predicting PYS opt-ins in funds 5 and 6

Conclusion and perspectives

The results of this study were used to justify the overcharging imposed after PYS, and will form part of the Framework adopted to understand the circumstances of regulatory changes in Australian superannuation such as PMIF.

This study can be improved in a number of ways, including using more data to strengthen our understanding of policyholder behavior. We can also propose new estimates of IBNR rates that do not overreact to earlier reporting by policyholders.

Remerciements

Je tiens à remercier toutes les personnes ayant contribué au bon déroulement de cette étude et qui m'ont aidée pendant la rédaction de ce mémoire.

Mes remerciements s'adressent d'abord au service des relations entreprises et des stages, pour leur disponibilité, leurs encouragements et leur accompagnement lors de la recherche de l'alternance.

Je tiens à exprimer toute ma reconnaissance à Mme Maryam Salehi pour m'avoir accordée cette opportunité. Sa disponibilité et sa confiance, ainsi que ses encouragements enthousiastes m'ont permis de progresser sans cesse pendant l'alternance.

J'adresse également mes sincères remerciements à Mr Hao Wu, mon tuteur chez Axa. En effet, sa disponibilité, son engagement dans notre projet et sa disposition inconditionnelle pour suivre minutieusement mes avancées et répondre à mes questions m'ont motivée d'avantage et m'ont permis de ne jamais lâcher prise de mes missions.

Je remercie très chaleureusement Mr Pierre Picard, mon référent pédagogique à l'ENSAE, pour ses sages recommandations et conseils. Son suivi m'a été très précieux et instructif.

Je remercie tout particulièrement et je témoigne toute ma reconnaissance à l'égard de l'équipe Technical Efficiency de ALHIS, avec qui j'ai eu l'opportunité d'effectuer mon alternance, pour leur accueil cordial, le partage de leurs expériences, l'effort qu'ils consacrent pour me faire sentir parfaitement en harmonie avec eux et finalement pour leur bonne humeur. Grâce à eux, j'ai pu vivre pleinement cette expérience.

Enfin mes remerciements vont à ma famille, mes amis et toutes les personnes qui m'ont aidée et encouragée lors de la réalisation de ce travail.

Liste des sigles et acronymes

PYS *Protect Your Super*

PMIF *Putting Members' Interest First*

IP *Income protection - La protection du salaire*

TPD *Total and Permanent Disability*

Table des matières

Introduction **20**

1 Contexte réglementaire australien et cadre de l'étude **22**

1.1 Les fonds de retraite ou superannuation funds 22

1.1.1 Principe de fonctionnement 22

1.1.2 Répartition par secteur 23

1.1.3 Cadre réglementaire 24

1.1.3.1 Superannuation Industry (Supervision) Act 1993 24

1.1.3.2 Financial Services Reform Act 2002 24

1.1.3.3 Superannuation Guarantee (Administration) Act 1992 . . 24

1.1.3.4 Autorités de régulation 24

1.2 La prévoyance en Australie 25

1.2.1 Death 26

1.2.2 Total and Permanent Disability 26

1.2.3 Income Protection 26

1.2.3.1 Waiting Period 27

1.2.3.2 Benefit Period 27

1.3 Changement réglementaire 27

1.3.1 PYS - Protect Your Superannuation 27

1.3.2 PMIF- Putting Members' Interest First 28

1.3.3 Chargement mathématique et d'antisélection 29
1.3.3.1 Chargement mathématique 29
1.3.3.2 Chargement d'antisélection 30

2 Présentation des données et statistiques descriptives 31
2.1 Terminologie utilisée 31
2.2 Présentation des bases "membership" et sinistres 32
2.2.1 Base membership 33
2.2.1.1 Variables caractéristiques 33
2.2.1.2 Variables PYS 34
2.2.2 Base sinistre 35
2.2.3 Période d'analyses 36
2.3 Statistiques descriptives 36
2.3.1 Profil général des fonds étudiés 36
2.3.2 Les différences à noter entre les fonds 37

3 Taux d'opt-in et illustration de l'antisélection 39
3.1 Calcul des taux d'opt-in et comparaison des fonds étudiés 39
3.1.1 Impact de l'âge et le genre sur le taux d'opt-in 40
3.1.1.1 L'évolution du taux d'opt-in avec l'âge 40
3.1.1.2 Caractéristiques de la population des opt-in 41
3.2 Asymétrie d'information et antisélection 42
3.2.1 L'asymétrie d'information 42
3.2.2 L'antisélection en assurance 42
3.2.3 L'antisélection liée au PYS 43
3.2.4 Aléa moral et fraude 44
3.2.4.1 Aléa moral 44
3.2.4.2 Fraudes 44

3.3 Régression logistique pour déduire l'antisélection 45
3.3.1 Significativité et interprétation des coefficients 46
3.3.2 Discussion des résultats 47
3.4 Comparaison du taux d'incidence des sinistres 51
3.4.1 Hypothèses 51
3.4.2 Limites des hypothèses 53
3.4.3 Evolution des taux d'incidences par année et par risque 54
3.4.3.1 Décès 54
3.4.3.2 TPD 54
3.4.3.3 IP 55

4 Modélisation du comportement des assurés 58
4.1 Problème de classification 58
4.1.1 Analyse des variables 58
4.1.1.1 Analyse de corrélation entre les variables qualitatives . . . 58
4.1.1.2 Analyse de corrélation entre les variables quantitatives . . 59
4.1.1.3 Analyse de corrélation entre les variables qualitatives et quantitatives 61
4.1.2 Pré-traitement des données 62
4.1.3 Etat de l'art des méthodes de classification 62
4.1.3.1 Régression logistique 62
4.1.3.2 Arbre de décision 63
4.1.3.3 Extreme Gradient Boosting ou Xg-boost 64
4.2 Imbalanced learning 67
4.2.1 Rééchantillonnage 67
4.2.1.1 Echantillonnage aléatoire/ Random sampling 69
4.2.1.2 SMOTE 69
4.2.1.3 Liens de TOMEK ou TOMEK LINKS 70

4.2.1.4 Smote-TL 70
4.2.2 Méthodes d'évaluation 70
4.2.2.1 Matrice de confusion 71
4.2.2.2 F1-score 72
4.2.2.3 Le choix et le calcul du f1 score 73
4.2.2.4 ROC et AUC 74
4.3 Modèle retenu et résultats 75
4.3.1 Illustration des résultats 76
4.3.1.1 Premiers résultats 76
4.3.1.2 Impact du sous-échantillonnage 77
4.3.2 Vérification de la cohérence 78
4.3.2.1 Comparaison des variables les plus importantes 79
4.3.2.2 Partial Dependence Plot 79
4.3.2.3 LIME 83

5 Application de l'étude 86
5.1 Prédiction du PMIF opt-in 86
5.1.1 Calcul des impactés 86
5.1.2 Application du modèle et résultats 88
5.2 Prédiction des PYS-opt-in d'autres fonds de superannuation 88
5.2.1 Les fonds 5 et 6 88
5.2.2 Adaptations et pré-traitement 89
5.2.3 Résultats de la prédiction 90
5.2.4 Pistes d'amélioration 90

Conclusion 92

Bibliographie 94

Introduction

Les changements réglementaires font partie du quotidien des assureurs. Ces changements peuvent avoir des impacts très conséquents sur toute l'activité assurantielle, d'où l'importance d'en étudier les conséquences.

Mon alternance chez AXA France s'est déroulée du 05/10/2020 jusqu'au 01/10/2021. Durant cette alternance, j'ai travaillé sur un changement réglementaire impactant le marché australien de superannuation dont une partie est réassurée par ALHIS (AXA Life and Health International Solutions).

En 2019, le régulateur australien a modifié les conditions d'éligibilité aux garanties décès, invalidité et incapacité temporaire liées aux membres des fonds de « superannuation ». Depuis juillet 2019, les membres qui n'ont pas contribué à leurs comptes depuis plus de 16 mois perdent automatiquement leurs couvertures, sauf s'ils expriment leur volonté de les garder. Cette réglementation est appelée PYS : Protect Your Super.

Cette situation a créé un déséquilibre d'information entre les assureurs et les assurés et pourrait donner lieu à un phénomène d'antisélection. En anticipation des conséquences de ce changement, ALHIS a appliqué un chargement tarifaire aux cédantes, ce chargement se décompose en un chargement mathématique, qui prend en compte le changement observable du profil de risque du portefeuille et un chargement d'antisélection, qui prédit la présence d'antisélection.

En plus du PYS, le PMIF est un autre changement réglementaire qui a pris effet en 2020. Ce changement modifie d'avantage les règles d'éligibilité aux garanties précédentes. Cependant, nous ne disposons pas de suffisamment de données pour pouvoir identifier cette fois-ci ceux qui ont choisi de garder leur couverture d'assurance.

C'est dans ce contexte que s'inscrit cette étude, dont l'objectif est double : d'une part analyser l'effet de cette réglementation sur la sinistralité des portefeuilles étudiés et dé-

tecter la présence éventuelle d'antisélection et d'autre part modéliser le comportement des individus qui choisissent de garder leur couverture versus ceux qui ne le choisissent pas.

D'abord une étude de l'environnement des « Superannuations » en Australie a été nécessaire. Dans cette même partie, nous avons également présenté le cadre réglementaire et les garanties proposées avec les comptes de superannuation. Ensuite, nous avons présenté les données qu'on doit traiter et exploiter. Par la suite, nous avons fait quelques illustrations de l'impact de la nouvelle réglementation sur le portefeuille et la sinistralité. Une régression logistique a été effectuée dans le but d'illustrer la présence d'informations cachées, ces informations cachées sont indicatrices d'un risque plus élevé. Conformément à la deuxième moitié de l'objectif, nous sommes passés à la modélisation du comportement des assurés à l'aide des méthodes classiques et du Machine Learning. Pour enfin terminer sur une application de cette étude dans le cadre du PMIF et une généralisation des résultats obtenus sur d'autres fonds.

Chapitre 1

Contexte réglementaire australien et cadre de l'étude

Nous allons présenter dans ce chapitre l'environnement des fonds de retraite en Australie. En effet, cet environnement est bien différent du système français de retraite de base. Nous allons alors expliquer le principe de fonctionnement des fonds de superannuation et le corps réglementaire qui les cadre. Nous allons ensuite étudier la prévoyance en Australie et ses produits de base qui sont souvent attribués avec les comptes de superannuation. Finalement, nous allons expliquer le PYS et le PMIF qui font l'objet de cette étude.

1.1 Les fonds de retraite ou superannuation funds

1.1.1 Principe de fonctionnement

La Superannuation, ou « super » est la caisse de retraite australienne. C'est un système de retraite par capitalisation qui est différent de celui de France où le régime de base est par répartition. En Australie, chacun est responsable de sa propre retraite et possède un compte dédié et bloqué.

Il s'agit d'un système soutenu par le gouvernement australien et il concerne tous les salariés qui gagnent plus de 450 AUD par mois. Les employeurs sont dans l'obligation de verser un montant déterminé sur les comptes de Superannuation des employés, ces derniers sont aussi encouragés à cotiser eux-mêmes dans le cadre de ce qui est appelé «

salary sacrifice arrangements ». Depuis juillet 2002, le taux de cotisation a augmenté de 9% à 9,5% en juillet 2020. Celui-ci est prévu d'être capé à 12% aux horizons 2027. Le rythme des contributions de la part des employeurs est trimestriel. C'est-à-dire, tous les trois mois, l'employeur verse le montant sur le compte de Superannuation des employés, ce montant est soumis à une taxe de la part de l'état qui s'élève à 15%.

1.1.2 Répartition par secteur

Il existe à peu près 500 fonds de superannuation en Australie, dont 362 fonds avec des actifs qui dépassent en valeur 50 million dollars. En total, les actifs de la superannuation ont atteint 2.7 trillions dollars en fin juin 2018, un montant qui rend l'Australie le 4ème pays du monde en termes d'actifs détenus par les fonds de retraite.

Les fonds sont souvent répartis par secteur et sont destinés à des publics spécifiques. Les principaux secteurs représentés sont :

— Retail funds (fonds de détail) : Ces fonds sont ouverts à tous et sont gérés par des sociétés d'investissements, des institutions financiers ou des assurances. La société propriétaire du fonds dégage une marge de profit sur les investissements.
— Industry funds (fonds des industries) : Les plus grands fonds des industries ne sont pas restreints, ils sont ouverts à tous tandis que certains sont réservés uniquement aux salariés d'une industrie particulière. A la différence de « Retail Funds », ce sont des fonds à but non lucratif, les membres sont les seuls à bénéficier des retours sur investissements.
— Public Sector funds (fonds du secteur public) : ce sont des fonds destinés aux employés des ministères du gouvernement fédéral, des États ou du Commonwealth. Les bénéfices sont réinvestis dans le fonds, les membres des fonds sont les seuls alors à y bénéficier ainsi.
— Corporate funds (fonds d'entreprise) : Ces fonds sont uniquement ouverts aux employés de certains groupes ou corporations. Ils sont aussi au but non lucratif.
— Self-managed super funds SMSF (fonds autogérés) : La différence entre un SMSF et les autres types de fonds est que les membres du SMSF sont généralement aussi les fiduciaires. Ils le gèrent pour eux-mêmes et sont responsables vis-à-vis du respect des lois.
— Eligible Rollover funds (fonds de roulement admissible) : Ce sont des comptes de dépôt pour les membres perdus ou inactifs ayant un solde de compte faible. Ces

fonds ne reçoivent pas de contributions patronales.

1.1.3 Cadre réglementaire

Les fonds de superannuation sont principalement régis par le « Superannuation Industry (Supervision) Act » de 1993 et le « Financial Services Reform Act » de 2002. Les cotisations patronales obligatoires sont quand-à-eux régies par le « Superannuation Guarantee (Administration) Act » de 1992.

1.1.3.1 Superannuation Industry (Supervision) Act 1993

Cette loi fixe toutes les règles auxquelles doivent se soumettre les fonds de superannuation. Ces règles concernent tous ce qui est lié au fiduciaire, aux investissements, à la gestion, aux comptes, à l'administrations ainsi qu'aux enquêtes et aux plaintes. Elle peut infliger des sanctions en cas d'infraction aux règles.

1.1.3.2 Financial Services Reform Act 2002

Cette loi vise à standardiser le secteur des services financiers. Les fiduciaires doivent à l'issue de cette réforme disposer d'une licence pour exercer ; chaque membre qui y travaille à besoin d'avoir un permis. Pour les fonds de superannuation, cette loi veille à : fournir des licences aux gestionnaires, superviser la formation des agents des fonds de pension, fixer les exigences sur les informations à partager avec les membres existants ou prospectifs et fixer les règles de bonne et mauvaise conduite pour les fonds de superannuation.

1.1.3.3 Superannuation Guarantee (Administration) Act 1992

Cette loi vise à assurer que les employés reçoivent bien une pension de retraite de la part de leurs employés. La législation fixe un montant minimum à fournir. Dans le cas où l'employeur ne respecte par cette obligation, il fait face à des sanctions sous la forme des frais supplémentaires à verser à l'administration fiscale australienne.

1.1.3.4 Autorités de régulation

Il existe quatre autorités de régulation qui régissent les fonds de pension :

— Australian Prudential Regulation Authority APRA (Autorité de régulation prudentielle australienne) : Contrôle le caractère prudentiel des fonds ainsi que leur conformité avec le Superannuation Industry (Supervision) Act.
— Australian Securities and Investments Commission ASIC (Commission australienne des valeurs mobilières et des investissements) : S'assure que les fonds honorent leurs engagements vis-à-vis aux members. Elle vise à protéger leur droit à l'information et à protéger leurs intérêts. Elle contrôle la conformité au Financial Services Reform Act 2002.
— Australian Taxation Office ATO (Bureau d'impôts australien) : S'assure que les fonds auto-gérés sont en conformité avec les lois existantes. Elle vise aussi à uniformiser les taxes déduites des superannuations de tous les australiens.
— Superannuation Complaints Tribunal SCT (Tribunal des pensions de retraite) : Vise à cadrer le règlement des plaintes entre les fonds de pensions et leurs membres. Cette autorité n'est plus active depuis le 31 octobre 2018.
— Australian Financial Complaints Authorities AFCA : Remplace Superannuation Complaints Tribunal SCT depuis novembre 2018 dans le règlement des plaintes.

1.2 La prévoyance en Australie

Plus de 70% des australiens qui disposent d'une assurance vie l'ont à travers leurs superannuations. La plupart des fonds de superannuation offrent des couvertures dites Death (Décès), Total and Permanent Disability (invalidité), et Income Protection (protection de salaire, arrêt de travail).

Les avantages d'avoir une couverture à travers la superannuation sont les suivants :

— Les primes sont généralement moins élevées puisque le fonds achète les polices en masse.
— Les primes sont automatiquement déduites du compte de superannuation.
— Moins de visites médicales : la plupart des fonds acceptent de couvrir un membre avec une couverture standard sans passer par aucune visite médicale. Ce point fera l'objet des discussions autour de l'asymétrie d'information entre les assureurs et les membres de superannuation.
— Avantages fiscaux : les cotisations patronales ainsi que les cotisations personnelles (salary sacrifice contributions) sont taxées à 15%. Ce qui est très bas en le compa-

rant à la taxe marginale normale. Ceci rend le paiement de l'assurance à travers la superannuation avantageux sur le plan fiscal.

Parmi les risques couverts par la prévoyance en Australie, on trouve :

— Total disability : ou incapacité totale. Lorsque l'assuré est incapable d'effectuer une ou plusieurs tâches liées à son travail et de pratiques les tâches régulières. Il n'est pas non plus actif sur une autre profession. L'assuré suit dans cas un spécialiste médical et est sous des soins réguliers.

— Partial disability : incapacité partielle. S'approche de la définition de la Total Disability. La différence est que l'assuré à capable de travailler mais à un niveau réduit. Il reste en suivi médical par un spécialiste agréé.

— Terminal Illness : Maladie en phase terminale. Lorsque l'assuré, avec la confirmation de deux médecins agréés, est dit atteint d'une maladie en phase terminale qui conduira très probablement à sa mort. Dans ce cas, l'assurance verse à l'assuré une somme forfaitaire. Cette couverture est généralement groupée avec l'assurance Décès.

1.2.1 Death

Cette garantie consiste à verser un montant de capital en cas de décès ou de maladie en phase terminale (Terminal Illness) de l'assuré.

1.2.2 Total and Permanent Disability

Cette garantie consiste à son tour au versement d'une somme (forfaitaire ou rente au choix) si l'assuré n'est plus capable de poursuivre son travail définitivement à cause d'une maladie ou d'un accident.

1.2.3 Income Protection

L'income protection consiste à verser un revenu de remplacement dans le cas d'une blessure ou une maladie qui conduit à l'arrêt de l'activité salariale., Ce revenu est souvent exprimé sous la forme d'un pourcentage du salaire brut de l'assuré, En générale, ce contrat couvre 75% du salaire et verse une contribution de 10% du salaire au fonds de superannuation. Les principales caractéristiques de ce contrat sont : le waiting period et le benefit

period.

1.2.3.1 Waiting Period

La période de franchise est la durée entre le premier jour de l'arrêt maladie et le premier jour de versement des prestations. Cette durée varie en fonction du produit et généralement est égale à 30, 60 ou 90 jours. Cette durée est inversement proportionnelle au montant de la prime. Cette franchise a comme but de pallier le risque de fraude ou d'antisélection.

1.2.3.2 Benefit Period

Cette durée correspond à la durée maximale pendant laquelle l'assuré peut recevoir des prestations. Cette durée peut être fixée à l'avance. Elle vaut dans ce cas 2 ans ou 5 ans. Cependant, elle peut aussi aller jusqu'à un âge limite où les prestations sont versées jusqu'à ce que l'assuré ait 60 ou 65 ans.

1.3 Changement réglementaire

1.3.1 PYS - Protect Your Superannuation

Le paquet « Protecting Your Super » a été adopté par le parlement australien pour une date d'effet du 1er Juillet 2019. Cette législation a pour but de protéger les soldes des comptes détenus par les membres de superannuation de certains frais. Ce paquet a inclus un certain nombre de changements, notamment une interdiction des frais de sortie, un cap sur les frais des comptes à faible solde, l'annulation de l'assurance pour les comptes inactifs ainsi que le transfert des comptes inactifs et de faible solde au Bureau d'impôts australien (Australian Taxation Office ATO).

Dans le cadre du PYS, on définit cette notion d'inactivité de la manière suivante. Un membre est dit inactif si son compte de superannuation n'a pas reçu de contribution dans les 16 mois précédents. Lorsqu'un compte reçoit une contribution, il est dit actif pour les 16 mois suivants.

Les contributions qui sont considérés dans cette notion d'activité sont :

— Les contributions patronales (y compris les cotisations obligatoires, volontaires et

salary sacrifice).
— Toute contribution de la part du membre (avant ou après impôts).
— Les contributions conjugales.
— Les transferts internes (des produits hors superannuations).
— Les contributions gouvernementales.
— Les contributions Low Income Superannuation Tax Offset (LISTO).
— Les roulements des autres fonds de superannuations.
— Contributions faites à un ancien compte de superannuation dans le cas de transfert vers un compte personnel à la suite d'une rupture du contrat de travail.

Il n'y a pas de montant minimum pour les contributions.

Il est probable que les personnes impactées par cette réglementation soient issues de ces cas de figures :

- Ceux qui ont pris un congé parental ou de prise en charge.
- Ceux qui ont travaillé plusieurs années à l'étranger.
- Ceux qui changent régulièrement de travail.
- Ceux qui disposent de plusieurs comptes de superannuation.

Pour les membres qui souhaitent garder leurs assurances, il suffit de faire un opt-in, c'est-à-dire exprimer au fonds de superannuation la volonté de garder les couvertures d'assurance ou d'effectuer une contribution au compte.

Le fait qu'un assuré choisisse de garder son contrat d'assurance, est susceptible de créer une situation d'antisélection.

1.3.2 PMIF- Putting Members' Interest First

La législation Putting Members Interest First ou PMIF est entrée en vigueur le 1er avril 2020. Elle a pour vocation de protéger les comptes de faible solde des primes non-nécessaires.

Depuis cette date, les fonds de superannuation ne seront plus capables de fournir une couverture par défaut aux nouveaux members qui ont moins de 25 ans ou à des membres avec un solde faible « low balance ». Un membre est dit ayant un low balance si, depuis le 1er novembre 2019, son solde n'a jamais dépassé 6000 AUD.

De la même manière les membres qui souhaitent rester couverts peuvent faire un opt-in ou effectuer une des changements suivants avant novembre 2019 :

— Remplir un formulaire de souscription pour l'assurance.
— Demander une augmentation de la couverture.
— Informer d'un changement du statut fumeur ou de l'occupation pour mettre à jour la couverture d'assurance.
— Effectuer un transfert d'assurance depuis un autre compte de superannuation.
— Effectuer un opt-in pour le PYS.

1.3.3 Chargement mathématique et d'antisélection

ALHIS, en tant que réassureur de quelques fonds de superannuation en Australie, est impliquée d'une manière directe par ces changements réglementaires. En effet, cet environnement augmente l'incertitude sur la sinistralité et met en cause la rentabilité des traités déjà mises en place.

De ce fait, des chargements tarifaires ont été appliqués lors du renouvellement du traité de réassurance du juillet 2019 pour prévoir l'effet de ce changement.

La nature des fonds ci-dessus sera expliquée dans la suite du mémoire.

	chargement	d'expérience	Chargement pure dû au	mathématique PYS	chargement l'antisélection	dû à PYS	Chargement	total
	Death et TPD	IP	Death et TPD	IP	Death et TPD	IP	Death et TPD	IP
Fonds 1	+7,1%		+2,5%		+3,1%		+13,2%	
Fonds 2			-5,9%	+10,1%	+0,0%	+0,0%	-5,9%	+10,1%
Fonds 3			+5,1%	+24,5%	+8,9%	+8,9%	+14,5%	+35,9%
Fonds 4	+33,8%	+41,3%	-7,0%	+21,0%	+22,0%	+1,0%	+51,8%	+72,6%

Table 1.1 – Taux de chargement appliqué suite au PYS sur les 4 fonds étudiés

On distingue 2 chargements liés au PYS : Le chargement mathématique et le chargement d'antisélection.

1.3.3.1 Chargement mathématique

Le chargement mathématique est lié aux variables observables par l'assureur et utilisées dans la tarification. Il traduit le changement des profils de risque connus à l'assureur. Comme sera discuté dans les parties suivantes, les tailles de tous les fonds ont été réduites. L'âge moyen, une des variables les plus significatives dans la tarification, a augmenté presque partout.

1.3.3.2 Chargement d'antisélection

Le chargement d'antisélection est lié à l'effet d'antisélection qui pourrait découler de ce changement réglementaire. Comme sera expliqué dans la partie dédiée à l'antisélection, nous sommes en présence d'une situation d'asymétrie d'information, l'assureur peut se trouver avec une population plus risquée que prévu par le chargement mathématique.

Le but de la première partie est de vérifier la justesse de ce chargement, et voir si l'impact sur les taux d'incidence traduit bien la présence d'antisélection.

Il est à noter que cette partie du chargement se calculait du cas par cas et a dépendu immensément des négociations avec les cédantes.

En Résumé :
Le marché australien de superannuation est un marché colossal. Plus de 70% des assurés disposant d'une couverture type décès l'ont à travers leurs fonds de superannuation. Ces fonds offrent généralement des garanties Death, TPD (Total and permanent disability) et IP (Income Protection).
Le changement réglementaire PYS a pour but de protéger les comptes de superannuation des australiens qui n'y versent plus de cotisations. Les members dont les comptes n'ont pas reçu de contributions pendant plus de 16 mois perdent leurs couvertures, sauf s'ils expriment leur volonté de les garder.
Ce changement a induit l'application d'un chargement tarifaire sur les primes de réassurance. Ce chargement se décompose en chargement mathématique, lié aux changements observables du profil de risque du portefeuille, et en chargement d'antisélection lié à la présence d'asymétrie d'information et du comportement probable d'antisélection chez les assurés.
L'un des buts de cette étude est de vérifier la justesse de ce chargement.

Chapitre 2

Présentation des données et statistiques descriptives

2.1 Terminologie utilisée

Afin de bien comprendre et faciliter le traitement de notre problème, nous avons utilisé des termes spécifiques pour désigner chaque catégorie à l'égard de la réglementation PYS :

- Active/Actif : Les membres qui activement reçoivent des cotisations, et ont reçu au moins une dans les derniers 16 mois.
- Impacted/impacté : Sont les membres dont les comptes n'ont pas reçu de contribution lors des derniers 16 mois et sont alors éligibles à effectuer un opt-in ou à perdre leurs couvertures d'assurance.
- Opt-in : Les membres impactés qui ont explicitement exprimé leur volonté de rester couverts conformément à la législation.
- Inactive/Inactif : les membres impactés, qui n'ont pas effectué un opt-in et ont par la suite perdu leurs couvertures d'assurance.

Alors, pour résumer nous avons :

$$Impacted = Opt - in + Inactive$$

On parle de opt-in rate comme étant le pourcentage des opt-in par rapport aux membres

impactés. C'est-à-dire :

$$Opt - in\ rate = \frac{Opt - in}{Opt - in + Inactive}$$

On parle de pourcentage d'une certaine catégorie, comme le pourcentage de cette catégorie sur l'ensemble des membres.

$$Pourcentage\ de\ x = \frac{x}{Active + Opt - in + Inactive}$$

2.2 Présentation des bases "membership" et sinistres

Cette section vise à présenter les données et les différentes nettoyages effectuées pour les fiabiliser. Nous disposons dans cette étape de données issues de 4 fonds australiens :

- Fonds 1 : Ce fonds a été créé pour gérer les superannuations des employés de travail à bureau. Il s'agit de l'un des plus grand fonds à destination des personnes de professions professionnelles, managériales, administratives et de services dans tous les secteurs. Les membres de ce fonds seront majoritairement des white-collars ou professionals.
- Fonds 2 : C'est un industry fund qui été créé à la base comme un fonds pour les travailleurs dans le secteur public ou les fonctionnaires.
- Fonds 3 : C'est un industry fund, conçu pour les travailleurs dans la vente en détail.
- Fonds 4 : Les members de ce fonds sont majoritairement des travailleurs de petites et moyennes entreprises.

Pour chaque fonds, nous disposons de deux bases ; une base membership, qui fournit, à une date d'extraction donnée, une image des members du fond. La base des sinistres fournit, quant à elle, les différents sinistres survenus entre 2 dates avec des informations sur le sinistré et le sinistre.

Pour cette partie nous avons exploité la base de membership pour calculer le taux d'opt-in par fonds et pour identifier quelques caractéristiques des opt-ins. En outre, nous avons exploité les bases de sinistres pour calculer les taux d'incidences.

2.2.1 Base membership

Les bases de données de chaque fonds à notre disposition ne sont pas forcément du même format. Cependant, elles se ressemblent à une certaine mesure. Nous avons très souvent un nombre unique associé au membre qu'on peut aussi retrouver dans les bases sinistres. En plus, elles sont toujours des données ligne par ligne.

On distingue deux types de variables, les variables caractérisant l'assuré et les variables permettant de repérer sa catégorie, à savoir s'il s'agit d'un opt-in, inactif ou actif.

2.2.1.1 Variables caractéristiques

— La date d'extraction : Correspond à la date d'extraction de la base de données. Nous avons choisi comme date le 1er juillet 2019 ou le 30 juin 2019 pour s'aligner avec la date d'effet du PYS.
— L'identifiant du membre de fonds : C'est un identifiant unique attribué à chaque membre du fonds. Il permet de lier éventuellement les bases de membership et de sinistre.
— La date de naissance : A partir de la date de naissance, on peut calculer le ANB : Age Next Birthday ou l'âge au prochain anniversaire au moment de l'extraction ou de l'incidence du sinistre.
— Le genre : Homme ou Femme.
— Le montant assuré : Il s'agit du montant assuré par risque : Décès, TPD ou IP. Nous avons créé alors un montant assuré tout risque confondu qui permet d'éliminer les membres du fonds non-assurés.
— Le montant de prime : de la même façon, nous disposons de la prime par risque assuré.
— La profession : Equivalente à la catégorie socio-professionnelle au sens français. Après traitement nous avons gardé les sous catégories suivantes :
  - Professionals : Regroupe les personnes qui exécutent des tâches de l'ordre analytique, conceptuelle et créative. Exemple : Actuaire, Programmeur ...
  - White collars : Est équivalente à la classe des cadres. Exécutant des tâches non manuelles et les fonctions managériales et administratives. Exemple : Agent immobilier, Avocat ...
  - Blue Collar/ Manual : La classe des ouvriers et des travailleurs exécutants des tâches manuelles et gagnant un salaire horaire. On subdivise cette catégorie en

3, Heavy, Medium (On les appellera Blue collars) et light manual. Comme leurs noms l'indique, light manual incluent les professions avec peu (moins de 20%) de travail manuel ou les professions de supervision de travaux manuels. Exemple de professions classés Light Manual : coiffeur, inspecteur dans l'industrie de la viande. Blue collar est constitué des professions des techniciens spécialisés et certifiés dans les travaux manuels. Exemple de professions classés Blue collars : Bijoutier, mécanicien. Heavy manual inclue les travailleurs de tâches très manuelles mais qui restent assurables (not hazardous, non-périlleux). Exemple de professions classés Heavy Manual : cheminot, peintre en bâtiment.

- Unknown : Catégorie inconnue, cette classe est attribuée lorsque l'information de la profession n'est pas disponible.

— Le solde du compte de superannuation : Correspond au solde du compte de superannuation au moment de l'extraction.

— Benefit period : Pour les individus qui ont une couverture IP, cette variable correspond à la durée maximale pendant laquelle ils peuvent recevoir les prestations income protection (arrêt de travail).

— Waiting period : De même pour les individus qui ont une couverture IP, elle correspond à la période de franchise.

2.2.1.2 Variables PYS

Comme indiqué précédemment, nous disposons des données issues de 4 fonds. Ces données contiennent des indicateurs qui nous permettent de repérer la catégorie d'un membre donné.

Sur les fonds 1 et 2, nous disposons de la date de la dernière contribution au fonds de superannuation. Ce qui nous permet de calculer le nombre de mois depuis la dernière contribution et de repérer ainsi les membres actifs et ceux qui sont impactés. Nous avons également un PYS Flag, qui est une indicatrice de l'opt-in. On peut alors facilement attribuer les catégories actifs, inactifs et opt-ins.

Sur les fonds 3 et 4 nous n'avons que le montant des contributions lors des 12 derniers mois. Alors, nous nous sommes trouvés contraints à considérer le seuil 12 mois au lieu de 16 mois pour repérer les membres impactés par la réglementation. Sur le fonds 4, on dispose des PYS flags, chose qui facilite la tâche de repérer les opt-ins. Les membres inactifs dans ce cas seront les membres impactés, qui ne sont pas des opt-ins.

Sur le fonds 3, on ne dispose pas des PYS flags malheureusement. Alors, nous avons utilisé un autre proxy ; on disposait de la base membership en novembre 2019. Alors nous avons considéré comme opt-ins, les membres impactés en juillet 2019 (pas de contributions lors des « 12 » derniers mois) qui restent assurés en Novembre 2019.

2.2.2 Base sinistre

Sur la base des sinistres nous disposons de :

- L'identifiant du membre : cette variable permet d'identifier le membre dans la base membership du fonds.
- Type de sinistre : Permet d'indiquer s'il s'agit d'un Death, TPD ou IP. Pour certains fonds chaque type de risque a sa propre base de données.
- Cause du sinistre.
- Date d'occurrence du sinistre.
- Date de déclaration du sinistre.
- Date de fin de sinistre.
- Statut du sinistre : Declined, Pending, Open, Paid.
  - Declined : Sinistre refusé (suite à une fraude ou non-respect des clauses du contrat), rien n'a été payé à l'assuré.
  - Pending : En cours de traitement, en attente d'acceptation ou de refus.
  - Open : Sinistre ouvert, les prestations sont en cours de paiement, il n'est pas clôturé encore.
  - Paid : Le sinistre a été accepté et toutes les prestations ont été payées. Le sinistre est clôturé.
- Genre.
- Date de naissance : On extrait de cette information l'âge au moment du sinistre.
- Montants assurés.

Pour les fonds 3 et 4 nous disposons des sinistres de 2015, 2016, 2017, 2018 et 2019. Alors que sur le fonds 2, nous disposons des sinistres de 2016, 2017, 2018 et 2019. Enfin sur le fonds 1, nous disposons des années, 2017, 2018 et 2019.

2.2.3 Période d'analyses

Pour les bases des assurés (membership), nous avons pris comme date d'analyse ponctuelle le 1er juillet 2019 ou le 30 juin 2019 selon la disponibilité de la base de données. Cette date a été choisie parce qu'elle correspond à la date d'entrée en vigueur de la règlementation PYS.

En ce qui concerne les sinistres, nous avons considéré deux périodes de temps distinctes. Une période ante-PYS où nous avons pris en compte les sinistres survenus en 2015, 2016, 2017, 2018 selon l'année de souscription (ne correspond pas forcément à l'année civile). Et une période Post-PYS où nous nous sommes concentrés sur les sinistres survenus entre le 1er juillet 2019 et le 31 décembre 2020 . Ceci en prenant en compte les sinistres surnevus mais non déclarés à l'aide des taux IBNR.

2.3 Statistiques descriptives

2.3.1 Profil général des fonds étudiés

Les 4 fonds étudiés sont très différents l'un de l'autre. Nous introduisons alors dans cette partie quelques points qui les caractérisent. Le fonds 3 est le plus grand fonds dans notre lot en termes du nombre de membres assurés. En 2018, 1.5 million membres de ce fonds disposaient d'une couverture décès. Le fonds 1 est le deuxième fonds sur ce même critère avec 200 000 membres couverts en Death. Ce dernier est suivi par le fonds 4, avec 160 000 membres. Le fonds 3 est le plus petit parmi les 4 fonds avec 50 000 membres assurés en death sur la même année. Nous avons considéré 2018 comme année de référence pour effectuer ces statistiques puisque c'est l'année la plus récente avant les changements réglementaires de PYS et PMIF.

Funds	Pourcentage de femmes	L'âge moyen
Fonds 3	59%	32.8
Fonds 1	56%	42
Fonds 4	44%	40.08
Fonds 2	60%	46.32

FIGURE 2.1 – Pourcentage de femmes et âge moyen par fonds

Sur le tableau ci-dessus, on peut voir que l'âge moyen est considérablement différent entre les fonds. Le fonds 3 est le fonds le plus jeune (âge moyen de 32) alors que dans le fonds 2, l'âge moyen atteint sa valeur la plus élevée de 46 ans. Pour rappel, le fonds 3 est constitué majoritairement des travailleurs en vente de détail et le fonds 2 a été historiquement un fonds à destination des fonctionnaires. L'âge moyen au sein du fonds 1 est de 42 alors qu'il est de 40 ans dans le fonds 4. Les femmes sont majoritaires dans les fonds 2 (60%), 3 (59%) et 1 (56%) alors que leur pourcentage dans le fonds 4 est de 44%.

2.3.2 Les différences à noter entre les fonds

Dans le fonds 2, la grande majorité (circa 97%) des contrats IP sont de court-terme. Ces contrats se caractérisent par un benefit period de 2 ans et une période de franchise de 60 jours. Pour le fonds 3 et pendant notre période d'analyse, 99% des contrats étaient payables jusqu'à l'âge de 60 ans, avec une période de franchise de 60 jours.

Dans le fonds 4, 65% des garanties IP sont de courte durée, 2 ans avec une période de franchise de 90 jours. Nous en trouvons également 12% de contrats long-terme payables jusqu'à 65 ans avec une période de franchise de 30 jours et 11% de contrats de benefit period de 2 ans avec 60 jours comme période de franchise.

Pour le fonds 1, nous ne disposons pas de bases de sinistres IP pour la simple raison que ce risque n'est pas réassuré par Axa France.

En Résumé :
Dans ce paragraphe nous avons défini la terminologie utilisée dans le cadre de cette étude. Les opt-ins sont notre catégorie d'intérêt ; ce sont les membres impactés par la nouvelle réglementation qui ont choisi de rester couverts.
Nous avons aussi présenté les données dont on dispose pour effectuer cette étude. Ces données incluent des informations sur les assurés, leurs sinistres ainsi que des indicateurs permettant de repérer les différentes catégories à savoir : les actifs, les opt-ins et les inactifs. Nous avons conclu ce paragraphe par la période d'analyse et quelques statistiques descriptives des fonds étudiés.

Chapitre 3

Taux d'opt-in et illustration de l'antisélection

3.1 Calcul des taux d'opt-in et comparaison des fonds étudiés

Fund	Member Count Pre-PYS	Impacted (Inactive+ Member count	Opt-in) %	Opt-in Member Count	Opt-in rate	Member count post-PYS	% Change in Member Count
Fund 1	183 652	55 627	30%	6 337	11%	134 362	-27%
Fund 2	52 503	8 632	16%	1 953	23%	45 824	-13%
Fund 3	1 174 296	318 174	27%	77 256	24%	933 378	-21%
Fund 4	136 047	58 683	43%	22 484	38%	99 848	-27%

Ce calcul a permis d'illustrer quelques différences en pourcentage des impactés et du opt-in rate entre les fonds. D'abord, nous remarquons que les fonds 2 et 3 ont des taux d'opt-in très proches l'un de l'autre (23% et 24% respectivement).

En plus, le fonds 1 dispose du taux d'opt-in le moins élevé à travers les fonds avec 11%. Il est à noter que ce fonds détient aussi le deuxième plus grand pourcentage des impactés (30%). Ceci peut indiquer la présence d'une grande population des assurés qui n'est pas engagée (mal informée ou possédant d'autres comptes de super).

Le fonds 4 est le fonds disposant simultanément du pourcentage d'impactés et du taux d'opt-in les plus élevés. Ceci peut être expliqué par une forte compagne de médiatisation et de sensibilisation du PYS.

3.1.1 Impact de l'âge et le genre sur le taux d'opt-in

3.1.1.1 L'évolution du taux d'opt-in avec l'âge

	Opt-ins		Inactives		Opt-in rate	
Age range:	Male	Female	Male	Female	Male	Female
16-20	0	1	19	22	0,00%	4,35%
21-25	22	26	990	935	2,17%	2,71%
26-30	74	100	2631	2675	2,74%	3,60%
31-35	163	241	3114	3612	4,97%	6,25%
36-40	246	464	3391	4494	6,76%	9,36%
41-45	311	532	3261	4119	8,71%	11,44%
46-50	368	598	3174	3743	10,39%	13,78%
51-55	403	606	2452	2730	14,12%	18,17%
56-60	349	590	1844	2212	15,91%	21,06%
61-65	290	392	1107	1406	20,76%	21,80%
66-70	205	356	543	816	27,41%	30,38%
Total:	2431	3906	22526	26764	9,74%	12,74%

FIGURE 3.1 – taux d'opt-in du fonds 1

	Opt-in		Inactive		Opt-in rate	
Age range:	Male	Female	Male	Female	Male	Female
16-20	72	59	65	105	52,55%	35,98%
21-25	381	456	788	637	32,59%	41,72%
26-30	670	849	2276	1737	22,74%	32,83%
31-35	1025	1262	3375	2657	23,30%	32,20%
36-40	1550	1647	4171	3074	27,09%	34,89%
41-45	1567	1561	3948	2754	28,41%	36,18%
46-50	1696	1467	3803	2265	30,84%	39,31%
51-55	1591	1297	2868	1637	35,68%	44,21%
56-60	1532	1061	1988	985	43,52%	51,86%
61-65	895	547	896	460	49,97%	54,32%
66-70	334	195	288	155	53,70%	55,71%
Total:	11313	10401	24466	16466	31,62%	38,71%

FIGURE 3.2 – taux d'opt-in du fonds 4

FIGURE 3.3 – Evolution du taux d'opt-in par genre et tranche d'âge dans les fonds 1 et 4

Les tableaux ci-dessus montrent l'évolution du taux d'opt-in par tranche d'âge et par genre. Nous avons utilisé une heatmap pour visualiser la croissance des taux. En effet, les valeurs les plus basses sont en vert, alors que les valeurs les plus élevées sont en rouge. On peut voir que les taux sont croissants en fonction des tranches d'âge, et que pour chaque tranche d'âge, les taux d'opt-in sont plus élevés pour les femmes. Nous aimerions revoir ces constats dans la partie modélisation.

3.1.1.2 Caractéristiques de la population des opt-in

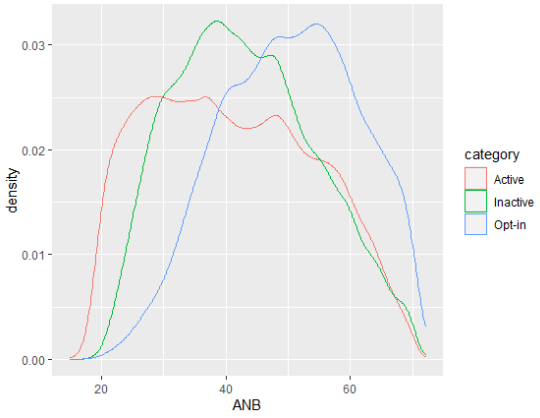

FIGURE 3.4 – La distribution de l'âge par catégorie Fonds 1

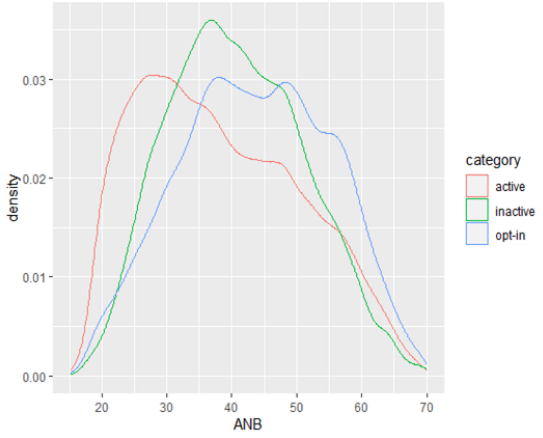

FIGURE 3.5 – La distribution de l'âge par catégorie Fonds 4

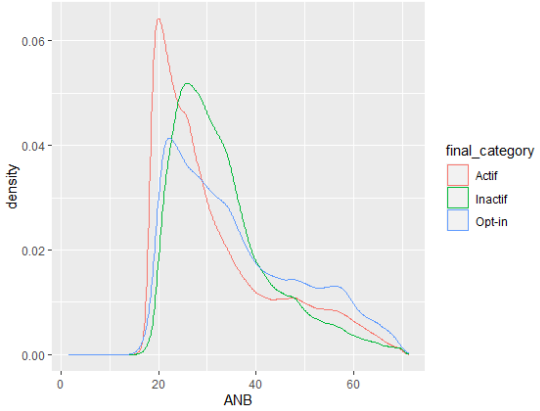

FIGURE 3.6 – La distribution de l'âge par catégorie Fonds 3

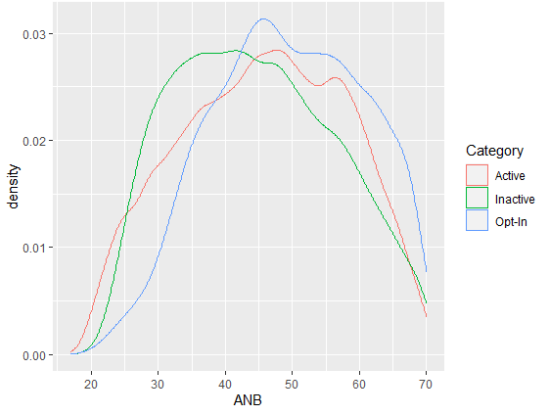

FIGURE 3.7 – La distribution de l'âge par catégorie Fonds 2

Sur les figures ci-dessus, nous avons tracé pour chaque fonds les distributions d'âge des trois catégories d'intérêt. Alors, nous pouvons voir en rouge la distribution de l'âge des membres actifs, en bleu celle des opt-ins et en vert celle des membres inactifs.

On en déduit que la population des opt-ins est plus âgée des deux autres. Ce constat est vérifié pour chacun de nos fonds. En effet, en calculant l'âge médian par catégorie, nous confirmons ce constat. Les opt-ins ont 3 ans de plus en moyenne que les autres catégories pour les fonds 2, 3 et 4 alors que pour le fonds 1, ils approximativement 8 ans en plus que les autres catégories. Donc, on peut confirmer que le changement réglementaire a incité une population plus âgée à garder sa couverture d'assurance, en dépit de ses primes plus élevés. La population dont le compte est inactif et qui a quitté le portefeuille d'assurance est dans la plupart des cas, plus jeune.

3.2 Asymétrie d'information et antisélection

3.2.1 L'asymétrie d'information

L'asymétrie d'information est un phénomène qui permet d'expliquer des comportements et des situations ayant lieu dans l'économie du marché. Ce phénomène se constate sur un marché à 2 agents, où l'un des deux dispose d'une meilleure information, c'est-à-dire qu'il connaît plus que l'autre sur les conditions de l'échange.

Ceci va à l'encontre d'une hypothèse fondamentale dans l'économie du marché, celle de la transparence d'information dans le modèle standard de la concurrence pure et parfaite. Les individus, qui sont dans ce cadre des individus rationnels veillant à optimiser leur utilité, sont alors amenés à avoir des comportements opportunistes qui nuisent à l'efficacité du marché.

3.2.2 L'antisélection en assurance

Le marché des assurances présente souvent des asymétries d'information entre les assureurs et les assurés. Ceci est lié au cycle inverse de production, c'est-à-dire que l'assureur s'engage à indemniser l'assuré avant que le sinistre ait lieu.

L'antisélection ou la sélection adverse est un phénomène statistique et économique où une offre faite sur un marché aboutit à des résultats inverses à ceux souhaités. L'antisélection se base sur l'asymétrie d'information en faveur de l'assuré, où ce dernier à une connaissance plus exacte de son propre risque. Dans une situation de parfaite symétrie d'information, les assureurs seraient capables de proposer des contrats complets (couvrant

la totalité des montants de sinistres) et actuariels reflétant le niveau exact du risque de chaque catégorie d'individus. Dans une situation d'antisélection, les primes souscrites refléteront une espérance de charge moyenne des sinistres des individus. Les individus dits à bas risque trouveront la prime proposée trop chère et inadéquate à leur profil de risque. Les assurés à haut risque quant-à-eux la trouveront faible et profiteront d'une couverture avantageuse.

Alors, les hauts risques seront demandeurs d'assurance alors que les bas risques seront moins demandeurs d'assurance pour la simple raison que leur prime est trop élevée.

3.2.3 L'antisélection liée au PYS

Dans notre situation, du moment où les individus choisissent ou non de garder leur couverture, on estime qu'en cas d'asymétrie d'information se seront les individus à haut risque qui seront le plus amenés à garder leur couverture.

Soient par exemple deux individus impactés par le PYS. Ces deux individus ont les mêmes caractéristiques observables à l'assureur. C'est à dire qu'ils ont le même âge, même genre, même catégorie socio-professionnelle ... . L'un des deux a eu un problème de santé qui accroît son risque. En absence de visite médicale, l'assureur n'a pas accés à cette information. Nous estimons dans le cadre d'antisélection que cet individu sera plus susceptible à faire un opt-in que l'autre individu considéré. Il connait son historique médical et sait qu'il encourt plus de risque de tomber malade dans le futur.

La composition du portefeuille sera ainsi changée : une part observable de ce changement inclue la réduction de la taille du portefeuille suite au départ des membres inactifs qui ne seront plus assurés ainsi que l'augmentation de l'âge moyen. Une part non observable en amont du changement concernera une sous population non connue par l'assureur. Cette population est soupçonnée d'être plus risquée et d'avoir plus tendance à effectuer un opt-in. L'assureur finira avec un profil de risque plus élevé que prévu.

Comme mentionné précédemment, un des avantages pour les assurés d'avoir une couverture offerte par leurs fonds de superannuation est l'absence de visites médicales. Ceci leur permet d'avoir des primes moins chers mais réduit la quantité d'information dont l'assureur dispose sur ses assurés et augmente les chances de présence de l'asymétrie d'information. En cas d'opt-in, les membres gardent leur couverture sans devoir passer par une visite médicale.

3.2.4 Aléa moral et fraude

On distingue généralement deux types d'aléa moral : l'aléa moral "ex ante" et l'aléa moral "ex post". Dans ce cadre, on parle d'aléa moral sans préciser lorsqu'il s'agit de l'aléa moral "ex-anté". Il y a aléa moral quand le fait d'être assurée pousse la personne à adopter des comportements susceptibles d'augmenter la probabilité de survenance d'un sinistre Décès, TPD ou IP.

L'aléa moral "ex-post" concerne l'exagération de l'ampleur des sinistres. Dans ce cas, cela revient à demander une indemnisation à laquelle on n'a pas droit. On appellera ce phénomène une fraude.

3.2.4.1 Aléa moral

L'aléa moral découle d'une situation où l'assuré est incapable d'observer parfaitement l'effort de l'assuré, qui peut conduire à l'augmentation du risque encouru. Il est impossible pour l'assureur de mettre en place une clause sur l'effort fourni par l'assuré, pour la simple raison que cet effort n'est ni observable, ni quantifiable. En général, l'aléa moral traduit la possibilité qu'un assuré augmente sa prise de risque, comparé à une situation où il supporterait entièrement les conséquences négatives d'un sinistre.

Dans notre cas, et par la nature des garanties offertes (décès, TPD et IP) et du PYS (pas de changement du niveau de couverture), il n'y a pas de vrai risque d'aléa moral.

3.2.4.2 Fraudes

Le troisième volet des phénomènes d'asymétrie d'information est la fraude. Comme indiqué précedemment, Cela est considéré comme de l'aléa moral "ex-post". La fraude en assurance est un acte malhonnête, réalisé par un assuré dans l'intention de tromper un assureur en contrevenant aux conditions générales ou particulières du contrat d'assurance.

Period Start	Period End	Decline Rate
01/12/2015	31/05/2016	13,12%
01/06/2016	30/11/2016	11,93%
01/12/2016	31/05/2017	11,63%
01/06/2017	30/11/2017	15,82%
01/12/2017	31/05/2018	14,21%
01/06/2018	30/11/2018	15,75%
01/12/2018	31/05/2019	20,11%
01/06/2019	30/11/2019	19,17%
01/12/2019	31/05/2020	18,05%
01/06/2020	30/11/2020	20,95%

FIGURE 3.8 – L'évolution du taux de refus pour le risque TPD

Period Start	Period End	Decline Rate
01/12/2015	31/05/2016	7,23%
01/06/2016	30/11/2016	6,83%
01/12/2016	31/05/2017	7,81%
01/06/2017	30/11/2017	9,09%
01/12/2017	31/05/2018	11,24%
01/06/2018	30/11/2018	12,41%
01/12/2018	31/05/2019	15,59%
01/06/2019	30/11/2019	17,89%
01/12/2019	31/05/2020	16,94%

FIGURE 3.9 – L'évolution du taux de refus pour le risque IP

FIGURE 3.10 – L'évolution du taux de refus pour les risques TPD et IP pour le fonds 3

Sur les garanties TPD et IP, les définitions des cas de couvertures sont bien précises. Après la date d'effet du PYS, nous avons remarqué une hausse des taux de refus des sinistres. Les individus, avec la campagne du PYS, ont repris conscience des couvertures et ont plus tendance à tenter leur chance avec les indemnités. Ce constat peut indiquer la hausse des cas de fraudes possibles.

3.3 Régression logistique pour déduire l'antisélection

Le principe de la régression logistique est expliqué dans le paragraphe 4.1.3.1. Dans ce cas, cette régression logistique permettra de modéliser la variable d'intérêt Y qui sera notre proxy de la sinistralité. Cette variable est égale à 1 lorsque l'individu a encouru un sinistre

lors du deuxième semestre de 2019 et 0 sinon.

Pour les variables explicatives nous allons utiliser les variables observables par l'assureur qui sont généralement utilisées dans la tarification. Ce sont ces variables qui constituent tous ce que l'assureur observe chez les assurés en cas normal. Ces variables seront l'âge, le account balance, le genre et la catégorie socio-professionnelle.

En plus de ces variables, nous allons ajouter la variable catégorie, qui vaut 1 lorsque l'individu est assuré en tant que opt-in et 0 si cet individu n'a pas été impacté par la réglementation, et est alors un membre actif. L'idée serait de prouver que cette variable impacte positivement la sinistralité, indépendamment des autres variables observables par l'assureur. Dans ce cas, les opt-ins incluront alors une sous-population plus risquée d'une manière inobservable par l'assureur, ce qui coïncidera avec la présence d'antisélection.

Cette approche n'est pas sans faille. En effet, les variables utilisées réagissent entre eux de plusieurs manières comme étudié dans la partie analyse des variables. On fera toutefois l'hypothèse que l'effet de la variable catégorie qu'on trouve avec la régression logistique effectuée est indépendant des autres variables explicatives.

3.3.1 Significativité et interprétation des coefficients

La régression logistique que nous faisons modélise linéairement ce qu'on a appelé le odd-ratio. Il s'agit de la fonction logit appliquée à la probabilité qu'un individu ait un sinistre sachant les réalisations de ses variables explicatives.

$$Logit(P(Y = 1|X)) = \beta X + \epsilon$$

Si le coefficient d'une certaine variable est significativement non nul, alors cette variable impacte la sinistralité. Nous avons effectué les tests de Wald et de rapport de vraisemblance pour prouver la non-nullité significative du coefficient de la variable catégorie.

Si le signe du coefficient est positif, alors la variable impacte positivement le logit de la probabilité étudiée, qui est une fonction croissante sur [0,1], l'intervalle auquel se situe toute probabilité.

3.3.2 Discussion des résultats

Nous avons effectué une régression logistique pour le risque TPD. En pratique, nous avons utilisé les données issues des 4 fonds. Les variables quantitatives ont été normalisées afin d'améliorer le modèle. Les variables catégorielles, dans cette régression, sont le gender qui vaut 1 si l'individu est un homme, 0 si c'est une femme et la catégorie socio-professionnelle. Nous avons gardé l'indicatrice des occupations white collar, professional, blue collar, heavy manual avec l'occupation light manual comme occupation de référence.

Nous avons effectué du sur-échantillonnage pour augmenter le nombre de sinistres, le rééchantillonnage sera expliqué dans le paragraphe 4.2.1.

Le schéma suivant montre les résultats de cette régression

Dep. Variable :	claim_tpd	**No. Observations :**	2821710
Model :	Logit	**Df Residuals :**	2821701
Method :	MLE	**Df Model :**	8
Date :	Thu, 30 Sep 2021	**Pseudo R-squ. :**	0.3470
Time :	22 :21 :46	**Log-Likelihood :**	-1.2771e+06
converged :	True	**LL-Null :**	-1.9559e+06

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-1.1296	0.003	-441.725	0.000	-1.135	-1.125
gender	-0.1750	0.003	-52.561	0.000	-0.181	-0.168
age	1.5671	0.002	843.852	0.000	1.563	1.571
account_balance	-0.1261	0.002	-80.536	0.000	-0.129	-0.123
category	0.9848	0.005	209.567	0.000	0.976	0.994
occupation_blue_collar	-1.6588	0.022	-74.498	0.000	-1.702	-1.615
occupation_professional	-0.1800	0.006	-31.739	0.000	-0.191	-0.169
occupation_white_collar	-0.1972	0.005	-42.503	0.000	-0.206	-0.188
occupation_heavy_manual	0.5385	0.008	65.758	0.000	0.522	0.555

En guise d'interprétation des résultats de cette régression, nous regardons la significativité et le signe des coefficients. Comme les variables quantitatives sont normalisées, il est plus difficile de comprendre l'ampleur des coefficients. La constante est également non interprétable dans ce cadre.

Pour les variables qualitatives, il est plus facile de comprendre. Par exemple, le passage au genre 1, c'est-à-dire le fait d'être un homme, réduit le odd ratio de 0.17. Comme la

catégorie socio-professionnelle de référence est l'occupation light manual, on voit bien que le fait d'avoir une occupation heavy manual (qui est une classe plus risquée) augmente la sinistralité de l'individu, alors que le passage en professional ou white collar réduit ce risque. La classe Blue collar est normalement plus risquée que light manual, mais ce constat n'est pas reflété sur le coefficient associé. Le coefficient de l'âge est naturellement positif, vu que le risque TPD croît effectivement avec l'âge. Le account balance quant à lui a l'effet inverse, on peut interpréter ce résultat par le fait que les individus ayant un account balance plus élevé sont à priori plus aisés ou plus anciens dans leurs métiers. De ce fait, ils n'ont pas de tâches trop risquées ou ont probablement accès à de meilleurs soins de santé.

Sur cette première régression avec la variable catégorie, on voit bien que la pvalue du test de Wald associée au coefficient de la variable catégorie est très petite. Ce qui indique que le coefficient de cette variable est significativement non nul.

On refait une autre régression sans cette variable, on remarque la décroissance du pseudo R carré. On refait un autre test de significativité du coefficient associé. C'est un test de vraisemblance.

Dep. Variable :	claim_tpd	**No. Observations :**	2821710
Model :	Logit	**Df Residuals :**	2821702
Method :	MLE	**Df Model :**	7
Date :	Thu, 30 Sep 2021	**Pseudo R-squ. :**	0.3353
Time :	22 :34 :56	**Log-Likelihood :**	-1.3000e+06
converged :	True	**LL-Null :**	-1.9559e+06

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-1.0305	0.002	-415.473	0.000	-1.035	-1.026
gender	-0.1470	0.003	-44.811	0.000	-0.153	-0.141
age	1.5543	0.002	847.345	0.000	1.551	1.558
account_balance	-0.1285	0.002	-83.405	0.000	-0.132	-0.126
occupation_blue_collar	-1.1173	0.020	-56.304	0.000	-1.156	-1.078
occupation_professional	-0.1621	0.006	-28.938	0.000	-0.173	-0.151
occupation_white_collar	-0.0084	0.004	-1.889	0.059	-0.017	0.000
occupation_heavy_manual	0.8905	0.008	114.896	0.000	0.875	0.906

La statistique de test est obtenue par l'écart entre les déviances. Sous H_0, le coefficient associé à la catégorie est nul, elle suit une loi du KHI-2 à 1 degré de liberté.

Dans notre cas ,cette statisque de test vaut donc 45874.8735 , ce qui correspond également à une p-value très faible (de l'ordre de 10^{-5}).

Nous voulions regarder si cet effet se reproduit en répartissant les assurés en 3 sous-groupes. Ceux avec un montant assuré faible, moyen et élevé. On veut vérifier si le coefficient associé à la variable category est toujours positif et s'il change en ampleur.

Groupe 1 : Montants assurés faibles

Dep. Variable :	claim_tpd	**No. Observations :**	634920
Model :	Logit	**Df Residuals :**	634911
Method :	MLE	**Df Model :**	8
Date :	Fri, 01 Oct 2021	**Pseudo R-squ. :**	0.3965
Time :	17 :01 :26	**Log-Likelihood :**	-2.6559e+05
converged :	True	**LL-Null :**	-4.4009e+05

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-1.5076	0.007	-224.062	0.000	-1.521	-1.494
gender	-0.2670	0.007	-36.702	0.000	-0.281	-0.253
age	1.3766	0.003	401.322	0.000	1.370	1.383
account_balance	-0.0544	0.003	-19.078	0.000	-0.060	-0.049
category	0.8739	0.013	67.685	0.000	0.849	0.899
occupation_blue_collar	-0.0003	0.041	-0.006	0.995	-0.080	0.080
occupation_professional	-1.6819	0.019	-89.430	0.000	-1.719	-1.645
occupation_white_collar	0.1816	0.009	20.337	0.000	0.164	0.199
occupation_heavy_manual	0.2033	0.016	13.111	0.000	0.173	0.234

Groupe 2 : Montants assurés moyens

Dep. Variable :	claim_tpd	**No. Observations :**	1466670
Model :	Logit	**Df Residuals :**	1466661
Method :	MLE	**Df Model :**	8
Date :	Fri, 01 Oct 2021	**Pseudo R-squ. :**	0.4337
Time :	17 :03 :58	**Log-Likelihood :**	-5.7574e+05
converged :	True	**LL-Null :**	-1.0166e+06

	coef	std err	z	P> |z|	[0.025	0.975]
const	-1.0761	0.004	-294.224	0.000	-1.083	-1.069
gender	-0.0779	0.005	-14.877	0.000	-0.088	-0.068
age	1.8918	0.003	627.643	0.000	1.886	1.898
account_balance	-0.1066	0.003	-31.145	0.000	-0.113	-0.100
category	1.0254	0.008	131.925	0.000	1.010	1.041
occupation_blue_collar	-18.2393	100.485	-0.182	0.856	-215.187	178.709
occupation_professional	-41.9245	6.06e+06	-6.91e-06	1.000	-1.19e+07	1.19e+07
occupation_white_collar	-0.7000	0.013	-53.195	0.000	-0.726	-0.674
occupation_heavy_manual	-18.4448	108.655	-0.170	0.865	-231.405	194.516

Groupe 3 : Montants assurés élevés

Dep. Variable :	claim_tpd	**No. Observations :**	720120
Model :	Logit	**Df Residuals :**	720111
Method :	MLE	**Df Model :**	8
Date :	Fri, 01 Oct 2021	**Pseudo R-squ. :**	0.1943
Time :	17 :04 :59	**Log-Likelihood :**	-4.0214e+05
converged :	True	**LL-Null :**	-4.9915e+05

	coef	std err	z	P> |z|	[0.025	0.975]
const	-0.9222	0.005	-181.058	0.000	-0.932	-0.912
gender	-0.0515	0.006	-9.048	0.000	-0.063	-0.040
age	1.5439	0.005	337.920	0.000	1.535	1.553
account_balance	-0.2787	0.003	-105.513	0.000	-0.284	-0.274
category	0.9689	0.007	135.017	0.000	0.955	0.983
occupation_blue_collar	-17.6993	101.081	-0.175	0.861	-215.814	180.415
occupation_professional	-0.2337	0.007	-33.082	0.000	-0.248	-0.220
occupation_white_collar	-0.4991	0.007	-70.513	0.000	-0.513	-0.485
occupation_heavy_manual	0.7486	0.011	66.736	0.000	0.727	0.771

Nous remarquons que le coefficient associé à la variable categorie est bien significativement non nul et positif, ce qui confirme notre doute que les opt-in contiennent une sous-population qui cache de l'information sur sa sinistralité.

Comme indiqué plus haut, le risque moral n'existe pas vraiment pour les garanties concernées. Un individu rationnel ne peut en aucun cas augmenter ses risques de décès

et d'invalidité d'une manière consciente. Le risque de fraude quant-à-lui n'a pas de raison d'être différent selon les catégories. Nous estimons que l'anti-sélection est bien le phénomène qui explique ce résultat.

3.4 Comparaison du taux d'incidence des sinistres

3.4.1 Hypothèses

Afin de pouvoir repérer l'impact du PYS sur la sinistralité, nous avons calculé le taux d'incidence des sinistres par tête sur les trois risques assurés, à savoir Death, TPD et IP. Nous avons utilisé les données de sinistres des 4 fonds et agrégé par année de souscription.

Le taux d'incidence que nous utilisons est le nombre des sinistres survenus dans l'année divisé par le nombre des assurés dans cette période. Nous n'avons pas pris en compte les sinistres refusés. Cependant, nous avons considéré les sinistres en cours de traitement en les diminuant légèrement avec le taux de refus historique considéré comme probabilité que le sinistre soit refusé.

Comme dans toute branche d'assurance, il existe pour les garanties étudiées des délais importants de déclaration. Nous voulions prendre en compte cet aspect. De ce fait, nous avons utilidé des taux IBNR calculé avec la méthode chain-ladder expliquée dans le paragraphe suivant. Ceci relève d'une hypothèse sous-jacente très forte, celle de supposer que les comportements passés de déclaration vont se reproduire. Ceci sera discuté dans le dernier paragraphe de cette section

IBNR

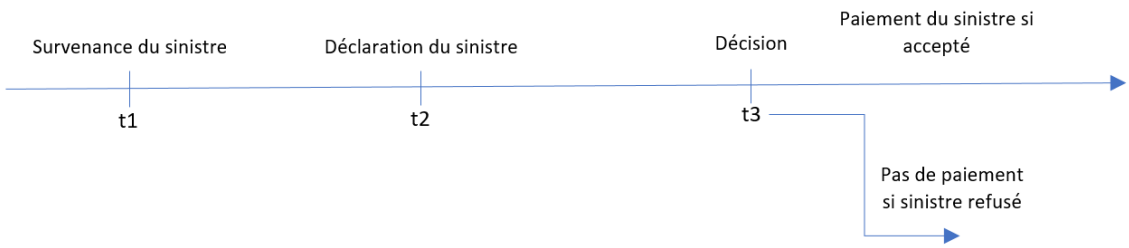

FIGURE 3.11 – Schéma simplifié de la vie d'un sinistre avant son paiement

Afin d'estimer les sinistres futurs à partir des sinistres existants, nous avons eu recours aux méthodes classiques d'évaluation des provisions. Il s'agit des méthodes basées sur les

historiques de sinistres pour estimer les provisions. Elles utilisent en majorité l'agrégation en triangle des données passées. Nous utilisons donc les notations suivantes :

— i : la date d'origine, correspond dans notre cas à la date de survenance du sinistre.
— j : la période de développement.
— $X_{i,j}$: la mesure d'intérêt pour l'année d'origine i et la période de développement j. Cette mesure présente dans notre cas le nombre de sinistres.

Ces montants sont souvent cumulés, on les note :

$$C_{i,j} = \sum_{k=0}^{j} X_{i,k} \ i.e \ X_{i,j} = C_{i,j+1} - C_{i,j}$$

Ce montant cumulé représente le nombre total des sinistres déclarés lors des j+1 premières années, pour les sinistres survenus à l'année i.

La méthode chain ladder fait partie des méthodes déterministes qui restent les plus utilisées sur le marché, grâce à leur simplicité et facilité d'implémentation. Elles sont néanmoins limitées cas elles ne permettent pas de proposer une distribution des provisions ni de calculer l'erreur de prédiction associé aux provisions futures.

Les résultats de la méthode chain ladder, comme indiqué précedemment, sont très facilement interprétables et leur implémentation reste simple et rapide. Elle suppose que les nombres de sinistres cumulés $C_{i,j}$ sont indépendants pour les différentes années d'origine i. Elle suppose également l'existence des facteurs de développement $f_0, f_1, \ldots\ldots..fJ-1 > 0$ tels que pour tous $0 \leq i \leq I$ et $0 \leq j \leq J$ nous avons :

$$\mathbf{E}(C_{i,j}/C_{i,0}, ..., C_{i,j-1}) = \mathbf{E}(C_{i,j}/C_{i,j-1}) = f_{j-1}C_{i,j-1}$$

La méthode de chain ladder consiste à estimer ses facteurs, ce qui revient à projeter les nombres de sinistres futurs. L'estimation est faite de la façon suivante, pour $j = 0, \ldots, J-1$:

$$\hat{f}_j = \frac{\sum_{i=0}^{I-j-1} C_{i,j+1}}{\sum_{i=0}^{I-j-1} C_{i,j}}$$

3.4.2 Limites des hypothèses

Une des plus fortes hypothèses que nous avons fait pour effectuer cette analyse est la possibilité d'utiliser la méthode chain-ladder et les ratios IBNR historiques. Cette hypothèse sous-entend une régularité dans les comportements de déclaration des sinistres avant et après le PYS. Cependant, on peut pertinemment soupçonner la population des opt-ins d'être plus engagée et plus consciente de sa couverture que la population totale considérée avant le PYS et que les membres actifs également. En effet, les opt-ins sont directement impactés par la réglementation et ont récemment effectué l'action de leur exprimer pour ne pas perdre leur couverture d'assurance.

Un autre facteur est la médiatisation énorme qui a accompagné le PYS, le gouvernement et les assureurs ont sensibilisé les assurés des impacts potentiels de la nouvelle réglementation et les ont rappelés des différents risques contre lesquels ils sont couverts. Ceci peut les inciter à être plus réactifs et rapides pour déclarer des sinistres après la date d'effet du PYS.

Il ne faut pas non plus négliger une part de l'augmentation au cours des dernières années des sinistres IP. Cette part d'augmentation est due à l'incidence plus courante des sinistres de maladies mentales comme la dépression, l'angoisse etc. Cette augmentation est indépendante de l'impact du PYS mais apparaît toutefois dans les taux d'incidence calculés.

3.4.3 Evolution des taux d'incidences par année et par risque

3.4.3.1 Décès

Death	2015	2016	2017	2018	2019 (Post-PYS)	PYS Impact on Frequency	GL Premium Loading For PYS	Opt-in as % of Post PYS Membership	2019 opt-ins (# of claims)	2019 actives (# of claims)
Fonds 1			0.11%	0.11%	0.13%	+23%	+6%	5%	0.50% (15)	0.11% (71)
Fonds 2		0.14%	0.15%	0.14%	0.14%	+4%	-6%	4%	0.58% (4)	0.12% (18)
Fonds 3	0.07%	0.08%	0.08%	0.08%	0.07%	-5%	+14%	8%	0.07% (30)	0.07% (325)
Fonds 4	0.09%	0.12%	0.10%	0.12%	0.14%	+20%	+13%	23%	0.23% (23)	0.11% (40)

TABLE 3.1 – Evolution des taux d'incidences du risque Décès

Les taux d'incidences avant le changement réglementaire PYS ont été stables. Les fonds 1 et 2 ont subi le plus grand écart entre les taux d'incidence des actifs et des opt-ins. Gardons en tête que le fonds 2 est de taille petite, ce qui peut donner lieu à plus de volatilité. Le fonds 3 présente une baisse légère du taux d'incidence après le PYS, on remarque toutefois que le taux d'incidence pour les opt-ins est supérieur à celui des actifs.

3.4.3.2 TPD

TPD	2015	2016	2017	2018	2019 (Post-PYS)	PYS Impact on Frequency	GL Premium Loading For PYS	Opt-in as % of Post PYS Membership	2019 opt-ins (# of claims)	2019 actives (# of claims)
Fonds 1			0.09%	0.10%	0.10%	-1%	+6%	5%	0.41% (4)	0.09% (25)
Fonds 2		0.15%	0.19%	0.22%	0.17%	-25%	-6%	4%	0.78% (3)	0.14% (12)
Fonds 3	0.08%	0.08%	0.08%	0.09%	0.10%	+6%	+14%	8%	0.14% (18)	0.10% (204)
Fonds 4	0.08%	0.09%	0.11%	0.11%	0.16%	+48%	+13%	23%	0.39% (21)	0.09% (93)

TABLE 3.2 – Evolution des taux d'incidences du risque TPD

Il est à noter que le nombre de sinistres du TPD est plus réduit que celui des autres risques. De même, nous remarquons que les taux d'incidence étaient stables avant le PYS toute en gardant une tendance vers la hausse. Pour tous les fonds, les opt-ins ont un taux d'incidence plus élevé que les actifs. C'est dans le fonds 4 qu'on détecte la hausse la plus importante, il s'agit d'ailleurs du fonds le plus impacté par le PYS, avec un pourcentage d'impacté et un taux d'opt-in élevés. Cette fois ci pour le fonds 3, on remarque une hausse du taux d'incidence total après le PYS.

Le tableau ci-dessus représente les résultats précédents pour le risque Group Life. C'est une combinaison à 50%-50% des risques Décès et TPD.

GL(50 :50 Death/TPD)	2015	2016	2017	2018	2019 (Post-PYS)	PYS Impact on Frequency	GL Premium Loading For PYS	Opt-in as % of Post PYS Membership	2019 opt-ins (# of claims)	2019 actives (# of claims)
Fonds 1			0.10%	0.10%	0.12%	12%	+6%	5%	0.45% (19)	0.10% (96)
Fonds 2		0.15%	0.17%	0.18%	0.15%	-14%	-6%	4%	0.68% (7)	0.13% (30)
Fonds 3	0.07%	0.08%	0.08%	0.08%	0.09%	+1%	+14%	8%	0.11% (48)	0.08% (529)
Fonds 4	0.09%	0.10%	0.11%	0.11%	0.15%	+34%	+13%	23%	0.31% (44)	0.10% (133)

TABLE 3.3 – Evolution des taux d'incidences du risque GL ou Group life (50% du décès et 50% du TPD)

3.4.3.3 IP

IP	2015	2016	2017	2018	2019 (Post-PYS)	PYS Impact on Frequency	GL Premium Loading For PYS	Opt-in as % of Post PYS Membership	2019 opt-ins (# of claims)	2019 actives (# of claims)
Fonds 2		0.42%	0.38%	0.44%	0.45%	+2%	+10%	4%	0.90% (4)	0.44% (48)
Fonds 3	0.23%	0.23%	0.25%	0.27%	0.28%	+4%	+36%	8%	0.21% (72)	0.29% (1367)
Fonds 4	0.20%	0.23%	0.26%	0.27%	0.36%	+33%	+22%	23%	0.88% (54)	0.21% (46)

TABLE 3.4 – Evolution des taux d'incidences du risque IP

Comme pour les autres risques assurés, les taux d'incidence IP étaient stables avant le PYS pour tous les fonds tout en présentant une légère croissance. Pour les fonds 2 et 4, les

taux d'incidence chez les opt-ins sont beaucoup plus élevés que ceux des actifs. Cependant, pour le fonds 3, on remarque l'inverse. Rappelons qu'il s'agit du fonds le plus jeune dont on dispose, avec un âge moyen de 33. Nous remarquons de nouveau que le fonds 4 est le plus sévèrement impacté en termes des taux d'incidence après le PYS.

Pour récapituler, nous avons abouti aux résultats suivants :

- Les taux d'incidence étaient stables dans les années avant PYS.
- Pour le fonds de taille plus petite, on remarque un saut des taux d'incidence dans la deuxième moitié de 2019. On note aussi que le taux chez les opt-ins est très élevé par rapport à celui des actifs. Cependant, il ne faut pas négliger la volatilité de ces résultats vu la taille petite de ces fonds.
- Pour le fonds le plus jeune, on ne remarque pas de différence notable pour le risque décès. Cependant pour les risques TPD et IP, on remarque bien une augmentation du taux d'incidence de la population des opt-ins par rapport à celui des actifs.
- C'est le fonds le plus impacté par la réglementation qui se voit augmenter ses taux d'incidence pour tous les risques après PYS. En effet, les opt-ins ayant un taux d'incidence considérablement plus élevé que les actifs, ont beaucoup plus d'impact dans ce cas sur le taux d'incidence total après le PYS.
- Dans presque tous les cas, les opt-ins ont un taux d'incidence des 3 risques plus élevé que les membres actifs. Par conséquence, cela a augmenté le taux d'incidence global après le PYS par rapport aux années précédentes.

Ces calculs, couplés avec la régression logistique effectuée pour illustrer la présence d'antisélection, montrent bien le bienfondé de la stratégie du réassureur. L'application des chargements tarifaires est justifiée.

En Résumé :
Dans ce paragraphe nous avons défini le phénomène d'antiséléction qu'on veut exhiber. Nous avons également montré la croissance des taux d'opt-in en fonction de l'âge, l'impact du genre sur ce taux ainsi que la distribution d'âge de chacune de nos catégories d'intérêt.
Nous avons effectué ensuite une régression logistique dans le but de repérer la présence d'une information cachée chez les opt-in. Nous avons voulu dissocier l'effet des caractéristiques observables par l'assureur et des caractéristiques non observables qu'on estime cachées dans le fait qu'un individu soit un opt-in sur la sinistralité. Nous avons réussi à trouver l'impact positif, indépendant de ce qui est observable par l'assureur, que la variable indicatrice d'opt-in a sur la sinistralité.
Nous avons par la suite calculé les taux d'incidence des sinistres Death, TPD et IP avant et après le PYS. Nous avons mis en évidence que les opt-ins ont un taux plus élevé que les membres actifs chose qui a déclenchée des changements sur les taux globaux. Dans ce paragraphe, nous avons discuté de la pertinence des hypothèses faites pour effectuer ce calcul.

Chapitre 4

Modélisation du comportement des assurés

4.1 Problème de classification

Dans la partie qui suit, les individus ayant effectué un opt-in sont nommés individus positifs. On leur a attribué la classe 1. De la même manière, les individus inactifs, qui sont impactés par la règlementation mais qui n'ont pas fait un opt-in, sont nommés individus négatifs. On leur a attribué la classe 0.

4.1.1 Analyse des variables

4.1.1.1 Analyse de corrélation entre les variables qualitatives

Soient X et Y deux variables qualitatives ayant respectivement r et s modalités et n réalisations. Ils décrivent les caractéristiques de n individus. Nous appelons :

— n_{ij} : le nombre des individus tels que $X = x_i$ (la i-ème modalité de X) et $Y = y_j$ (la j-ème modalité de Y).

— $n_{i.}$: le nombre des individus tels que $X = x_i$.

— $n_{j.}$: le nombre des individus tels que $Y = y_i$.

Le test d'indépendance χ^2 est basé sur la statistique d^2 qui est définie de la manière suivante :

$$d^2 = \sum_i \sum_i \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

La valeur de d^2 est comprise entre 0 (indépendance) et 1 (dépendance fonctionnelle). Cependant, cette mesure dépend du nombre d'observations et des modalités des variables explicatives en question. Alors, on ne peut pas, par le billais de cette mesure, comparer et ordonner deux à deux des variables explicatives s'ils ont des nombres de modalités différents.

Le V de Cramer est défini de la manière suivante :

$$V = [\frac{d^2}{n \inf((s-1);(r-1))})]^{\frac{1}{2}}$$

Idem, le coefficient de Cramer est compris entre 0 et 1. Il compense l'effet du nombre d'observations et des modalités.

Variable 1	Variable 2	V de Crammer
fund	occupation	0.719
fund	gender	0.080
fund	benefit period	0.632
fund	waiting period	0.597
gender	occupation	0.090
occupation	benefit period	0.468
occupation	waiting period	0.458

TABLE 4.1 – Le V de Cramer sur les variables qualitatives dans notre jeu de données

Comme attendu, les variables occupation, waiting period et benefit period sont fortement liées à la variable fond. Ceci se justifie par le fait que chaque fonds a des caractéristiques spécifiques pour les contrats IP, et aussi par la répartition historique des fonds de pension par secteur d'activité.

4.1.1.2 Analyse de corrélation entre les variables quantitatives

Pour ce faire nous allons utiliser le coefficient de corrélation de Pearson ρ , qui est défini de la manière suivante :

Etant donné deux variables aléatoire continues X et Y :

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

où :

- $cov(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$
- $\sigma_X^2 = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
- $\sigma_Y^2 = \mathbb{E}((Y - \mathbb{E}(Y))^2) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$

Pour un échantillon de taille n, et avec $(x_1, y_1), ..., (x_n, y_n)$ les réalisations des variables X et Y, le coefficient de corrélation de Pearson, désigné alors r_{xy} est défini par :

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

Le coefficient de corrélation de Pearson est compris entre -1 et 1. Si sa valeur absolue vaut 1, cela implique une colinéarité entre les deux variables en question. Le signe du coefficient indique comment les deux variables varient l'une par rapport à l'autre, un signe positif indique que lorsque X croît Y croît également, alors qu'un signe négatif indique que lorsque X augmente, Y diminue. Une valeur de 0 implique qu'il n'y a pas de dépendance linéaire entre les variables étudiées.

	age	account balance	SI DTH	SI TPD	SI IP
age	1.000000	0.448412	0.117424	0.050353	-0.002969
account balance	0.448412	1.000000	0.145253	0.216284	0.085707
SI DTH	0.117424	0.145253	1.000000	0.528564	0.299200
SI TPD	0.050353	0.216284	0.528564	1.000000	0.116703
SI IP	-0.002969	0.085707	0.299200	0.116703	1.000000

TABLE 4.2 – Le coefficient de corrélation de Pearson des variables quantitatives sur notre jeu de données

Les variables age et account balance ont un coefficient de corrélation important. Les membres plus âgés ont passé généralement plus de temps sur le marché de travail et ont par la suite fait plus d'économie. Les variables sum insured Death et TPD sont aussi considérablement corrélées.

4.1.1.3 Analyse de corrélation entre les variables qualitatives et quantitatives

Pour mesurer l'effet d'une variable qualitative sur une variable quantitative, nous avons choisi la mesure η^2. Soit X une variable continue et Y une variable quantitative à K modalités. La variabilité totale de X peut se décomposer en variabilités expliquées par l'appartenance aux groupes des K modalités de Y, et en variabilités résiduelles (intra-groupes). η constitue la racine de la proportion de la variable expliquée.

$$\eta^2 = \frac{\sum_{i=1}^{K} n_k(\bar{x_k} - \bar{x})^2}{\sum_{i=1}^{K}(x_i - \bar{x})^2}$$

où :

- $\bar{x}$ est la moyenne empirique des $\{x_1, ..., x_n\}$
- $\bar{x_k}$ est la moyenne empirique des $\{x_1, ..., x_n\}$ qui appartiennent au groupe de la k-ème modalité de Y.
- n_k le nombre des x_i qui appartiennent au groupe de la k-ème modalité de Y.

η^2 est compris entre 0 et 1. Pour déterminer si le coefficient η^2 est élevé au faible, il n'existe pas de manière formelle, cependant une règle d'index est proposée par Cohen(1998) :

— $0 \leq \eta^2 \leq 0.01$ négligeable
— $0.01 < \eta^2 \leq 0.06$ faible
— $0.06 < \eta^2 \leq 0.14$ moyen
— $0.14 < \eta^2 \leq 1$ élevé

La variable catégorielle	La variable continue	η^2
fund	age	0.097
occupation	age	0.091
fund	account balance	0.055
occupation	account balance	0.067
fund	death sum insured	0.027
fund	tpd sum insured	0.179
fund	ip sum insured	0.251
gender	age	0.003
gender	account balance	0.002

TABLE 4.3 – Le coefficient η^2 de quelques variables de notre jeu de données

Comme indiqué précédemment, une valeur de plus de 0.14 est considérée comme élevée. Ceci correspond avec l'impact de la variable fonds sur les montants assurés IP et TPD. Il n'y a pas d'impact notable de la variable gender sur l'âge et l'account balance

4.1.2 Pré-traitement des données

Pour assurer de bonnes performances des méthodes de classification, nous avons effectué des démarches de traitement des données :

— Pour les variables continues : Nous les avons normalisées à l'aide du Standard Scaler. Il s'agit de déduire la moyenne et diviser par l'écart type. Cette méthode permet d'assimiler chaque variable continue à une variable aléatoire normale centrée réduite.

— Pour les variables discrètes ou catégorielles : Nous les avons transformées en plusieurs variables binaires. Par exemple pour la variable gender, on avait deux instances Male et Female. On a attribué 1 pour les hommes et 0 pour les femmes. Ceci garantie un format adapté pour entraîner facilement les différents modèles.

4.1.3 Etat de l'art des méthodes de classification

4.1.3.1 Régression logistique

La régression logistique sert à modéliser une variable qualitative à 2 modalités : 1 ou 0, présence où absence d'une maladie, bon ou mauvais client, fraude ou non fraude... Dans notre cas, elle vise à modéliser 1 : opt-in, 0 : inactif parmi les impactés par le PYS.

Les modèles de régression linéaire visent à l'explication d'une variable d'intérêt quantitative à l'aide d'une variable explicative. En effet, le régresseur linéaire usuel $X\beta$ ne prend pas forcément des valeurs binaires. L'objectif alors est adapté à la situation, en essayant de modéliser plutôt les probabilités :

$$\pi = \mathbb{P}(Z = 1) \ ou \ 1 - \pi = \mathbb{P}(Z = 0)$$

Ou une transformation de cette probabilité. En pratique, on fait intervenir une fonction réelle monotone g agissant de [0,1] dans $\mathbb{R}$. L'objectif est donc de chercher un modèle linéaire sous la forme :

$$g(\pi_i) = x_i'\beta$$

Pour des raisons théoriques et pratiques, le choix de cette fonction est le logit, défini comme :

$$g(\pi) = logit(\pi) = \ln \frac{\pi}{1-\pi} \; avec \; g^{-1}(x) = \frac{e^x}{1+e^x}$$

Le rapport $\frac{\pi}{1-\pi}$ s'appelle l'odd, il exprime une cote. La régression logistique s'interprète donc comme la modélisation linéaire du *log* odd tandis que certains modèles expriment des rapports de cotes (odds ratio) c'est à dire l'influence d'un facteur qualitatif sur le risque d'un échec ou la chance d'un succès de notre variable d'intérêt Z.

4.1.3.2 Arbre de décision

Dans un problème de classification, l'arbre de décision repose sur le principe de segmenter un échantillon de données. Cette segmentation est faite à l'aide des règles binaires successives, l'idée est d'associer à chaque sous-groupe une instance de notre variable d'intérêt.

Cette segmentation est présentée sous la forme d'un arbre orienté dont :

— Les noeuds internes sont étiquetés par un test applicable à tout individu sur une des variables explicatives, ils correspondent aux séparations que l'arbre effectue.

— Les arcs contiennent les résultats du test, souvent l'arc gauche amène à la suite de la décision en cas de vérification de la condition du nœud, l'arc droit au cas contraire.

— Les feuilles correspondent aux segmentations finales, elles sont étiquetées par l'une des deux modalités de la variable d'intérêt.

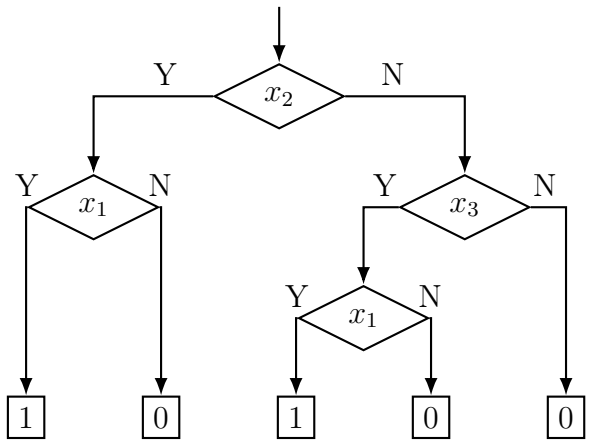

FIGURE 4.1 – Exemple d'un arbre de décision : les losanges correspondent aux noeuds et les carrés aux feuilles.

Un arbre de décision est donc un classifieur organisé d'une manière arborescente. Cette construction d'arbre lui attribue l'un de ses points les plus forts, celui d'être facilement interprétable. Il est aussi rapide à entraîner et capable de traiter à la fois des variables quantitatives et qualitatives. Il est construit suivant les étapes ci-dessus :

- Trouver la variable explicative et la condition qui permettent d'obtenir la meilleure répartition dans le cadre d'initiation du premier nœud.
- Décider si le nœud courant est terminal auquel cas une classe sera affectée à l'issue de chaque arc. Sinon, retrouver une variable explicative et une condition qui permettent de partitionner d'une manière optimale les individus.

Ceci en se basant sur des critères bien définis à savoir : le critère d'arrêt, de partitionnement et d'affectation de classe.

Le critère d'arrêt permet de contrôler la croissance de l'arbre. L'arbre ne doit pas être trop simpliste, par peur de ne pas bien apprendre les caractéristiques des deux classes ni trop compliqué, par peur de sur-apprentissage.

Le critère de partitionnement permet de choisir la variable explicative qui sera utilisé dans le nœud courant pour effectuer la segmentation. Les mesures les plus souvent utilisées pour juger de la qualité de partitionnement sont les suivantes : Notons p_0 et p_1 la proportion d'individus respectivement de la classe 0 et 1. L'indice de Gini est défini comme :

$$G(P) = p_0(1 - p_0) + p_1(1 - p_1) = 1 - (p_0^2 + p_1^2)$$

L'entropie de Shannon est définie comme :

$$E(P) = -p_0 \ln(p_0) - p_1 \ln(p_1)$$

4.1.3.3 Extreme Gradient Boosting ou Xg-boost

Boosting

Le boosting est une technique ensembliste qui consiste à agréger des classifieurs élaborés séquentiellement sur un échantillon d'apprentissage dont les poids des individus sont corrigés au fur et à mesure. Les classifieurs, où le i-ème classifieur est noté h_i , sont assemblés

avec une somme pondérée de la façon suivante :

$$\sum \alpha_i.h_i$$

α_i correspond au coefficient de la pondération. Il est une fonction de ϵ_i qui est l'erreur du i-ème classifieur :

$$\alpha_i = \frac{1}{2}\ln(\frac{1-\epsilon_i}{\epsilon_i})$$

α_i est donc une fonction décroissante de l'erreur du classifieur. En d'autres termes, plus l'erreur d'un classifieur est élevé, moins son importance dans la classification du modèle. On donne plus d'importance aux classifieurs plus performants.

Adaboost

Adaboost, abrévié du mot Adaptive boosting, est l'un des algorithmes de boosting les plus célèbres. Il a été utilisé pour la première fois par Yoav Freund et Robert Shapire et utilise les arbres de décision. Le « weak learner » ou classifieur faible que cet algorithme utilise est généralement un arbre de décision à seulement 2 braches et 2 feuilles. Les étapes de construction de ce weak learner sont les suivantes :

— On attribue le même poids à chaque instance du dataset d'apprentissage.
— On entraîne le weak learner de façon à minimiser l'erreur.
— Comme expliqué dans le paragraphe précédent, on attribue un poids à ce learner qui est inversement proportionnel à son erreur, ce poids sera utilisé dans le vote final.

De plus, cet algorithme pousse les prochains prédicteurs à corriger l'erreur du prédicteur précédent. Ceci est fait en augmentant le poids des instances où le prédicteur précédent s'est trompé et diminuer le poids des instances où le prédicteur a correctement prédit.

Formellement, notons les observations (ou les individus) par $(x_1, y_1), ..., (x_m, y_m)$ où x_j décrit les caractéristiques de l'individu j, et y_j est la variable à prédire (dans notre cas, y_j traduit si l'individu a fait un opt-in ou pas).

Le poids est initialisé comme un poids uniforme sur tous les individus, c'est à dire :

$$D_i(j) = \frac{1}{m}, j \in 1, .., m \ et \ i \in 1, ..., I$$

Alors, pour $i = 1, ..., I$, c'est à dire qu'on parcourt l'ensemble des prédicteurs :

1. Nous trouvons une fonction h_i qui minimise l'erreur de classification ϵ_i en prenant en compte la pondération D_i. Cette fonction vérifie le programme de minimisation

suivant :

$$h_i = argmin_{h \in \mathbf{H}} \sum_{j=1}^{m} D_i(j) E(y_j, h(x_j))$$

$$\epsilon_i = \sum_{j=1}^{m} D_i(j) E(y_j, h_i(x_j))$$

2. On sélectionne cette fonction si $\epsilon_i < 0.5$.

3. On calcule le poids du modèle dans le vote final :

$$\alpha_i = \frac{1}{2} ln(\frac{1 - \epsilon_i}{\epsilon_i})$$

4. On met à jour les poids des individus de la manière suivante :

$$D_{i+1}(j) = \frac{D_i(j) \exp^{-\alpha_i E(y_j, h_i(x_j)}}{Z_i}$$

Avec Z_i un facteur de normalisation pour que la somme des poids soit égale à 1.

Gradient Boosting

Cet algorithme, inventé par le statisticien américain Jerome Friedman, partage beaucoup de traits avec l'algorithme Adaboost. Les deux sont des ensembles de weak learners construits l'un après l'autre. En outre, les deux font que chaque weak learner est entraîné pour corriger les erreurs de son précédent. Cependant, le Gradient Boosting n'attribue pas de poids différents aux prédicteurs dans le vote final, ils ont tous la même importance quelle que soit leur performance.

En gardant les mêmes notations que précédemment, la fonction du coût s'écrit :

$$J(y, h) = \sum_{j=1}^{m} E(y_j, h(x_j))$$

La descente de gradient est une technique itérative permettant d'approcher la solution d'un problème d'optimisation. Ceci est fait en approchant la solution suivant la pente (ou le gradient) de la fonction objective.

En utilisant le principe de cette méthode, le prédicteur de l'étape i, serait construit à

partir de celui de l'étape précédente en ajoutant une amélioration. Ceci s'écrit comme :

$$h_i(x_j) = h_{i-1}(x_j) - \eta \times \nabla E(y_j, h_{i-1}(x_j))$$

avec :

— η la constante d'apprentissage, elle permet de piloter le processus jusqu'à convergence. Elle est un hyper paramètre du modèle.

— ∇ le gradient, dans ce cas la dérivée partielle de la fonction de coût par rapport au modèle.

Xg-boost

Xg-boost ou Extreme Gradient Boosting est une amélioration de l'algorithme Gradient Boosting. Sa particularité réside dans son choix de weak learners, il utilise les arbres de décision. Les arbres qui ne sont pas assez bons sont « élagués », c'est-à-dire qu'on leur coupe des branches et ils sont complétement supprimés s'ils ne peuvent pas être améliorés. Par le biais du cette façon, le Xg-boost ne garde que les bon weak learners.

De surcroît, le Xg boost s'arme avec une optimisation informatique de très haut niveau, ce qui rend les différents calculs nécessaires à l'apprentissage du modèle extrêmement rapides.

Il propose également un panel d'hyperparamètres très conséquent. Il est possible via ces paramètres de choisir la combinaison adéquate pour maximiser le score. Toutes ces caractéristiques ont motivé notre choix d'utiliser le Xg-boost.

4.2 Imbalanced learning

4.2.1 Rééchantillonnage

Avoir une base déséquilibrée, c'est-à-dire une base où la répartition des deux classes présente une disparité, peut nuire aux apprentissages et performances des modèles. Les modèles donneront plus d'importance à la classe majoritaire. Cette classe peut ne pas nous intéresser. Ce problème est très présent dans la littérature, par exemple pour classifier les cas de fraude et de non-fraudes, la plupart des cas sont heureusement des non-fraudes. Cependant, arriver à caractériser et modéliser les cas frauduleux est plus intéressant sur le plan pratique. Il est donc judicieux de mettre en avant des bases équilibrées où les 2 classes

à identifier sont à proportion égale.

Pour ce faire, il existe plusieurs méthodes possibles. Elles comprennent le sur-échantillonnage et le sous-échantillonnage. Le sur-échantillonnage consiste à augmenter la taille de la classe minoritaire en produisant des individus issus de cette même classe jusqu'à l'obtention d'une base équilibrée. Un avantage majeur de cette méthode est qu'on ne perd pas d'information. Cependant, lorsqu'on dispose d'une base de données volumineuse, on ajoute davantage de volume ce qui peut devenir très coûteux en termes de stockage et de temps d'entrainement des modèles. Un autre désavantage est qu'on ajoute de la dépendance des modèles au jeu de données initial, chose qui peut conduire au surapprentissage.

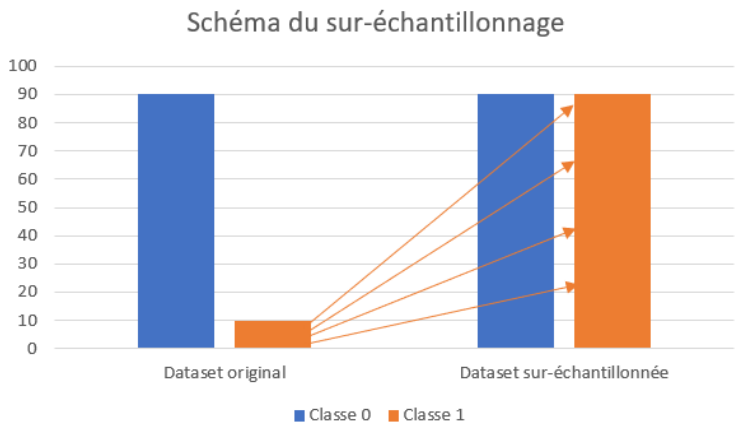

FIGURE 4.2 – Schéma simplifié du sur-échantillonnage

Le sous-échantillonnage quant à lui, consiste à réduire le nombre des instances de la classe majoritaire afin d'avoir la même proportion des classes. Cette méthode est plus consistante avec notre cas d'usage, elle permet d'économiser le temps d'apprentissage et le coût de stockage. L'inconvénient majeur de cette méthode est qu'on perd de l'information en se débarrassant d'un volume conséquent de données.

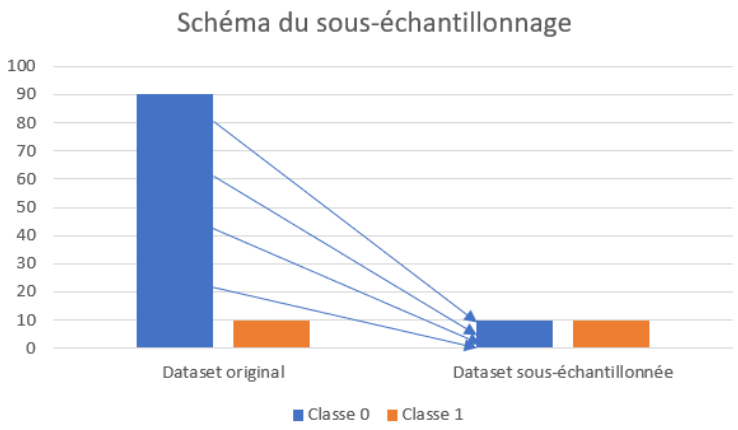

FIGURE 4.3 – Schéma simplifié du sous-échantillonnage

Ces deux méthodes peuvent être combinées, en sur-échantillonnant la classe minoritaire et en sous-échantillonnant la classe majoritaire. Ceci sera un compromis et diminuera l'impact des inconvénients des deux méthodes.

4.2.1.1 Echantillonnage aléatoire/ Random sampling

Cette méthode est la plus simple. Elle consiste à supprimer aléatoirement des échantillons de la classe majoritaire. C'est l'une des premières méthodes construites pour pallier le problème du déséquilibre des données. Elle peut cependant augmenter la variance du classifieur.

4.2.1.2 SMOTE

Cette méthode de sur-échantillonnage se concentre sur la classe minoritaire, qui est augmentée en créant des exemples « synthétiques ». C'est l'un des algorithmes les plus utilisés pour le sur-échantillonnage.

La méthode est basée sur les étapes suivantes :

— Un individu de la classe minoritaire est choisi au hasard (on l'appellera parent 1).
— On désigne les k plus proche voisins de cet individu parmi lesquels on choisit aléatoirement un autre individu que l'on appelle parent 2.
— On crée un nouvel individu synthétique (descendant) à un point choisi au hasard sur la ligne entre les deux parents dans l'espace des caractéristiques. Le nouvel

individu dépend linéairement de ces parents, mais celui-ci reste quand même un nouvel individu généré.

4.2.1.3 Liens de TOMEK ou TOMEK LINKS

C'est une méthode de sous-échantillonnage très connue pour ses bienfaits sur les performances des classifieurs. Elle s'appelle liens de Tomek. Ces liens sont des paires de données qui sont les plus proches autour de la ligne de séparation. Plus formellement un lien de Tomek est défini de la manière suivante :

Soit un pair d'individus (x_i, x_j) tel que $x_i \in S_{min}$ et $x_j \in S_{max}$, avec S_{min} et S_{max} respectivement la classe minoritaire et majoritaire, et $d(x_i, x_j)$ la distance entre x_i et x_j. Ce pair est appelé un lien de Tomek s'il n'y a aucun individu x_k qui vérifie :

$$d(x_i, x_k) < d(x_i, x_j) \; ou \; d(x_j, x_k) < d(x_i, x_j)$$

Si deux individus forment un lien de tomek, alors il n'y a que deux possibilités : l'un des deux est un bruit, il se trouve dans une zone caractérisant une autre classe, ou les deux se trouvent près d'une frontière et sont susceptibles d'être problématiques lors de la classification. En supprimant ces paires de données, la séparation entre les deux classes sera élargie, et les classifieurs deviendront plus performants.

4.2.1.4 Smote-TL

La méthode hybride combine les approches de sur-échantillonnage et celles de sous-échantillonnage en éliminant des données dans la classe majoritaire et ajoutant des données dans la classe minoritaire afin de rééquilibrer la distribution des classes. L'approche Smote-TL est la combinaison des algorithmes Smote et Tomek Links.

4.2.2 Méthodes d'évaluation

Afin de mesurer la performance de nos modèles, nous avons réparti notre base de données en une base d'apprentissage (train set) et une base de test (test set). Les modèles seront entraînés sur le train set et sont ensuite examinés sur le test set. Nous avons alors besoin d'une métrique qui permettra de comparer nos modèles entre eux et de juger leur

performance d'une manière mesurable.

Notre cadre est le suivant :

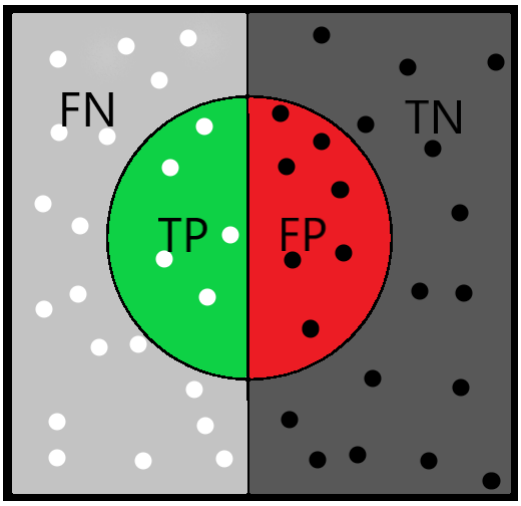

FIGURE 4.4 – Cadre de l'évaluation du problème de classification binaire

— Les points en blanc représentent les éléments positifs, les points en noir représentent quant-à-eux les éléments négatifs
— les points qui se trouvent à l'intérieur du cercle sont les éléments que le modèle prédit comme positifs
— Les éléments qui sont situés dans la partie verte du cercle sont correctement prédits positives, ils sont donc des vrais positifs
— Les éléments qui se trouvent dans la partie gris foncé du carré sont prédits négatives correctement, ils sont donc des vrais négatifs
— Les éléments qui sont inclus dans la partie rouge du cercle, c'est à dire ceux que le modèles prédit à tort comme positifs, sont libellés faux positifs
— d'une manière symétrique, les éléments qui se trouvent dans la partie gris clair du carré, sont des éléments positifs que le modèle n'a pas réussi à identifier comme tels, ils sont alors les faux négatifs

4.2.2.1 Matrice de confusion

La matrice de confusion regroupe les différentes classes décrites ci-dessus, de la façon suivante :

		Classe Prédite		
		n	**p**	**total**
classe réelle	**n′**	True Negative	False Positive	N′
	p′	False Negative	True Positive	P′
	total	N	P	

FIGURE 4.5 – Matrice de confusion

4.2.2.2 F1-score

Cette répartition a permis la définition de plusieurs mesures. L'accuracy est définie comme suit :

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

En présence de classes déséquilibrées, cette mesure peut ne pas refléter l'intelligence réelle du modèle, par exemple si la classe minoritaire ne constitue que 10% du total des individus alors un modèle très naïf qui prédit toujours la classe négative aura une accuracy de 90% alors qu'il n'a aucun pouvoir prédictif.

On définit alors la précision (precision) et le rappel appelé aussi sensitivité (recall) :

$$Precision = \frac{TP}{TP + FP}$$

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

La précision mesure le pourcentage des vrais positifs sur le nombre de positifs prédits. Le rappel quant-à-lui mesure le pourcentage des vrais positifs sur le nombre des positifs réel présent sur les données. Ces mesures sont plus intéressantes que l'accuracy, car ils ne

prévilégient pas forcément la classe majoritaire.

Le f1-score correspond à la moyenne harmonique de la précision et le rappel, ce qui s'écrit comme :

$$f1 - score = \frac{Precision \times Rappel}{Precision + Rappel}$$

Lorsque le f1-score est élevé, cela signifie que la précision et le rappel sont tout ensemble élevés ; lorsque le f1-score est faible, à ce moment on regarde la matrice de confusion pour visualiser de plus près comment le modèle se comporte.

On tient à indiquer que la mesure f1-score n'est pas symétrique. Elle accorde plus d'importance à la prédiction correcte des éléments positifs. si on permute les deux classes, le f1-score changera de valeur.

Nous aimerions aussi regarder le taux des faux positifs (False Positive Rate), qui permet de regarder parmi les éléments réellement négatifs, le pourcentage des éléments prédits positifs à tort :

$$Taux\ des\ faux\ positifs = \frac{FP}{FP + TN}$$

4.2.2.3 Le choix et le calcul du f1 score

La précision et le rappel constituent les deux degrés de liberté de la matrice de confusion. Cette matrice contient les TP, TN, FP et FN. Les relations qui diminuent les degrés de libertés sont $n = TP + TN + FP + FN$ et le fait que les vraies proportions des classes sont déjà connus.

L'utilisation du f1 score est ainsi justifiée. C'est une combinaison de ces deux degrés de libertés qui restent, à savoir la précision et le rappel. Cette mesure est devenue très populaire dans le contexte des classifications, et plus particulièrement dans les problèmes de classification déséquilibrée. ce qui correspond parfaitement à notre cas.

Pour la régression logistique, qui fournit en output des probabilités ,on utilise le seuil 0.5 pour déterminer à partir des sorties si un individu est un opt-in ou pas. On obtient un vecteur avec les catégories prédites. Le f1 score est calculé ensuite à partir de ce vecteur et du vecteur des vraies valeurs dont on dispose.

Pour l'arbre de décision et le Xg-boost, la catégorie est fournie à la sortie du modèle, elle correspond à la classe majoritaire de la dernière feuille de l'arbre. On refait le même

calcul pour obtenir le f1 score.

4.2.2.4 ROC et AUC

Dans une autre vision d'évaluation du modèle, on peut ne pas choisir de privilégier les éléments positifs, c'est à dire les opt-ins. Et plutôt choisir d'avoir une répartition équilibrée. En effet, les individus inactifs sont beaucoup plus présents en réalité. On peut envisager un modèle qui traite les deux classes de la même manière, sans avoir recours au rééchantillonnage ni au f1-score. Cette vision sera intéressante pour la généralisation du modèle pour calculer un taux d'opt-in PMIF.

Nous considérons alors le AUC, défini ci-dessus.

ROC

La courbe ROC est un outil graphique permettant d'évaluer et de comparer les modèles. En particulier, les modèles de classification. Cette courbe permet de mesurer la qualité du modèle sans dépendre des proportions de classe.

La sortie du modèle de classification est en effet une probabilité d'attribution à la classe positive et pas l'attribution directe à une classe. Elle correspond à cette probabilité $P(Y = 1/X)$, avec X l'ensemble des variables explicatives.

La règle d'attribution la plus utilisée est la suivante : L'individu est attribué à la classe où la probabilité conditionnelle d'appartenance est la plus grande. C'est-à-dire, qu'on attribue à un individu la classe 1 si $P(Y = 1/X) > 0.5$.

L'idée derrière la courbe ROC est de faire varier ce seuil s de 0 à 1. Nous obtiendrons à chaque fois des nouvelles matrices de confusion et donc de nouvelles sensibilité et spécificités. Ce sont ces valeurs qui seront utilisée pour créer la courbe ROC où on met en relation la sensibilité en ordonnée (ou taux de vrais positifs TVP) et 1- spécificité en abscisse qui correspond au taux de faux positifs.

AUC

L'aire sous la courbe ROC (ou AUC Area Under The Curve) est utilisé afin d'évaluer la qualité d'un classifieur. Elle correspond à l'intégrale de la fonction ROC. Théoriquement $AUC = P(Y > X)$.

Plus l'aire est élevée plus la probabilité de placer un positif devant un négatif en classant les individus par score est élevée. Dans le meilleur des cas, tous les positifs auront un score

supérieur aux négatifs, c'est-à-dire que l'AUC vaudra 1. De plus si le score classe au hasard les individus, l'aire sous la courbe ROC est égal à 0.5.

4.3 Modèle retenu et résultats

À la suite de l'analyse faite dans la partie précédente, nous avons divisé notre jeu de données en deux groupes : Un groupe pour l'apprentissage, qu'on appelle couramment train set et un groupe pour le test, qu'on appelle test set. En conséquence, nous allons entrainer les modèles suivants sur le train set et les comparer à l'aide de leur f1-score et leur AUC calculés sur le test set. Ces modèles sont :

— Une régression logistique avec une RFE (Recursive Feature Elimination) ou une pénalisation type Lasso ou Ridge.
— Un arbre de classification sous R.
— Un arbre de décision sous Python avec un grid search pour choisir les meilleurs hyperparamètres.
— Un XG-boost aussi avec un grid search pour un choix optimal d'hyperparamètres.

4.3.1 Illustration des résultats

4.3.1.1 Premiers résultats

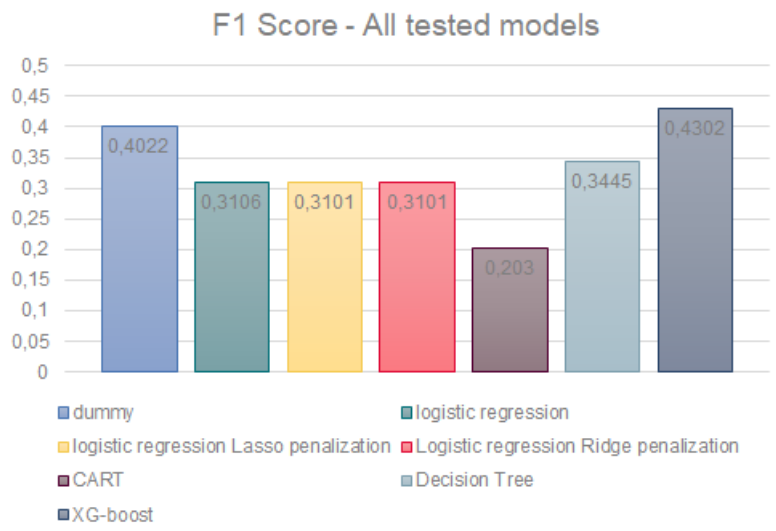

FIGURE 4.6 – F1 score par modèle sur un train set non modifié

Comme indiqué précédemment, la métrique principale utilisée pour évaluer les modèles est le f1-score. Le modèle dummy est un modèle naïf qui prédit toujours qu'un individu fera un opt-in. Ce modèle est utilisé ici comme benchmark. Si un autre modèle a un score inférieur à celui du modèle dummy on estime qu'il n'a pas de pouvoir prédictif.

On remarque ici que presque tous les modèles scorent pire que le modèle dummy, sauf le XG-boost. Une explication potentielle de ce problème serait le déséquilibre des classes. Nous avons dans notre jeu de données 75% des inactives et que 25% des opt-ins.

Nous introduisons par la suite le sous-échantillonnage de notre train set.

4.3.1.2 Impact du sous-échantillonnage

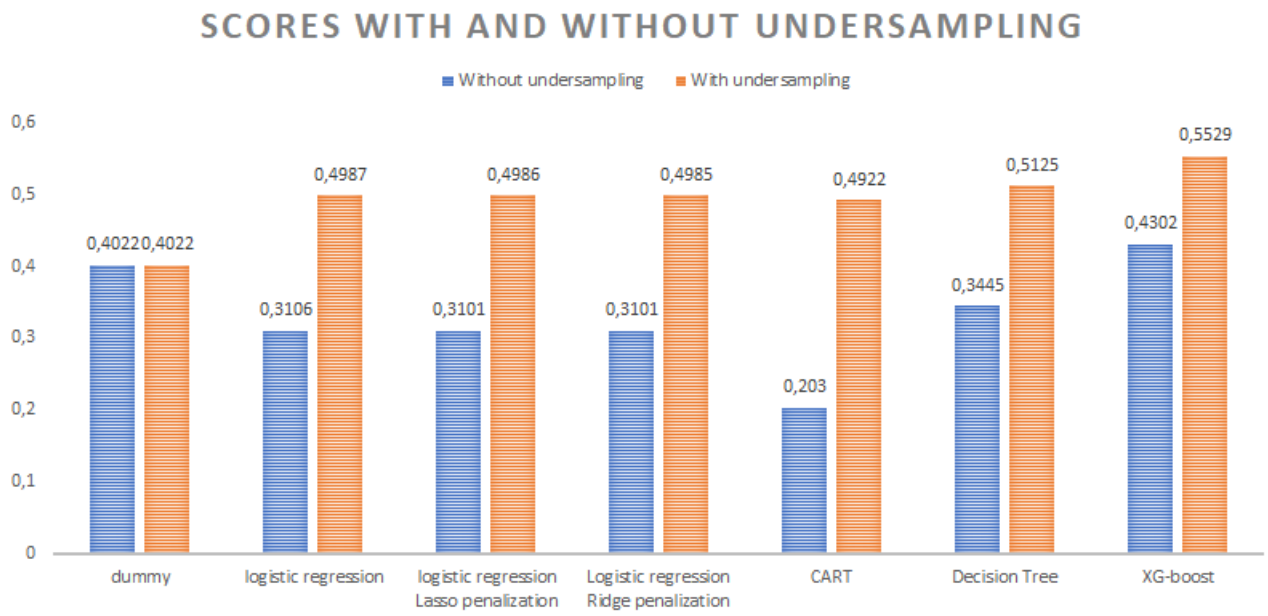

FIGURE 4.7 – En bleu : Le f1-score sans sous-échantillonnage sur le train set
En Orange : Le f1-score avec sous-échantillonnage sur le train set

Tous nos modèles sont améliorés grâce à l'utilisation du sous-échantillonnage. Comme dit avant, le fait d'avoir un train set équilibré en termes de classes aide les modèles à mieux distinguer entre nos deux classes d'intérêt, les opt-ins et les inactifs. Le modèle qui a le meilleur f1-score est le Xg-boost.

En outre du critère f1-score, qu'on adopte lorsqu'on veut privilégier la classe 1 qui est la classe des opt-ins, on compare le score AUC des modèles entrainées sur un train set sans rééchantillonnage. Cette deuxième vision, comme expliqué précédemment, nous aide à traiter les deux classes de la même manière et à garder leur proportions réelles. C'est ce modèle qu'on utilise pour prédire l'effet du PMIF et pour comparer son taux d'opt-in prédit avec le taux d'opt-in réel sur d'autres fonds.

Méthode utilisée	Score AUC obtenu
Régression logistique	0,7158
Arbre de décision	0,5955
Xg-boost	0,7519

TABLE 4.4 – Score AUC des modèles testées sans rééchantillonnage

Tous les modèles ont un score supérieur à 0.5. De ce fait, ils disposent tous d'un pouvoir prédictif. Le Xg boost est de nouveau le meilleur modèle en termes du score AUC. Nous allons consacrer la partie suivante à la démystification de ce modèle. Dans cette partie, nous nous intéressons plus à la compréhension de l'opt-in. Alors, nous utilisons le modèle entrainé avec sous-échantillonnage qui maximise le f1-score.

4.3.2 Vérification de la cohérence

Dans plusieurs cas, y compris le notre, un modèle complexe peut aboutir à des prédictions plus précises et un score meilleur qu'un modèle plus simple mais interprétable. Il s'agit de faire un compromis entre la performance et l'interprétabilité.

Nous voulons quand même comprendre les raisons qui sont derrière les décisions du XG-boost. Une première approche a consisté à regarder le graphe de l'importance des variables. Une seconde approche a consisté à visiter le graphe de dépendance partielle et l'algorithme LIME.

4.3.2.1 Comparaison des variables les plus importantes

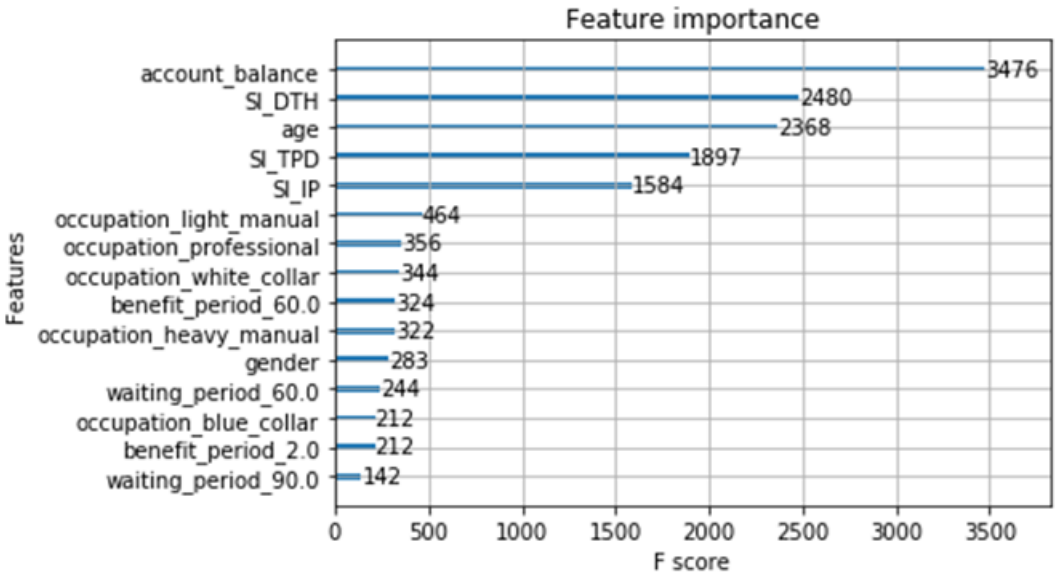

FIGURE 4.8 – Le graphe de l'importance des variables selon le XG-boost

Ce graphique représente l'importance relative des variables. Cette importance se calcule en utilisant le nombre de fois où la variable en question est sélectionnée pour la séparation d'un nœud à l'intérieur du modèle. De ce fait, elle traduit le poids de la variable dans la prise de décision.

Les variables account balance, les montants assurés et l'âge apparaissent en tête de ce classement. Ce constant est consistent avec la classification faite par l'arbre de décision, où ces variables sont utilisées pour faire les premières segmentations.

On souligne le fait que les variables account balance et age sont corrélées entre eux, un membre plus âgé aura un solde de superannuation plus élevé grâce à son ancienneté au sein du marché de travail.

4.3.2.2 Partial Dependence Plot

Le graphique de dépendance partielle ou Partial dependence graph (PDP), est un graphique qui montre l'effet marginal d'une variable sur la prédiction d'un classifieur. Sa force réside dans sa capacité d'identifier non seulement le sens (positif ou négatif) mais aussi la forme (linéaire, non linéaire, par paliers) de la relation. Il permet de comprendre l'impact

d'une variable en parcourant les différentes valeurs qu'elle peut prendre tout en moyennant l'influence des autres variables explicatives.

Formellement, considérons un modèle f qu'on essaye d'interpréter. Soit x_S, la variable dont on souhaite de repérer l'impact et x_C les autres variables explicatives que le modèle utilise pour prédire. Alors la fonction de dépendance partielle est définie par :

$$f_{x_S}(x_S) = \mathbb{E}[f(x_S, x_C)] = \int f(x_S, x_C) d\mathbb{P}(x_S)$$

C'est une fonction qui dépend uniquement de la variable x_S à laquelle on s'intéresse. En pratique, cette fonction est calculée d'une manière empirique en remplaçant l'intégrale par une moyenne :

$$\hat{f}_{x_S}(x_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

Avec n le nombre des instances dans notre jeu de données.

Dans notre cas de problème de classification binaire, la fonction f du modèle est généralement la probabilité que la variable cible vaut 1. Alors, le graphique de dépendance partielle affiche la probabilité attribuée à la classe 1 en fonction des différentes valeurs que la variable explicative d'intérêt x_S prend.

Ci-dessous nous traçons le graphique de dépendance partielle du Xg-boost optimisé et entraîné sur la base d'apprentissage sous-échantillonnée avec SMOTE. Il existe deux types du graphique qu'on peut visualiser et interpréter d'une façon intuitive : Le graphique à une dimension avec une seule variable explicative, et un graphique à deux dimensions où on regarde la réponse du modèle à deux variables explicatives en même temps.

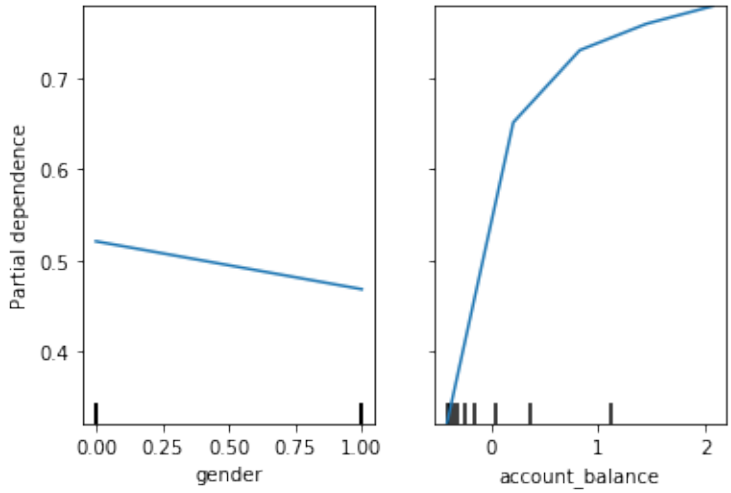

FIGURE 4.9 – Le partial dependence plot en fonction des variables genre et account balance

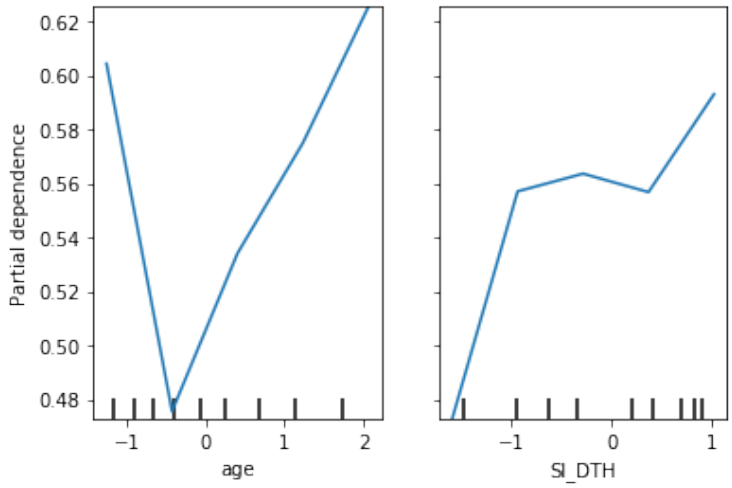

FIGURE 4.10 – Le partial dependence plot en fonction des variables âge et montant assuré décès

FIGURE 4.11 – Le partial dependence plot à une dimension en fonction du genre, account balance, âge et montant assuré décès

A partir de ce graphique il est possible de remarquer, par exemple, que le taux d'opt-in décroît en passant du genre 0 (femme) au genre 1 (homme). Il est possible aussi de voir que ce taux croît en fonction de la variable account balance. L'effet de la variable âge n'est pas monotone, le taux d'opt-in décroît puis recroît à partir de l'âge moyen des fonds.

Maintenant passons au graphique croisé, ci-dessous nous visualisons la dépendance partielle du modèle en fonction de l'account balance et le genre d'un côté et de l'account

balance et l'âge d'un autre. Le account balance, pour rappel, est l'une des variables clés de cette modélisation.

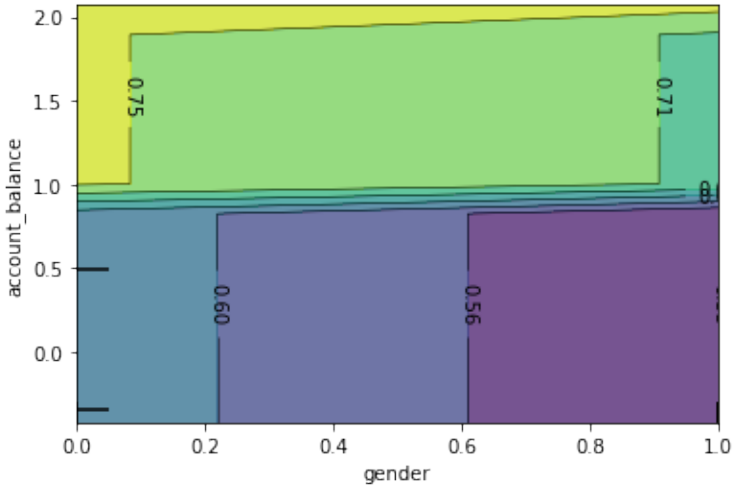

FIGURE 4.12 – Le partial dependence plot croisé des variables genre et account balance

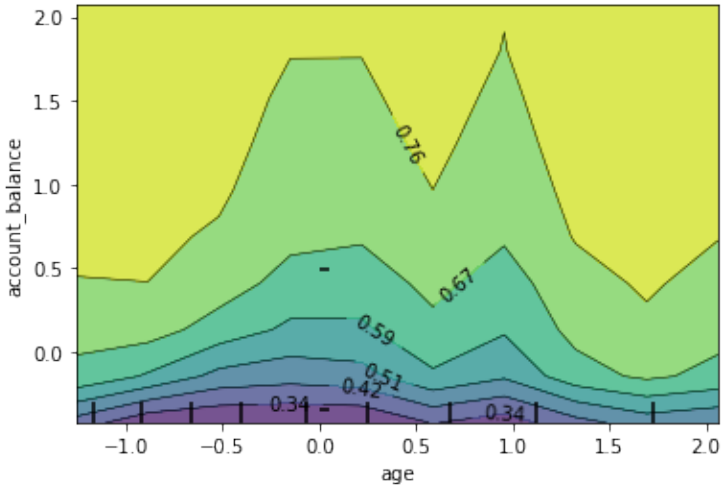

FIGURE 4.13 – Le partial dependence plot croisé des variables âge et account balance

FIGURE 4.14 – Le partial dependence plot à deux dimensions

Les couleurs foncés correspondent à des probabilités d'opt-in moins élevées. Sur le premier graphe nous pouvons voir la croissance des probabilités de prédiction d'opt-in avec le account balance et le genre. Le account balance de 1 en valeur normalisée coupe quasiment en deux les valeurs prédites.

L'effet croisé de l'âge et le account balance est moins linéaire et simple. En effet, ces

deux variables sont corrélées comme on a prouvé dans la partie dédiée à l'analyse des variables. On peut néanmoins discerner la croissance monotone en fonction du account balance et le changement de tendance en fonction de l'âge aux alentours de l'âge moyen.

4.3.2.3 LIME

L'algorithme LIME (ou Local interpretable Model-agnostic Explanations) est un modèle local qui cherche à porter des explications à la prédiction d'un individu par l'analyse de son voisinage. Cette méthode est agnostique dans la mesure où elle peut être utilisée avec n'importe quel modèle. LIME est interprétable et simple dans le voisinage d'un point qu'on choisit.

Soit x_0 une observation (un individu) au voisinage du quel on souhaite interpréter la prédiction. Nous disposons de $x -> f(x)$ qui est la fonction de décision du modèle complexe et difficile à interpréter. LIME crée de nouvelles données dans un voisinage proche de x_0 et construit $x -> g(x)$ un modèle interprétable à ce voisinage.

Le modèle g est plus transparent et joue le rôle de modèle de substitut pour interpréter les résultats du modèle complexe d'origine. Il est donc possible d'afficher pour certains individus choisis le modèle interprétable qui a permis de substituer le Xg-boost au voisinage de ce point. On peut voir les variables utilisées pour faire le choix et comment elles ont influé le choix du modèle.

Il est donc possible d'afficher pour certains individus choisis le modèle interprétable qui a permis de substituer le Xg-boost au voisinage de ce point. On peut voir les variables utilisées pour faire le choix et comment elles ont influé le choix du modèle.

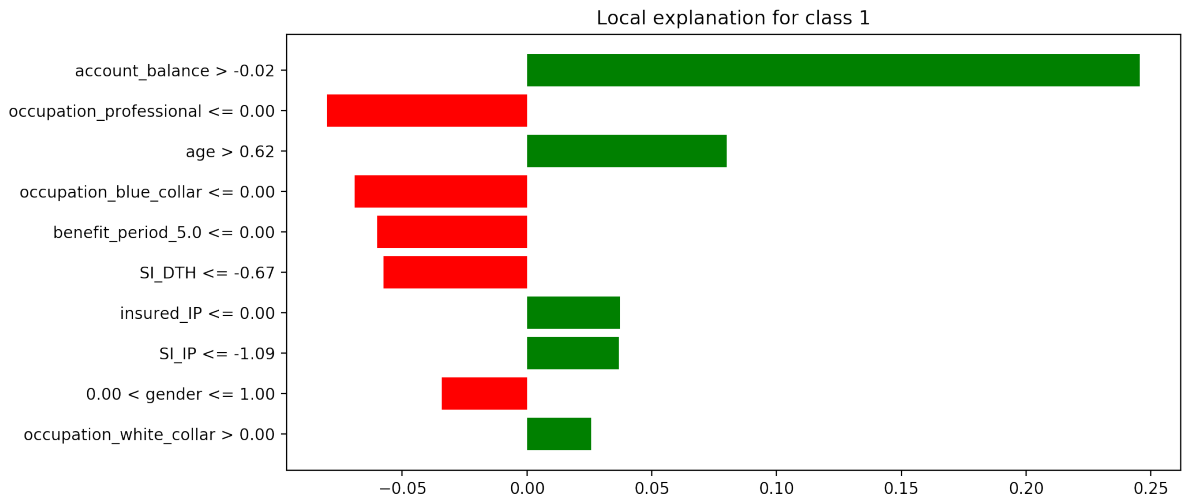

FIGURE 4.15 – Application de la méthode LIME sur un individu pour le modèle Xg-boost optimisé et entraîné sur la base souséchantillonnée.
En rouge : contredit l'opt-in. En vert : soutient l'opt-in.

Cet individu est correctement prédit comme opt-in. Le fait que son account balance en valeur normalisée est supérieur à 0 (c'est-à-dire, son vrai account balance est supérieur au solde de superannuation moyen) a joué en faveur de son comportement de faire un opt-in. Le fait que son âge en valeur normalisée est supérieur à 0.62 a aussi agi en faveur de son comportement, ce qui rejoint ce que nous avons remarqué au début : le taux d'opt-in est plus important pour les âges élevés. On peut voir aussi que le fait que gender=1, c'est-à-dire que cet individu est homme, a contredit le comportement d'opt-in. Ceci rejoint aussi notre constat que les hommes ont un taux d'opt-in moins élevé que les femmes.

En Résumé :
Dans cette partie nous avons essayé de modéliser le comportement d'opt-in chez les individus étudiés. Nous avons testé plusieurs modèles de classification en les comparant entre eux à l'aide du f1-score et AUC. Nous avons utilisé les techniques du sous-échantillonnage pour équilibrer les modèles et améliorer leurs scores.
Nous avons choisi le modèle Xg-boost pour modéliser ce comportement. Ce modèle étant performant, manque d'interprétabilité. Nous avons utilisé alors le graphe d'importance des variables, le graphique de dépendance partielle et LIME pour interpréter ses résultats.

Chapitre 5

Application de l'étude

Nous allons à présent présenter les applications de la modélisation. Une première application, et qui constituait d'ailleurs une des motivations de cette modélisation, est la prédiction du taux d'opt-in PMIF. En effet, Nous ne disposons pas de données suffisantes pour identifier les membres qui ont effectué un opt-in pour la réglementation PMIF. De ce fait, ce serait très important de pouvoir les approximer afin de prévoir leurs effets sur le profil de risque ainsi que sur la sinistralité future.

Une deuxième application, serait de voir si nous arrivons à généraliser les résultats obtenus sur d'autres fonds. L'idée serait de tester le modèle d'indentification des opt-ins basé sur les fonds 1, 2, 3 et 4 et voir si ses résultats peuvent être généralisés. Nous disposons en l'occurrence de données membership de deux autres fonds de superannuation qu'on appellera fonds 5 et 6.

5.1 Prédiction du PMIF opt-in

5.1.1 Calcul des impactés

Pour faire ce calcul, nous avons exclu les membres inactifs et les opt-ins de la base des données. Comme indiqué dans la première partie, les membres opt-ins du PYS qui ont un solde de superannuation inférieur à 6000 dollars australiens sont automatiquement considérés comme des opt-ins au PMIF.

A partir des actifs du PYS, nous avons identifié les impactés PMIF à l'aide de la variable

account balance. Les membres qui ont un solde inférieur à 6000 dollars australiens sont ainsi considérés comme impactés par la réglementation PMIF.

Fonds	Pourcentage des impactés
Fonds 1	28,3%
Fonds 2	14,7%
Fonds 3	41,7%
Fonds 4	27,8%

TABLE 5.1 – Pourcentage des impactés du PMIF par fonds

Le tableau ci-dessus récapitule le pourcentage des impactés par le changement réglementaire PMIF pour chacun nos 4 fonds étudiés depuis le début. On remarque que le pourcentage des impactés détenu par le fonds 3 est très élevé. On rappelle qu'il s'agit du fonds à destination des travailleurs dans la distribution (Retail) et qui est par conséquent le fonds le plus jeune du lot. Les membres de ce fonds n'ont pas passé à priori assez de temps dans le marché de travail et n'ont pas, en conséquence, accumulé assez d'argent dans leurs comptes de superannuation.

On rappelle que les pourcentages des impactés par le PYS sont les suivantes ; Fonds 1 : 30%, Fonds 2 : 16%, Fonds 3 : 27% et Fonds 4 : 43%.

Similairement au PYS, le Fonds 2 est celui qui détient le pourcentage des impactés le moins élevé. Pour rappel, il est le fonds à destination des fonctionnes et son âge moyen est de 46 ans. Vue la nature de leur occupation et leur âge relativement élevé ,dont on a prouvé la corrélation positive avec le account balance, les membres de ce fonds ont tendance à avoir des soldes de superannuation plus élevés.

Le fonds 1 dispose d'un pourcentage d'impactés par le PMIF relativement élevé. Cependant, ce pourcentage reste légèrement moins élevé que celui des impactés par le PYS. Le fonds 4 qui était le fonds le plus impacté par le PYS en termes de pourcentage des impactés, n'est plus en tête de ce classement. Il détient un pourcentage des impactés valant 27% ce qui s'approche du pourcentage obtenu dans le fonds 1.

5.1.2 Application du modèle et résultats

Fonds	Taux d'opt-in PMIF prédit	Taux d'opt-in PYS
Fonds 1	11,6%	11%
Fonds 2	2,8%	23%
Fonds 3	41,6%	24%
Fonds 4	27,4%	38%

TABLE 5.2 – Taux d'opt-in pour le PMIF prédit par le modèle et taux d'opt-in PYS observé par fonds

L'observation du marché a montré que le taux d'opt-in PMIF observé sur le marché est inférieur au taux d'opt-in PYS. Sur nos résultats, ce constat est vérifié pour le fonds 4, qui voit son taux se dégrader de 38% à 27%. Dans le cas du fonds 1, le taux d'opt-in du PMIF est très proche de celui du PYS. C'est un taux faible que l'on a associé auparavant à la présence d'une population non-engagée. Le taux obtenu chez le Fonds 2 est extrêmement faible et bien plus faible que le taux d'opt-in PYS. On sait aussi que ce fonds détient le pourcentage d'impactés le plus faible. Il est possible de présumer que ce fonds ne sera pas très impacté par le PMIF.

Chez le fonds 3, le modèle prédit un taux d'opt-in très grand. Ce fonds, déjà très touché en termes de pourcentage d'impactés, est le fonds qu'on juge le plus impacté par le PMIF en raison de sa démographie jeune.

5.2 Prédiction des PYS-opt-in d'autres fonds de superannuation

5.2.1 Les fonds 5 et 6

Nous disposons de nouvelles données issues de deux fonds de superannuation. On les nommera fonds 5 et fonds 6. Le tableau ci-dessous présente le profil d'âge et de genre de ces deux fonds.

Fonds	Pourcentage des femmes	Age moyen
Fonds 5	69%	45
Fonds 6	56%	42

TABLE 5.3 – Statistiques descriptives des fonds 5 et 6

Le fonds 5 est un fonds constitué généralement des fonctionnaires et travailleurs dans le secteur public. Son âge moyen s'approche de celui du fonds 2, ce dernier est à son tour à destination des fonctionnaires. La proportion des femmes membres de ce fonds est très élevée, elle dépasse les deux tiers. En 2019, ce fonds a eu approximativement 550 000 membres disposant d'une garantie décès. Par ailleurs, le fonds 6 est constitué de travailleurs dans les secteurs de santé, de l'éducation ainsi que du secteur public. Son âge moyen est de 42 ans et en 2019 avant le PYS, ce fonds avait près de 160 000 membres assurés contre le décès.

5.2.2 Adaptations et pré-traitement

L'assureur de ces deux fonds a exclu quelques individus de l'éligibilité au PMIF. En effet, les membres qui disposent des couvertures d'assurance optionnelles ou additionnelles sont considérés comme ayant déjà effectué le choix de rester couverts. Ils sont par suite exclus dans l'identification des impactés et des opt-in.

Et pareil, comme on ne disposait pas de flag d'opt-in, nous avons utilisé deux bases membership avant et après le PYS et la date de la dernière cotisation afin d'identifier les impactés par le PYS et les opt-in. Nous avons obtenu les valeurs suivantes :

Fonds	Pourcentage des impactés	Taux d'opt-in
Fonds 5	11%	33,49%
Fonds 6	6%	18,68%

TABLE 5.4 – Pourcentage des impactés et taux d'opt-in PYS des fonds 5 et 6

Nous remarquons que les deux fonds disposent d'un faible pourcentage de membres impactés. Le fonds 5 voit 11% de ces membres impactés alors que le fonds 6 a 6% de membres impactés. Concernant le taux d'opt-in, il est très élevé chez le fonds 5 qui est en même temps le fonds le plus âgé des deux. Ce taux de 33% s'approche des 38% du fonds

4. Le fonds 6 quant à lui a un taux d'opt-in de 18%. Ce taux est faible en regardant les taux d'opt-in obtenus sur nos fonds étudiés précédemment.

5.2.3 Résultats de la prédiction

Le tableau ci-dessus présente les résultats de l'application du modèle pour prédire les opt-ins sur les données des fonds 5 et 6. Sur ce tableau, nous montrons le taux d'opt-in réel, le taux d'opt-in prédit ainsi que le score AUC et le f1 score (à comparer avec le f1 score dummy) sur nos 2 fonds.

Fonds	Taux d'opt-in réel	Taux d'opt-in prédit	Score AUC	f1 score	f1 score dummy
Fonds 5	36,34%	44,87%	0,6039	0,5148	0,5018
Fonds 6	18,68%	18,32%	0,6102	0,3634	0,3148

TABLE 5.5 – Résultats et scores du modèle pour la prédiction du PYS opt-ins dans les fonds 5 et 6

Le modèle réussit à prédire pour le fonds 6 un taux d'opt-in qui s'approche beaucoup de la réalité. Son AUC est de 0.61 et son f1 score est de 0.36 sachant que le f1 score du modèle dummy est de 0.31. Cependant, pour le fonds 5, le modèle surestime beaucoup le taux d'opt-in. Il prédit un taux d'opt-in de 44% contre un taux réel de 36%. Pour rappel, le taux d'opt-in le plus haut obtenu sur nos 4 fonds est celui du fonds 4. Ce taux valait 38%. Le modèle prédit néanmoins que ce fonds présentera un taux d'opt-in important. En regardant les scores AUC et f1 score, on voit que son f1 score s'approche beaucoup du modèle dummy. En récapitulant, Notre modèle réussit bien à estimer une forte tendance d'opt-in chez les membres impactés du fonds 5. Cependant, il n'arrive pas à très bien généraliser ses résultats à cause des mauvais scores et de la valeur prédite trop amplifiée d'opt-in.

5.2.4 Pistes d'amélioration

Afin de pallier les faiblesses du modèle, nous pouvons d'une part envisager l'utilisation de plus de données. Ces données seront idéalement issues de plusieurs fonds de natures différentes.

D'autre part, nous pouvons customiser le score d'évaluation du modèle. Nous avons utilisé jusqu'à maintenant les scores AUC et le f1 score pour cette fin. Mais ces scores ne sont pas très intuitifs, et peuvent ne pas correspondre à ce qu'on veut.

Nous pourrions également utiliser les méthodes dites de voting. Ce sont des modèles qui utilisent les prédictions de plusieurs modèles (allant des modèles classiques de régression jusqu'aux modèles complexes de machine learning). Comme leur nom l'indique, elles tirent profit de plusieurs prédictions pour fournir une prédiction (une sorte de vote) alliant les points forts de tous les modèles utilisés. Le modèle sortant sera encore moins interprétable, mais plus puissant.

En Résumé :
Dans ce paragraphe, nous avons appliqué notre modèle pour pouvoir prédire l'impact du changement réglementaire PMIF. Nous avons analysé les résultats obtenus et jugé le fonds 3 comme le fonds le plus impacté par le PMIF.
Dans une seconde partie de ce chapitre, nous avons confronté notre modèle entraîné sur les fonds 1, 2, 3, 4 avec le comportement d'opt-in PYS observé sur deux nouveaux fonds 5 et 6. D'abord, nous avons analysé les taux d'opt-ins réels observés, nous les avons comparés avec ce que notre modèle prédit et enfin nous avons proposé quelques pistes d'amélioration pour encore polir le pouvoir généralisateur du modèle.

Conclusion

Cette étude avait pour but d'étudier l'impact de la nouvelle réglementation australienne PYS et de modéliser le comportement des assurés vis-à-vis de choisir de rester couvert ou pas.

Nous avons remarqué les premières tendances d'opt-in en regardant les taux d'opt-in par âge qui est un indicateur de risque, et en calculant les taux d'incidence des sinistres avant et après la réglementation et par catégorie.

Le changement du taux d'incidence avant et après le PYS nous sert de repère pour vérifier la cohérence des chargements déjà appliqués aux cédantes. Nous pouvons voir sur le tableau ci-dessus que le taux appliqué bouge généralement dans la bonne direction en termes de signe. Pour le fonds 4, qui est le fonds le plus touché par le PYS, on voit, à ce stade très tôt certes, que les chargements appliqués ne sont pas suffisants.

Fonds	PYS Impact on GL Frequency	GL Premium Loading for PYS	PYS Impact on IP Frequency	IP Premium Loading for PYS
Fonds 1	+12%	+13%		
Fonds 2	-14%	-6%	+2%	+10%
Fonds 3	+1%	+14%	+4%	+36%
Fonds 4	+34%	+52%	+33%	+22%

Table 5.6 – Récapitulatif et comparaison de l'impact PYS sur les taux d'incidence avec les taux de chargement appliqués

Nous avons isolé l'effet des caractéristiques observables par l'assureur de ce qui est caché derrière le fait de faire un opt-in sur la sinistralité. Ceci en effectuant une régression logistique ayant comme variable cible l'incidence d'un sinistre invalidité. Nous avons

prouvé alors la présence d'antisélection dans la mesure où les opt-ins contiennent une sous population qui est plus risquée de ce qui est prévu par l'assureur.

Pour la modélisation, nous avons testé plusieurs modèles et utilisé le sous-échantillonnage pour les renforcer. Notre choix est tombé sur un Xg-boost en nous basant sur sa performance. Ce modèle souvent synonyme d'un modèle boîte noire, n'était pas facile à interpréter. Nous avons alors utilisé le graphe d'importance des variables, le graphique de dépendance partielle ainsi que LIME pour divulguer ses mystères.

Comme application des résultats du modèle, nous avons appliqué le modèle obtenu pour prédire les opt-ins du PMIF et essayé de généraliser ses résultats sur le PYS pour d'autres fonds de superannuation.

Les résultats de cette étude ont été utilisés pour justifier la surtarification imposée après le PYS, et feront partie du Framework adopté pour comprendre les circonstances des changements réglementaires en superannuation australienne comme le PMIF. Certaines améliorations de cette étude pourront être mises en place.

Cette étude peut être améliorée de plusieurs manières, notamment par le biais de l'utilisation de plus de données pour renforcer notre compréhension du comportement des assurés. Nous pouvons également proposer de nouvelles estimations des taux IBNR qui ne réagit pas excessivement aux déclarations plus rapides par les assurés.

Au-delà des problématiques techniques et scientifiques liées à mon alternance, j'ai également pu me former sur l'environnement de la réassurance internationale et pu échanger avec des agents venant de l'autre bout du monde, chose qui m'a été très enrichissante.

Bibliographie

[1] *Superannuation : tout savoir sur la caisse de retraite australienne.* Article publié en Juin 2018 disponible sur lien.

[2] *Superannuation in Australia.* Page Wikipedia disponible sur lien.

[3] *Insurance through super : A financial safety net through your super.* Article disponible sur : lien.

[4] *'Protecting your super package' announced in the Federal Budget 2018-2019 ; Exposure Draft legislation released for consultation.* Article publié en mai 2018 disponible sur : lien.

[5] *Treasury Laws Amendment (Protecting Your Superannuation Package) Act 2019.* Texte de législation issu de Federal Register of Legislation disponible sur : lien.

[6] *What does Protecting Your Super mean for your insurance ?.* Article de Metlife Blog publié en juin 2019 disponible sur : lien.

[7] *Occupational ratings guide for insurance.* Guide de MLC insurance disponible sur : lien.

[8] *Séléction Adverse.* Page Wikipédia disponible sur : lien.

[9] Milan Vujisic. *L'asymétrie d'information.* Article publié en décembre 2017

[10] Mehdi Ben Hamouda. *Présence de l'asymétrie informationnelle sur le marché de l'assurance automobile.* Article publié en décembre 2017

[11] Yizhe Wang Damien Nouvel Marguerite Leenhardt Gaël Patin . *Apprentissage déséquilibré pour la détection des signaux de l'implication durable dans les conversations en parfumerie.*

[12] WikiStat. *Régression logistique ou modèle binomial* Université de Toulouse.

[13] Rémi Eyraud. *Classification, Apprentissage, Décision* Aix-Marseille Université.

[14] Ricco Rakotomalala. *Gradient Boosting : Technique ensembliste pour l'analyse prédictive* Université Lumière Lyon 2.

[15] Christoph MOLNAR. *Interpretable Machine Learning : A guide for making black box models explainable*. Février 2019.

[16] Manon VALDIQUIÉ. *Mesure du risque d'anti-sélection en Assurance Santé Collective*. Mémoire d'Actuariat présenté en Janvier 2017.

[17] Sophie FRANQUET. *Modélisation de la fréquence des sinistres graves en assurance automobile : apports et interprétabilité des méthodes d'apprentissage statistique*. Mémoire d'Actuariat présenté en Septembre 2018.

[18] Thomas BOUCHÉ. *Modèle de propension des assurés par rapport aux risques de sinistres corporels graves en assurance automobile.*. Mémoire d'Actuariat présenté en Septembre 2014.

[19] Héloise YVROUD. *Construction de lois de maintien en arrêt de travail en Australie par segmentation non-paramétrique de la population*. Mémoire d'Actuariat présenté en Novembre 2018.

[20] Ali ALICHE. *Invalidité et arrêt de travail en Australie*. Mémoire d'Actuariat présenté en Novembre 2018.

[21] Guillaume SCHLOSSER. *Modélisation de la mortalité dans le cadre du provisionnement de la garantie plancher en cas de décès : utilisation des méthodes de scoring*. Mémoire d'Actuariat présenté en 2019.

[22] Henri DESCAMPS. *Etude de méthodes de provisionnement individuelles et application sur un portefeuille auto*. Mémoire d'Actuariat présenté en 2017.

Table des figures

1 Le graphe de l'importance des variables selon le XG-boost 8

2 Application de la méthode LIME sur un individu pour le modèle Xg-boost optimisé et entraîné sur la base souséchantillonnée. En rouge : contredit l'opt-in. En vert : soutient l'opt-in. 8

3 The variable importance graph according to the XG-boost 12

4 Application of the LIME method on an individual for the optimized Xg-boost model trained on the under-sampled database. In red : contradicts the opt-in. In green : supports the opt-in. 12

2.1 Pourcentage de femmes et âge moyen par fonds 37

3.1 taux d'opt-in du fonds 1 40

3.2 taux d'opt-in du fonds 4 40

3.3 Evolution du taux d'opt-in par genre et tranche d'âge dans les fonds 1 et 4 40

3.4 La distribution de l'âge par catégorie Fonds 1 41

3.5 La distribution de l'âge par catégorie Fonds 4 41

3.6 La distribution de l'âge par catégorie Fonds 3 41

3.7 La distribution de l'âge par catégorie Fonds 2 41

3.8 L'évolution du taux de refus pour le risque TPD 45

3.9 L'évolution du taux de refus pour le risque IP 45

3.10 L'évolution du taux de refus pour les risques TPD et IP pour le fonds 3 . . 45

3.11 Schéma simplifié de la vie d'un sinistre avant son paiement 51

4.1 Exemple d'un arbre de décision : les losanges correspondent aux noeuds et les carrés aux feuilles. 63
4.2 Schéma simplifié du sur-échantillonnage 68
4.3 Schéma simplifié du sous-échantillonnage 69
4.4 Cadre de l'évaluation du problème de classification binaire 71
4.5 Matrice de confusion 72
4.6 F1 score par modèle sur un train set non modifié 76
4.7 En bleu : Le f1-score sans sous-échantillonnage sur le train set En Orange : Le f1-score avec sous-échantillonnage sur le train set 77
4.8 Le graphe de l'importance des variables selon le XG-boost 79
4.9 Le partial dependence plot en fonction des variables genre et account balance 81
4.10 Le partial dependence plot en fonction des variables âge et montant assuré décès 81
4.11 Le partial dependence plot à une dimension en fonction du genre, account balance, âge et montant assuré décès 81
4.12 Le partial dependence plot croisé des variables genre et account balance . . 82
4.13 Le partial dependence plot croisé des variables âge et account balance . . . 82
4.14 Le partial dependence plot à deux dimensions 82
4.15 Application de la méthode LIME sur un individu pour le modèle Xg-boost optimisé et entraîné sur la base souséchantillonnée. En rouge : contredit l'opt-in. En vert : soutient l'opt-in. 84