Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 15/03/2022

Par : **Vincent Habib**

Titre : **Création d'un score d'appétence à la diversification pour les clients épargne**

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 : Nicolas Baradel

Signature :

Membres présents du jury de l'Institut des Actuaires

Entreprise : Crédit Agricole Assurances

Directeur du mémoire en entreprise :

Nom : Merwan Amimeur
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

Remerciements

En premier lieu, je tiens à remercier mon tuteur Merwan Amimeur, pour sa supervision et ses précieux conseils tout au long de mon stage de fin d'études qui a débouché sur ce mémoire.

Je souhaite aussi remercier Jérôme Burnod, Responsable du service *Risques Suivi et Pilotage*, Stéphanie Brugirard, Responsable de l'équipe *Data & Risks Monitoring*, Margaux Humeau et Gaetan Beaud-De-Brive pour leur aide et leurs conseils avisés durant la conduite de ce projet.

J'adresse mes remerciements à l'ensemble de l'équipe pédagogique de l'ENSAE pour l'encadrement durant mon cursus d'ingénieur.

Je tiens également à remercier Allnassir Rajabaly, Directeur Financier de Crédit Agricole Creditor Insurance, pour ses conseils et de m'avoir orienté vers le domaine de l'actuariat.

Enfin, je remercie Vincent Drevot pour sa relecture minutieuse ainsi que Alexandra Elbakyan pour l'accès aux articles de recherche utilisés dans ce mémoire.

Résumé

La crise économique de 2008 ainsi que celle du Covid-19 ont eu des impacts importants sur le niveau des taux, entraînant par exemple en juin 2019 l'apparition de taux négatifs pour les OAT françaises. Cet environnement de taux bas a fortement diminué le rendement et la rentabilité des fonds euros en épargne pour les assurés et les assureurs. C'est dans ce cadre que Crédit Agricole Assurances souhaite aiguiller ses clients vers les fonds en Unité de Compte (UC) présentant une rentabilité plus élevée notamment vis-à-vis des fonds propres et un meilleur taux de rendement pour les assurés. Toutefois, cette incitation à investir dans les UC ne peut être proposée à l'ensemble des clients de Crédit Agricole Assurances, nécessitant donc un ciblage des assurés les plus à même de réaliser cette action. Ce mémoire a pour but de créer un score permettant d'accompagner les équipes marketing sur le ciblage client afin d'augmenter la part d'UC au sein du portefeuille de Crédit Agricole Assurances.

Pour créer ce score, nous avons tout d'abord procédé à la définition d'un critère d'appétence à la diversification, les clients vérifiant ce critère sont les clients de référence pour notre score. Nous avons ensuite cherché à prédire ce critère à l'aide d'un modèle de machine learning, l'*eXtrem Gradient Boosting* (XGBoost). L'interprétation des résultats du modèle a été effectuée à l'aide des valeurs de Shapley, nous permettant de connaître l'importance globale et locale des variables pour le modèle. Afin de créer, pour chaque variable, les catégories de notre score, nous avons eu recours, en partie, à l'algorithme de clustering BIRCH. Toutes ces étapes, nous ont permis d'obtenir à partir des valeurs de Shapley un score sur 1 000. Ce mémoire analyse ensuite l'adéquation du score créé, une fois celui-ci appliqué à l'ensemble des clients de notre base de données.

Abstract

The 2008 economic crisis and the one of the Covid-19 had significant impacts on the level of rates, leading for example to the appearance in June 2019 of negative rates for French bonds. This low interest rate environment reduced the yield and profitability of euro savings funds. It is in this context that Crédit Agricole Assurances wishes to direct its clients towards Unit of Account (UA) funds that have higher yield and profitability. However, this incentive to invest in UAs cannot be offered to all Crédit Agricole Assurances customers, thus requiring the targeting of policyholders most able to carry out this action. The purpose of this dissertation is to create a score for the marketing teams allowing them to assess the customers' appetite for diversification.

To create this score, we first of all defined a criterion of appetite for diversification, the customers verifying this criterion are the reference customers for our score. We then tried to predict this criterion using a machine learning model, the *eXtrem Gradient Boosting* (XGBoost). The interpretation of the model results was performed using Shapley values, allowing us to know the global and local importance of the variables fro the model. In order to create, for each variable, the categories of our score, we used, partly, the BIRCH clustering algorithm. All these steps allowed us to obtain a score out of 1000 from the Shapley values. This thesis then analyzes the adequacy of the score created, once it has been applied to all the clients in our database.

Note de synthèse

Le contexte actuel de taux bas, conséquence de la crise de 2008 et celle du Covid-19, a eu un fort impact sur l'activité des assureurs. En épargne, cela a entraîné une chute importante du taux de rendement des fonds euros, fonds sur lesquels sont placés la majorité des encours. Ainsi, Crédit Agricole Assurances souhaite inciter ses assurés à augmenter leur taux d'Unité de Compte (UC). Les UC ont l'avantage d'avoir un taux de rendement pour le client et de rentabilité pour l'assureur plus élevés que les fonds euros, en moyenne, et de voir le risque de marché porté par l'assuré, et non l'assureur, sur les UC. Ce mémoire a pour but de créer un score d'appétence à la diversification qui permettra de cibler les assurés sur lesquels mener des opérations marketing afin de pousser ces assurés à augmenter la part d'UC au sein de leurs polices d'épargne.

Afin de créer ce score, nous avons dû définir les clients de référence, les clients appétents à la diversification. Ces derniers sont ceux qui vont au-delà de la politique de collecte de Crédit Agricole Assurances soit les personnes ayant :

— Plus de 30% d'UC si l'encours est inférieur à 40 000€

— Plus de 40% d'UC si l'encours est supérieur à 40 000€

Les personnes appétentes à la diversification représentent 18% des clients de la base de données servant à construire le score. Cette dernière fait de 5,2 millions de ligne et est à la maille client. La période d'étude de notre base de données s'étale du 31/03/2019 au 31/03/2021. Elle est découpée en 2 parties, une première allant du 31/03/20219 au 31/03/2020 d'où sont issues les données de la base et une période s'étalant du 31/03/2020 au 31/03/2021 où est calculée l'appétence à la diversification du client. La base de données contient une centaine de variables ayant trait aux caractéristiques du client, à l'état de ses assurances épargne et prévoyance ainsi qu'aux mouvements qu'il a pu effectuer sur ses assurances épargne durant la période d'étude.

Le profil du client moyen correspond à une personne ayant 55 ans, 14 ans d'ancienneté, et un taux d'UC de 15%. L'encours médian sur l'ensemble des clients de la base est de 8 554€. Les clients appétents vont se distinguer du client moyen notamment sur leur taux d'UC moyen qui est de 56,2% et l'encours médian des clients appétents est de près de 18 000€.

Afin de créer ce score d'appétence à la diversification, nous avons tout d'abord cherché à prédire l'appétence future de nos clients. Les 10 variables que nous avons retenues pour effectuer la modélisation sont les suivantes :

— Pourcentage d'UC au 31/03/2020

— Nombre de polices d'épargne multi-support détenues par le client

— PM totale au 31/03/2020

— Montant du revenu fiscal

— Âge

— Ancienneté

— Code Profession

— Montant des versements programmés entre le 31/03/20219 et le 31/03/2020

— Pourcentage d'action parmi les UC

— Pourcentage d'obligation parmi les UC

Dans un premier temps, nous avons dû rééquilibrer la base de données à l'aide de l'algorithme SMOTE qui mélange le sur et sous-échantillonage. La modélisation de l'appétence future des clients s'est effectuée via le modèle de machine learning XGBoost, qui est un algorithme de Gradient Boosting. Le XGBoost a été très performant dans la prédiction de l'appétence avec un AUC-PR obtenu de 0,96. Un AUC-PR aussi élevé nous permet de nous assurer que le score que nous créons par la suite sera précis.

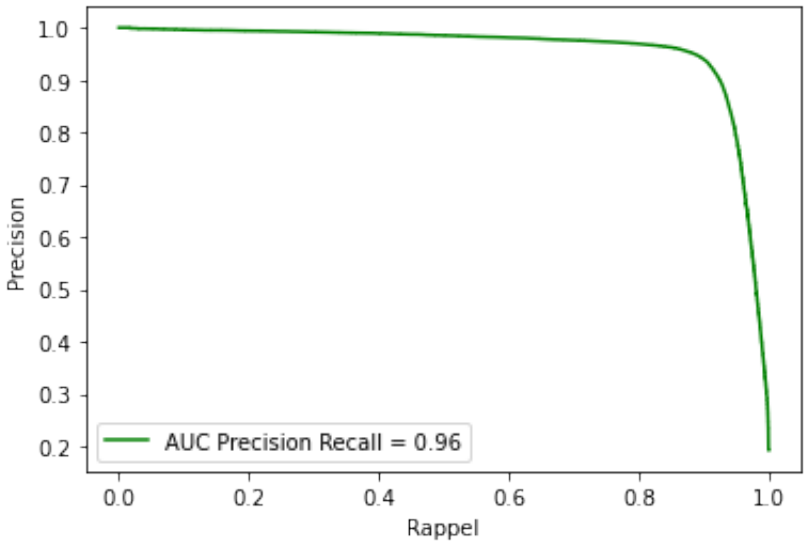

FIGURE 1 – Courbe Précision-Rappel obtenue pour notre modèle

Le XGBoost n'étant pas un modèle linéaire, nous avons recours aux valeurs de Shapley afin d'interpréter les résultats de la modélisation. Les valeurs de Shapley permettent de connaître l'importance globale des variables pour le modèle et comment cette importance évolue en fonction des valeurs prises par les variables. Afin de passer des valeurs de Shapley à un score, nous avons dû discrétiser les variables continues utilisées. Cette discrétisation s'est faite en fonction des observations effectuées sur la courbe montrant l'évolution des valeurs de Shapley en fonction des valeurs prises par la variable et à l'aide de l'algorithme de clustering BIRCH. En prenant la valeur de Shapley moyenne pour chacune des catégories créées et avec quelques retraitements, nous obtenons un score sur 1 000.

Lorsqu'on applique ce score sur 1 000 à l'ensemble des clients de notre base, on distingue, tout d'abord, que la répartition du score pour les clients appétents et pour les clients non-appétents sont distinctes l'une de l'autre. Ceci nous permet d'affirmer que notre score est bien capable de différencier nos clients de référence.

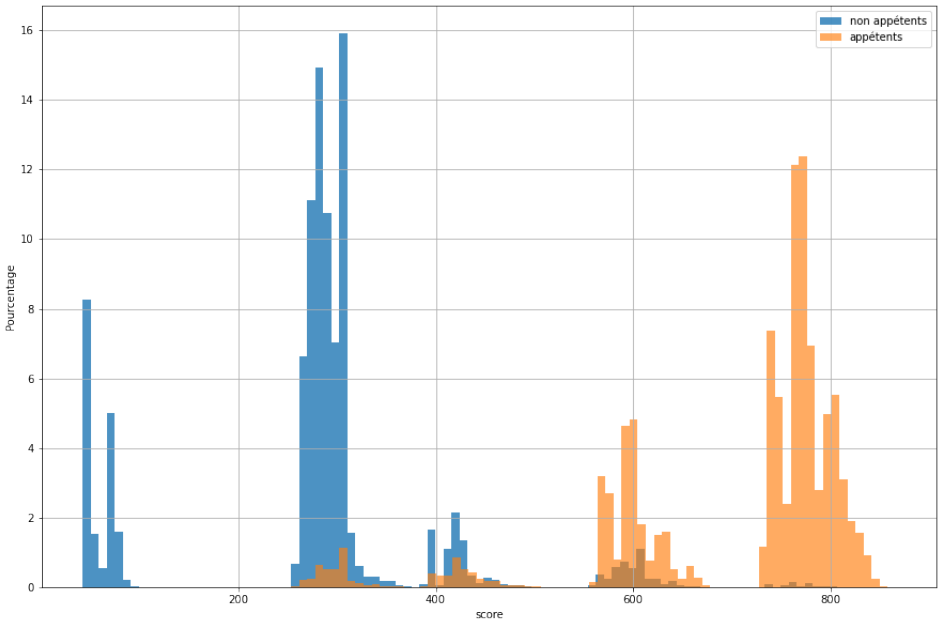

FIGURE 2 – Répartition des scores selon l'appétence des clients

L'affichage de la répartition du score au sein des clients de notre base nous permet de discerner 5 pics, qualifiés de classes d'appétence. On identifie :

— Le groupe des retraités, qui correspond aux personnes ayant un score entre 40 et 100. Il est dénommé ainsi car il est majoritaire composé de retraités. Il contient 14% des clients de notre base. Les membres de ce groupe ne possèdent pas d'UC et il semble peu probable de les voir prendre des UC.

— Le groupe des petits épargnants contient les personnes ayant un score allant de 250 à 330. Ce groupe est ainsi nommé du fait que le montant de l'encours médian est de près de 4 600€, soit presque 2 fois moins que l'encours médian sur l'ensemble de la base. Ce groupe contient 58% des clients de notre base. Les membres de ce groupe possèdent très peu ou pas d'UC. Il paraît donc peu probable de réussir à les amener à placer une partie de leur encours sur les UC.

— Le groupe du client moyen comprend les personnes ayant un score entre 380 et 450 points. Les caractéristiques des clients de ce groupe est très proche de celui du client moyen. 6% des clients de la base appartiennent à ce groupe. Seul le taux d'UC moyen est un peu plus élevé, de 23% pour ce groupe contre 15% pour l'ensemble de la base.

— Avec un score entre 550 et 680, on retrouve le groupe des petits appétents, qui représente 8% des clients de la base. Les clients de cette classe ont, en moyenne, un taux d'UC de 32%. Les petits appétents sont un potentiel groupe cible dont les membres peuvent être amenés à augmenter leur taux d'UC.

— Le dernier groupe identifié est celui des grands appétents. Il regroupe les clients ayant entre 720 et 850 soit 13% des clients de la base. Le taux d'UC moyen des membres de ce groupe est de 66% et l'encours moyen est de 74 420€. Ces 2 valeurs sont particulièrement élevées comparées à l'ensemble de la base. Ce groupe est donc composé de clients que l'ont peut potentiellement inciter à augmenter leur taux d'UC.

Le score que nous avons obtenu nous a permis de bien séparer les clients appétents des non-appétents et d'identifier 5 classes d'appétences qui ont toutes des caractéristiques propres. Parmi ces 5 classes d'appétences, 2 semblent pouvoir être aisément ciblées par les équipes marketing afin de les pousser à augmenter leur taux d'UC.

Executive Summary

The current context of low interest rates, a consequence of the 2008 and the Covid-19 crisis, has had a strong impact on the activity of insurers. In savings, this has led to a significant drop in the rate of return on euro funds, funds in which the majority of outstanding amounts are placed. Thus, Crédit Agricole Assurances wishes to encourage its policyholders to increase their Unit of Account (UA) rate. UAs have the advantage of having a higher rate of return and of profitability than euro funds, on average. The purpose of this dissertation is to create a diversification appetite score that will allow to target savings policyholders on whom to conduct marketing operations in order to encourage them to increase the share of UAs in their savings policies.

In order to create this score, we had to define the reference customers, customers that have an appetite for diversification. These are defined as those who go beyond Crédit Agricole Assurances collection policy, i.e. people with :

— More than 30 % UAs if the outstanding amount is less than 40,000€

— More than 40 % UAs if the outstanding amount is greater than 40,000 €

People with an appetite for diversification represent 18% of clients in the database used to build the score. The latter is composed of 5.2 million lines and is at the customer level. The study period runs from the 31/03/2019 to the 31/03/2021. It is divided into 2 parts, the first going from 31/03/20219 to 31/03/2020 from which the data are taken and a period spanning from 31/03/2020 to 31/03/2021 where the customer's appetite for diversification is calculated. The database contains around a hundred variables relating to the characteristics of the client, the state of his savings and provident insurance as well as the movements he may have made on his savings insurance during the study period.

The profile of the average customer corresponds to a person being 55 years old, having 14 years of seniority and a UAs rate of 15%. The median outstanding amount for all of the base's customers is of 8,554€. Appetite customers will stand out from the average customer in particular on their average UA rate which is of 56.2% and the median outstanding amount of appetite customers is nearly 18,000 €.

In order to create this diversification appetite score, we first tried to predict the future appetite of our customers. The 10 variables we retained to carry out the modeling are the following :

— Percentage of UA on the 31/03/2020

— Number of multi-medium savings policies held by the client

— Total outstanding amount on the 31/03/2020

— Amount of tax income

— Age

— Seniority

— Job

— Amount of payments scheduled between 31/03/20219 and 31/03/2020

— Percentage of stock among UAs

— Percentage of bonds among UAs

First, we had to rebalance the database using the SMOTE algorithm which mixes over and under-sampling. The modeling of future customer appetite was done via the machine learning model XGBoost, which is a Gradient Boosting algorithm. The XGBoost performed very well in predicting palatability with an AUC-PR obtained of 0.96. Such a high AUC-PR allows us to make sure that the score we create afterwards will be accurate.

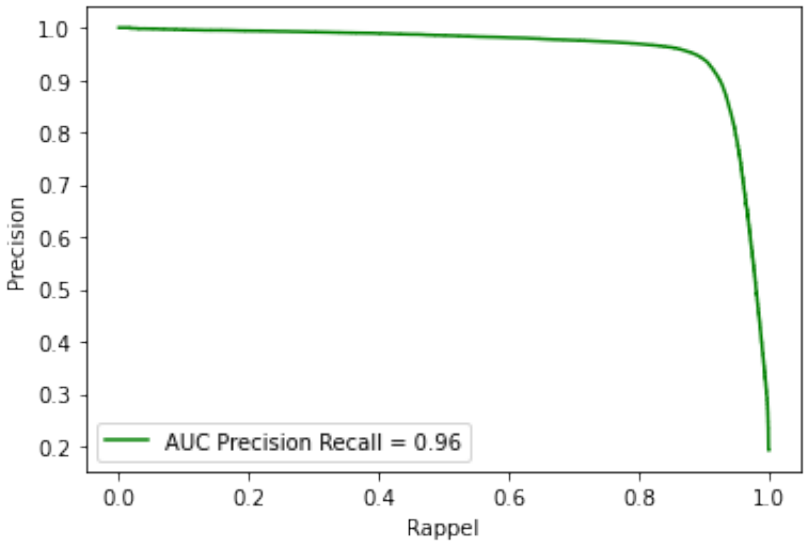

FIGURE 3 – Precision-Recall curve obtained for our model

As XGBoost is not a linear model, we use Shapley values to interpret the modeling results. The Shapley values make it possible to know the global importance of the variables for the model and how this importance evolves according to the values taken by the variables. In order to pass from the Shapley values to a score, we had to discretize the continuous variables used in the model. This discretization was done according to the observations made on the curve showing the evolution of the Shapley values given the values taken by the variable and using the BIRCH clustering algorithm. By taking the average Shapley value for each of the categories created and with some reprocessing, we get a score out of 1000.

When we apply this score out of 1000 to all the customers in our database, we can, first of all, distinguish that the distribution of the score between customers that have an appetite for diversification and those who don't are distinct from one another. This allows us to affirm that our score is able to differentiate our reference customers.

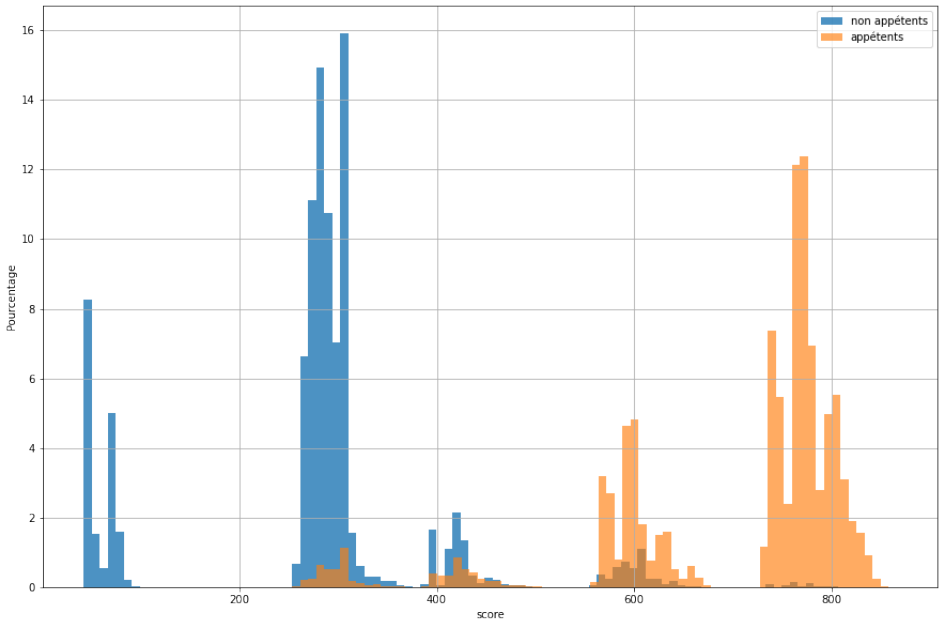

FIGURE 4 – Distribution of scores according to customer appetite for diversification

The display of the distribution of the score among the clients of our database allows us to discern 5 peaks qualified as appetite classes. We identify :

— the group of retired people, which corresponds to people with a score between 40 and 100. It is so called because it is mainly composed of retired persons. It contains 14% of our database customers. The members of this group do not have UAs and it seems difficult to be able to convince them to get UAs.

— the group of small savers contains people with a score ranging from 250 to 330. This group is so named because the median amount of outstanding is almost of 4,600€, which is almost 2 times less than the median outstandings across the database. This group contains 58% of the clients in our database. Members of this group have very little or no UAs. It therefore seems unlikely to be successful in getting them to place part of their assets in UAs.

— the average customer group includes people with a score between 380 and 450 points. The characteristics of customers in this group is very close to the ones of the average customer on the entire databse. 6% of the database's clients belong to this group. Only the average UA rate is a little higher, 23% for this group against 15% for the entire base.

— with a score between 550 and 680, we find the group of small appetites, which represents 8% of the base customers. Customers in this class have, on average, a UAs rate of 32%. Small appetites are a potential target group whose members may be encouraged to increase their UA rate.

— the last group identified is that of the great appetites. It groups together clients with between 720 and 850, i.e 13% of the base's clients. The average UA rate of the members of this group is of 66% and the average outstanding amount is of 74,420 €. These 2 values are particularly high compared to the whole database. This group is therefore made up of customers who have the potential to increase their UA rate given the right incentives.

The score we created allowed us to separate the customers with an appetite for diversification from the who don't and to identify 5 classes of diversification appetite which all have their own characteristics. Among these 5 diversification appetite classes, 2 seem to be able to be easily targeted by the marketing teams in order to push them to increase their UA rate.

Table des matières

1 Introduction

En 2020, la France s'est financée pour la première fois, en moyenne sur l'année, à un taux négatif de -0,14% pour ses obligations de long et moyen terme [3]. Ce taux négatif est la conséquence des programmes de Quantitative Easing mis en place par la BCE depuis 2015 et des nouveaux outils de politique monétaire accommodante mis en oeuvre suite à la pandémie du Covid-19. Ce contexte de taux bas, voire négatif, en place depuis plusieurs années déjà, a un fort impact sur les activités des assureurs.

L'assurance-vie, premier secteur en terme d'encours pour les assureurs français, est fortement affectée par cet environnement de taux bas. En effet, la majeure partie des encours est placée sur des fonds euros, composés à plus de 75% d'obligations [2]. Ces dernières, et notamment les obligations souveraines, ont vu leur taux de rendement diminuer fortement ces dernières années, ce qui a une forte incidence sur la rentabilité des produits d'épargne pour les assureurs et les assurés.

Pour compenser cette baisse, les assureurs cherchent à orienter leurs clients vers les Unités de Compte (UC) qui ont l'avantage d'être plus rentables pour les assureurs [8], d'avoir un meilleur taux de rendement, en moyenne, pour les assurés et de faire porter le risque par l'assuré. Pour cela, Crédit Agricole Assurances souhaite pouvoir accompagner ses équipes marketing sur le ciblage client dans le but d'augmenter la part d'UC au sein de son portefeuille. Ce mémoire a donc pour but de créer un score d'appétence à la diversification pour les clients épargne à partir de modèles de machine learning et à destination des équipes marketing. Ces travaux s'inscrivent dans la continuité des études menées, au sein de Crédit Agricole Assurances, sur les comportements des clients épargne, afin d'accroître la part d'UC au sein de son portefeuille, et dans la politique interne de l'entreprise de développement de l'utilisation des outils de Data Science.

Après la présentation du contexte et de l'objectif de cette étude, ce mémoire abordera les caractéristiques de la base de données construite à partir des données épargne et prévoyance de Crédit Agricole Assurances. Dans un troisième temps, sera présenté les modèles utilisés pour la prédiction de l'appétence de nos clients ainsi que les différentes méthodes employées pour interpréter ces résultats et construire le score. Enfin, les résultats obtenus seront détaillés.

2 Contexte et objet de l'étude

2.1 Présentation de l'assurance-vie

En 2020, l'épargne financière des français a atteint un montant record de 180 milliards d'euros [8]. Parmi les supports utilisés par les ménages français pour le placement de leur épargne, on retrouve l'assurance-vie. Il s'agit d'un contrat permettant le versement d'un capital ou d'une rente au souscripteur ou au(x) bénéficiaire(s) désigné(s) par le contrat.

Les contrats d'assurance-vie peuvent être décomposés en 2 familles :

— Les contrats d'épargne : ils permettent de faire fructifier le capital de l'assuré à travers le temps. Les primes versées sont investies par l'assureur sur les marchés financiers. Ces contrats permettent, selon leurs caractéristiques, de profiter, via la participation aux bénéfices, des bons résultats des placements de l'assureur, de sécuriser le capital et les intérêts versés. Ils peuvent permettre de préparer sa retraite et/ou de transmettre son patrimoine.

— Les contrats prévoyance : ce sont les contrats couvrant le "risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail ou d'invalidité ou du risque chômage" (Loi EVIN 1989). Ils servent à protéger les personnes couvertes par le contrat face aux aléas de la vie.

Dans ce mémoire, nous nous intéresserons plus précisément aux contrats d'épargne. L'argent versé par l'assuré sur son contrat d'épargne peut être placé sur 2 types de fonds différents :

— Les fonds euros : ce sont des fonds sans risque pour l'assuré garantissant la sécurité du capital, il est impossible de perdre le capital présent dessus (hors frais de gestion) et possédant un effet cliquet, les intérêts obtenus sur les fonds euros le sont définitivement. Le risque pour les fonds euros est porté par l'assureur. Plus précisément, l'assureur fait face à un risque de taux et de marché alors que l'assuré porte seulement le risque de contrepartie vis-à-vis de l'assureur.Le rendement de ces fonds est, en général, peu élevé.

— Les fonds en Unité de Compte (UC) : A l'inverse des fonds euros, le risque est porté par l'assuré dans les fonds en Unité de Compte. Le montant de l'encours n'est pas garanti, l'assureur garantit seulement le nombre de parts des supports UC et non leur valeur. En effet, les primes versées permettent de disposer d'un certain nombre de parts de support d'investissement dont la valeur fluctue avec les marchés financiers. Il est donc possible d'obtenir un rendement négatif sur les fonds UC. Donc l'assuré fait face au risque de marché en plus du risque de contrepartie. Il existe plusieurs sortes de supports UC selon l'actif financier sur lequel ils sont investis (actions, obligations, immobilier,...) et selon leur niveau de risque. Les fonds UC offrent, en général, un rendement plus élevé que les fonds euros [8].

Le marché de l'épargne contient 2 sortes de contrat : les contrats monosupports et les contrats multi-supports où une partie de l'encours est placée sur des fonds euros et une partie sur des fonds UC. Pour les contrats multi-supports, l'assuré a la possibilité d'effectuer des arbitrages, c'est-à-dire modifier la répartition du capital entre les différents supports, afin d'obtenir le taux de rendement maximum selon l'évolution de son profil de risque et de la situation sur les marchés financiers. Il existe plusieurs modes de gestion des contrats d'épargne :

— La gestion libre : l'assuré choisit seul la répartition de son capital entre les différents fonds et supports selon son profil de risque. C'est le mode de gestion par défaut des contrats multi-supports.

— La gestion pilotée : des options d'arbitrage automatiques sont proposées à l'assuré en plus de la gestion libre. Parmi les options les plus courantes, on retrouve celles de sécurisation des plus values et de limitation des pertes (*stop loss* en anglais).

— La gestion profilée : la gestion de votre contrat multi-supports est confiée à une société de gestion (peut être l'assureur ou une entreprise tiers homologuée par l'assureur) qui adapte sa gestion au profil de risque de l'assuré.

— La gestion à horizon : l'assuré communique une date à laquelle il souhaite racheter son contrat et l'assureur adapte sa gestion en fonction de l'approche de la dite date. Par exemple, pour un départ à la retraite, la gestion va être plus offensive sur les premières années avant de devenir plus prudente avec le temps.

— La gestion sous mandat : comme pour la gestion profilée, la gestion du contrat est confiée à une société de gestion mais l'assuré n'a pas la possibilité d'adapter la répartition sur les supports en fonction de son profil de risque.

Outre la performance des supports UC et les intérêts garantis des fonds euros, la rentabilité des contrats d'épargne, pour l'assuré, va dépendre de la participation aux bénéfices. C'est une partie des bénéfices de l'assureur que la réglementation oblige à reverser à ses clients.

Les assurés ont la possibilité d'effectuer des rachats partiels ou totaux de leur contrat d'épargne. Ceci correspond au fait de récupérer une partie ou tout l'encours de leur contrat d'épargne. Si un rachat total est effectué, le contrat prend fin. En cas de rachats partiels ou totaux, seuls les intérêts sont taxés.

Une fiscalité spécifique s'applique sur les contrats selon l'ancienneté. Pour les produits provenant de versements effectués avant le 27/09/2017, la fiscalité appliquée est celle d'avant la réforme de 2018 : un taux de 7,5% pour plus de huit ans d'ancienneté, 15% lorsque l'ancienneté est entre 4 et 8 ans et 35% dans les autres cas. Pour les produits provenant de versements effectués après le 27/09/2017, un taux de 12,8% est appliqué pour les produits ayant moins de 8 ans d'ancienneté et 7,5% ceux ayant une ancienneté supérieure [15].

En cas de décès de l'assuré, la fiscalité sur les contrats d'épargne dépend aussi de l'âge auquel l'assuré a effectué des versements. En effet, après l'âge de 70 ans, les versements réalisés sont taxés selon la fiscalité sur les droits de succession après un abattement de 30 500€. Toutefois, les intérêts constitués à partir des versements effectués après l'âge de 70 ans sont exonérés des droits de successions [13].

Aujourd'hui, les assureurs et notamment leurs activités en épargne font face à un contexte particulier qui est celui des taux bas, voire négatifs, actuellement constatés sur les marchés financiers. Cette situation particulière ainsi que ses effets sur les assureurs est décrite dans la partie ci-dessous.

2.2 Contexte des taux bas et objectifs de l'étude

Depuis le milieu des années 80, on constate une baisse quasi-continue des taux d'intérêts qui s'est accélérée depuis la crise financière de 2008 avec la mise en place de politiques monétaires accommodantes et d'un programme de Quantitative Easing par la BCE depuis 2015. Ainsi, en janvier 2021, le taux de dépôt de la zone euro est à -0,5% et le taux de refinancement à 0%. Cette situation de taux bas voire négatifs risque de continuer pendant plusieurs années, du fait de prévisions d'inflation basses ainsi que le prolongement des politiques monétaires accommodantes suite à la crise liée à la Covid-19 avec par exemple, la mise en place par la BCE du Pandemic Emergency Purchase Programme (PEPP).

Cet environnement de taux bas voire négatifs est problématique pour les assureurs. En effet, de par l'inversion du cycle de production en assurance, les assureurs perçoivent d'abord les primes et versent ensuite les prestations. Ainsi, afin de pouvoir réaliser les promesses faites aux assurés, ils possèdent des provisions techniques sous forme d'actifs. Ces derniers doivent être peu risqués, rentables et liquides pour pouvoir ensuite permettre de régler les prestations que l'assureur doit au titre des ses engagements le moment venu.

L'assurance-vie est de loin le premier secteur, en terme d'actifs détenus, pour les assureurs français avec près de 91% des actifs détenus par des assureurs français placés par des assureurs-vie [2]. Les assureurs privilégient pour les fonds euros, les investissements à revenus réguliers et à valeur de remboursement stable. Ainsi, ces derniers sont composés à plus de 75% d'obligations [2]. Pour les fonds UC, l'investissement se fait principalement via des Organismes de Placement Collectifs (OPC) qui vont mettre aux alentours de 50% des encours sur des actions [2]. La part des contrats en UC augmente d'années en années (14,8% en 2018, 15,8% en 2019 [2]) mais les contrats euros restent très largement majoritaires dans les produits d'épargne choisis par les épargnants français.

Ainsi, cet environnement de taux bas a des retombées très importantes sur les fonds euros, ces derniers privilégiant les obligations. Cependant, il y a un décalage entre la baisse des taux obligataires et celle des taux de revalorisation des supports en euros,

la première étant plus importante. En effet, les assureurs possèdent des obligations anciennes datant d'il y a plusieurs années et ayant un taux de rendement relativement plus élevé que les obligations actuelles. Toutefois, plus le temps passe, plus ces obligations arrivent à échéance. Les assureurs sont donc dans l'impossibilité d'avoir accès à des obligations avec des taux aussi intéressants. Ainsi, la baisse des taux sur les nouveaux titres acquis entraîne une baisse du rendement des actifs détenus par l'assureur. La Banque de France estimait qu'en 2019, "69 % du portefeuille d'obligations à taux de coupon fixe dont la maturité résiduelle est supérieure à quatre ans affichent un taux inférieur à 3 %, contre 40 % pour le portefeuille dont la maturité résiduelle est strictement inférieure à quatre ans" [2]. En 2017, elle calculait aussi que « Le réinvestissement des nominaux obligataires arrivant à échéance par des obligations à rendement nul (NDLA : les taux d'intérêts actuels sur les obligations souveraines des grands pays sont pratiquement nuls voire négatifs), la baisse du taux de rendement de l'actif au cours des 10 prochaines années serait de l'ordre 20 points de base par an. » [1].

Le taux de rendement des actifs connaît une forte baisse ces dernières années, en passant de 5,1% en 2006 à 3,4% en 2015 avec la baisse la plus forte enregistrée sur les obligations souveraines passant de 3,9% à 0,9% en moyenne sur cette période [1]. Cette baisse entraîne mécaniquement celle du taux servi mais celle-ci n'a pas été aussi forte que la diminution du taux de rendement des actifs. Cela conduit donc au risque de voir les taux de rendements du marché devenir inférieurs au taux garanti aux assurés et donc que la rentabilité du portefeuille ne soit plus suffisante pour financer les promesses faites dans les contrats d'assurance. Ceci engendre ainsi une pression sur les marges et les coûts des assureurs. De plus, les réinvestissements des titres et primes à des taux plus bas rend vulnérable l'assureur à une remontée des taux qui provoquerai des moins-values latentes et des risques de rachat, le tout faisant peser un risque de solvabilité pour l'assureur. Ainsi, les assureurs vie ont diminué le taux de participation aux résultats de 4,5% en 2006 à 3,2% en 2015, le taux de participation aux bénéfices de 2,7% à 2% sur cette période [1].

Cette baisse des taux a incité les assureurs à accélérer la transformation de leur modèle. Concernant les placements des assureurs vie, ils ont augmenté leurs investissements sur les actifs non amortissables. La part des obligations dans les placements hors UC en valeur nette comptable des 15 principaux assureurs vie et mixtes (avant transpa-

risation) est passée de de 68 à 64 % de fin 2013 à 2015 pendant que la part des titres est passée de 26 à 29 % [1].

Les assureurs ont aussi cherché à agir sur leur passif en développant la part des contrats UC dans leur portefeuille et plus particulièrement ceux à revenus variables. On constate en effet une décollecte des supports euros au profit des supports UC depuis mi-2016 car le taux de rendement net de frais moyen sur 10 ans pour les supports UC est de 2,8% tandis que le celui des support euros est passé d'une valeur supérieure à 2,8% à 1,3% en 2020. Pour les nouveaux contrats en euros souscrits, les assureurs ont diminué les taux techniques pour être en adéquation avec les évolutions du marché.

Crédit Agricole Assurances a donc imposé récemment une politique de collecte à ses assurés souscrivant un nouveau contrat épargne avec un taux minimum de 30% d'UC pour les personnes amenées à dépasser les 40 000€ d'encours.

Ainsi, c'est dans cet environnement de taux bas et de volonté pour Crédit Agricole Assurances de voir augmenter la part d'UC au sein des contrats épargne de ses clients que s'inscrit le sujet de ce mémoire. Ce dernier aura pour but de créer un score d'appétence à la diversification permettant aux équipes marketing de cibler leurs actions sur les clients pouvant être amenés à augmenter la part d'UC au sein de leur contrat épargne. Le ciblage de ces clients se fait via un score, et non directement via les prédictions d'un modèle de machine learning, pour des questions de compréhensions et d'intelligibilités pour ses utilisateurs.

3 La base de données

Après avoir présenté le contexte de cette étude ayant amené Crédit Agricole Assurances à mettre en place une politique de collecte afin d'augmenter le taux d'UC dans ses contrats d'épargne, nous allons détailler dans cette partie les caractéristiques de notre base de données représentant le portefeuille épargne de Crédit Agricole Assurances ainsi que notre critère d'appétence à la diversification.

3.1 Critère d'appétence à la diversification

La première étape de ce projet a nécessité de définir un critère binaire d'appétence à la diversification chez les clients épargne de Crédit Agricole Assurances. Ce critère va nous permettre de définir les clients de référence pour notre score.

Pour qu'un assuré soit considéré comme appétent à la diversification, ce dernier doit aller au-delà de la politique de collecte définie par Crédit Agricole Assurances, qu'il y soit soumis ou non. Cette dernière requiert un taux minimum de 30% d'UC pour les nouveaux entrants de la catégorie haut-de-gamme, c'est-à-dire, les personnes dépassant 40 000€ d'encours. On définit donc qu'une personne vérifie le critère d'appétence à la diversification si elle possède :

— Plus de 30% d'UC si l'encours est inférieur à 40 000€

— Plus de 40% d'UC si l'encours est supérieur à 40 000€

Selon notre critère, 18% des clients de la base de données utilisée pour construire le score sont considérés comme appétents, cela représente près de 40% des clients ayant des UC.

3.2 Période d'étude

La période d'étude de la base de données s'étale du 31/03/2019 au 31/03/2021. Les données présentes dans la base sont issues de la période allant du 31/03/2019 au 31/03/2020 et le critère d'appétence à la diversification est, quant à lui, calculé sur la période du 31/03/2020 au 31/03/2021.

FIGURE 5 – Période d'étude de la base de données

3.3 Description et Volumétrie

La base de données utilisée afin de créer ce score d'appétence à la diversification a été construite à la maille client. Elle contient 5 239 658 de lignes, 123 variables et regroupe des informations sur tous les clients épargne de Crédit Agricole Assurances présents durant toute la période d'étude. On trouve 6 types de variables différentes :

- Variables concernant les caractéristiques du client (Situation familiale, Sexe, ...)
- Variables résumant l'état de l'assurance épargne du client (Nb polices épargne, montant encours, type police, ...)
- Variables résumant l'état de l'assurance prévoyance du client (Nb polices prévoyance, montant encours, type police, ...)
- Variables comportementales (Nombre d'arbitrages et montant selon les différents types d'arbitrages réalisés sur la période 31/03/2019-31/03/2020)
- Variables signalétiques (Indicatrice de détention d'un produit de la gamme Pro/Agri, ...)
- Variable cible d'appétence au risque (variable binaire)

La construction de la base de données a nécessité l'agrégation et la jointure entre plusieurs tables. Celle-ci s'est déroulée en 6 étapes :

1. Récupération des informations sur les fonds UC : Durant cette étape, nous récupérons les informations sur les fonds UC, notamment l'historique de classification des fonds selon la nomenclature définie par l'AMF. Cette dernière est agrégée en 6 catégories (Fonds à Formule, Monétaire, Immobilier, Obligations, Diversifiés, Actions) à des fins de compréhension.
2. Jointure avec les données sur les polices épargne : Les données sur les polices épargne sont stockées à la maille police x fonds, on récupère ces informations auxquelles sont rajoutées celles obtenues à l'étape 1.

3. Création d'une table comportementale comprenant les actions effectuées sur les polices durant la période d'observation des données : les données sont construites par comptage des différentes types d'opérations observées (versement initial/exceptionnel, rachat partiel/total, transfert Fourgous, etc) ainsi que leur montant sur l'ensemble de la période d'observation à la maille police.
4. Jointure et agrégation des tables créées aux étapes 2 et 3 à la maille client : on procède tout d'abord à une partition des tables créées lors de l'étape 2, contenant les informations au niveau police, en fonction de la catégorie et du type de produit auquel appartient la police. Cette partition sert à effectuer un retraitement spécifique, selon le type de produit en épargne et en prévoyance, aboutissant la création de plusieurs variables. Par exemple, une police UC dont la PM UC est placée sur un fonds action verra la PM UC attribuée sur la variable comptabilisant les montants placés sur des actions. Les tables, constituées suite à la partition précédente, concernant les produits d'épargne sont ensuite regroupées et une jointure est effectuée avec les tables créées aux étapes 2 et 3. Une agrégation des données à la maille client est par la suite réalisée.
5. Ajout des données sur les clients et tests de qualité de données : On ajoute les données sur les caractéristiques clients à la la table créé lors de l'étape 4. Des tests de qualité des données sont effectués (ex : pas d'âge et d'ancienneté incohérent, nombre de lignes = nombre de clients uniques, etc)
6. Calcul du critère

Durant la construction de la base, différents filtres ont été appliqués :

— Exclusion des produits concernant des fonds dédiés et le bons de capitalisation : Les contrats concernant les fonds dédiés sont supprimés de par leur spécificité. S'agissant des bons de capitalisation, ces derniers peuvent être détenus anonymement, ce qui ne nous permet pas d'obtenir une vision complète sur ce type de produit et nous oblige donc à ne pas les prendre en compte dans notre base

— Exclusion des contrats ayant un encours nul ou manquant durant la période d'étude

— Exclusion des clients non présents durant toute la période d'étude

A des fins de contrôle sur nos données à l'aide d'une source de données tierce, nous avons effectué une comparaison entre la somme des encours de notre base de données au

31/12/2019 et la somme des encours recensés par la Direction du Contrôle de Gestion à cete même date. Nous obtenons un écart de -5,23% par rapport aux données du contrôle de gestion. Nous pouvons expliquer cet écart par la suppression de près de 5% des clients au moment de l'application des différents filtres listés ci-dessus.

3.4 Statistiques Descriptives

Sur les plusieurs millions de lignes qui composent la base de données construites afin de créer ce score d'appétence à la diversification, nous distinguons un certain nombre de caractéristiques sur nos clients qui sont décrites dans la partie ci-dessous.

En effet, parmi les 5 239 658 clients de notre base de données, on retrouve 51,7% de femmes pour un âge moyen de 55 ans et une ancienneté moyenne de 14 ans. 82% des clients de la base de données possèdent une seule police épargne et 13% en possèdent deux. Seuls 20% des clients de notre base détiennent également au moins une police prévoyance chez Crédit Agricole Assurances. L'encours moyen par client sur ses polices d'épargne est de 41 893€, toutefois, l'encours médian n'est que de 8 554€. Pour l'ensemble des clients de la base ayant moins de 100 000€ d'encours total, la répartition de ce dernier au 31/03/2020 est la suivante :

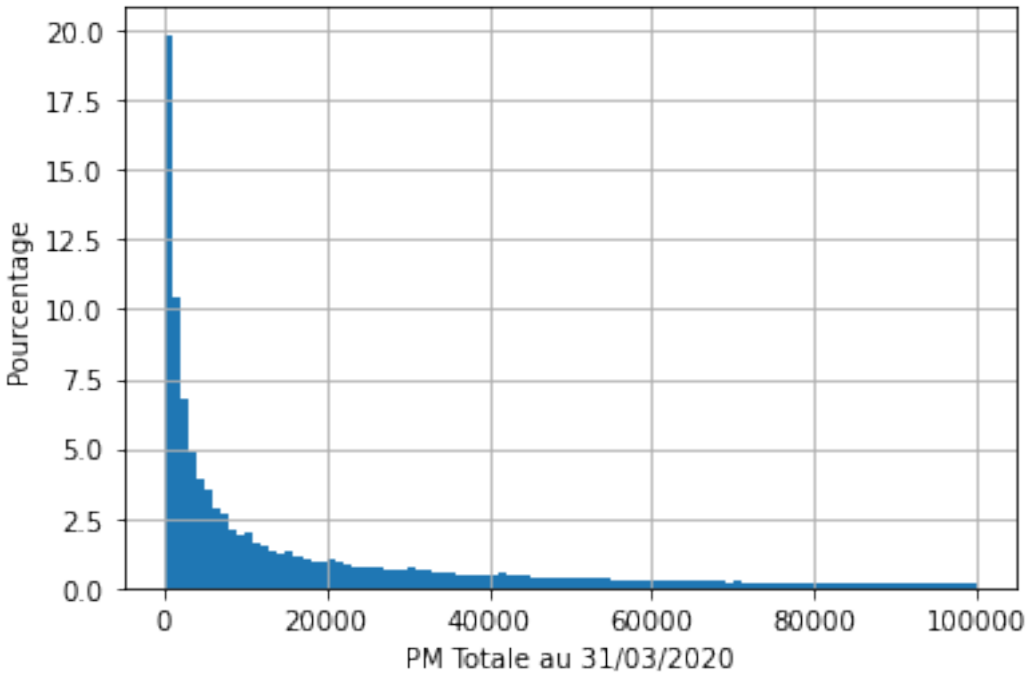

FIGURE 6 – Répartition de l'encours total pour les assurés ayant encours total inférieur à 100 000€

Celle-ci est décroissante avec le montant de l'encours total. Le premier quartile de

cette variable pour l'ensemble de la population est à 1 730€ et le troisième est à 37 545€. Seuls 10% des clients de la base ont plus de 100 000€ d'encours.

Le taux d'UC moyen sur notre base est de 15%. Cependant, seuls 45% des clients de notre base possèdent des UC. Chez ces clients, le taux d'UC moyen est de 33,5%. La répartition du taux d'UC selon que les clients aient plus ou moins de 40 000€ est la suivante :

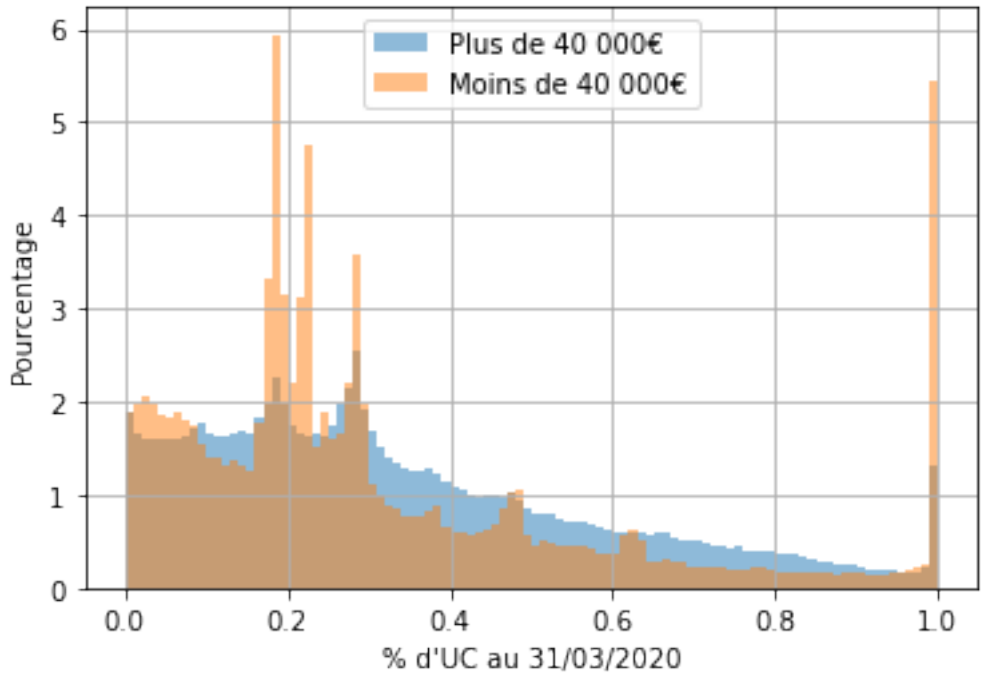

FIGURE 7 – Répartition du taux d'UC selon que le client ait plus ou moins 40 000€ d'encours pour les personnes ayant des UC

Pour les clients ayant moins de 40 000€ d'encours, on constate que la répartition est décroissante avec le pourcentage d'UC avec cependant 3 pics :

— un entre 17% et 23% d'UC qui est attribuable, en partie, au minimum des 20% UC nécessaires lors d'un transfert Fourgous sur une police d'épargne multi-support.

— un autour des 30% d'UC

— un dernier à 100% d'UC qui regroupe un peu plus de 5% des clients ayant moins de 40 000€ d'encours

Pour les personnes ayant plus de 40 000€ d'encours, approximativement les mêmes pics sont observables. Toutefois, ces derniers sont beaucoup moins importants. Le pic autour des 30% est imputable à la politique de collecte mise en place par Crédit Agricole Assurances.

Au sein de notre base, 24% de nos clients possèdent plus de 40 000€ d'encours. Parmi ces derniers, ils sont 62% à avoir des UC. Ce taux tombe à 40% parmi les personnes détenant moins de 40 000€ d'encours.

Les clients possédant de l'UC et ceux considérés comme appétents à la diversification selon notre critère possèdent en moyenne des caractéristiques proches de celles du client moyen, exception faite du niveau des encours et du taux d'UC. Les clients appétents selon notre critère possèdent un taux d'UC beaucoup plus élevé en moyenne que celui des clients ayant de l'UC, avec un taux d'UC moyen de 56,2%. Les clients appétents se distinguent aussi des autres clients de par leur revenu fiscal plus élevé, le revenu fiscal médian des clients appétents est de 33 100€ contre 25 000€ pour l'ensemble de la base et 28 800€ pour les clients ayant de l'UC.

Ces caractéristiques sont résumées dans le tableau suivant :

	Profil Moyen	Clients ayant de l'UC	Appétents
% de la base		45%	18%
% de femmes	51,7%	50,2%	51,8%
Âge moyen	55 ans	55 ans	57 ans
Ancienneté moyenne	14 ans	14 ans	14 ans
Nb Police Epargne moyen	1,24	1,39	1,32
Nb Police Prev moyen	0,29	0,32	0,31
Encours moyen	41 893€	61 364€	65 556€
Encours médian	8 554€	15 943 €	17 972€
% UC moyen	15%	33,5%	56,2%
Revenu fiscal médian	25 000€	28 800€	33 100€

Si l'on regarde, en détails, la répartition de l'âge des clients pour les clients appétents et non-appétents, celle-ci est la suivante :

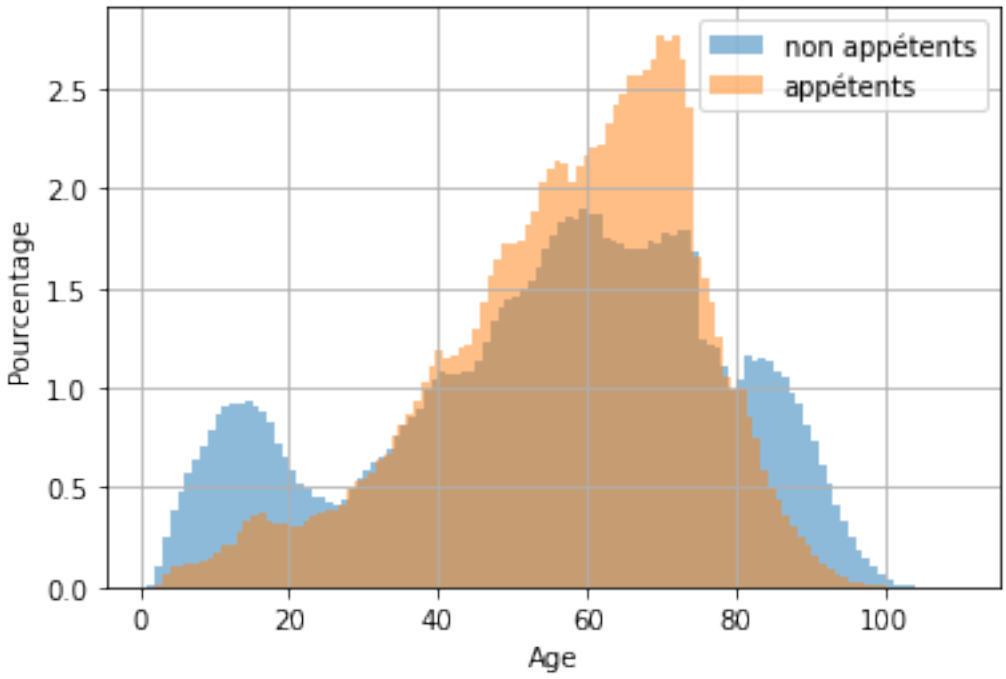

FIGURE 8 – Répartition de l'âge des assurés selon qu'ils soient appétents ou non

On constate un pic d'assurés appétents à la diversification entre 60 et 70 ans. L'âge moyen entre les clients appétents et l'ensemble de la population étant le même, ce pic est une conséquence de la faible présence d'appétents avant 18 ans, compte tenu des particularités des contrats pour les mineurs, et après 70 ans, du fait que les contrats d'épargne sont gérés dans une optique de succession.

De la même manière, la répartition de l'ancienneté des clients selon leur appétence est la suivante :

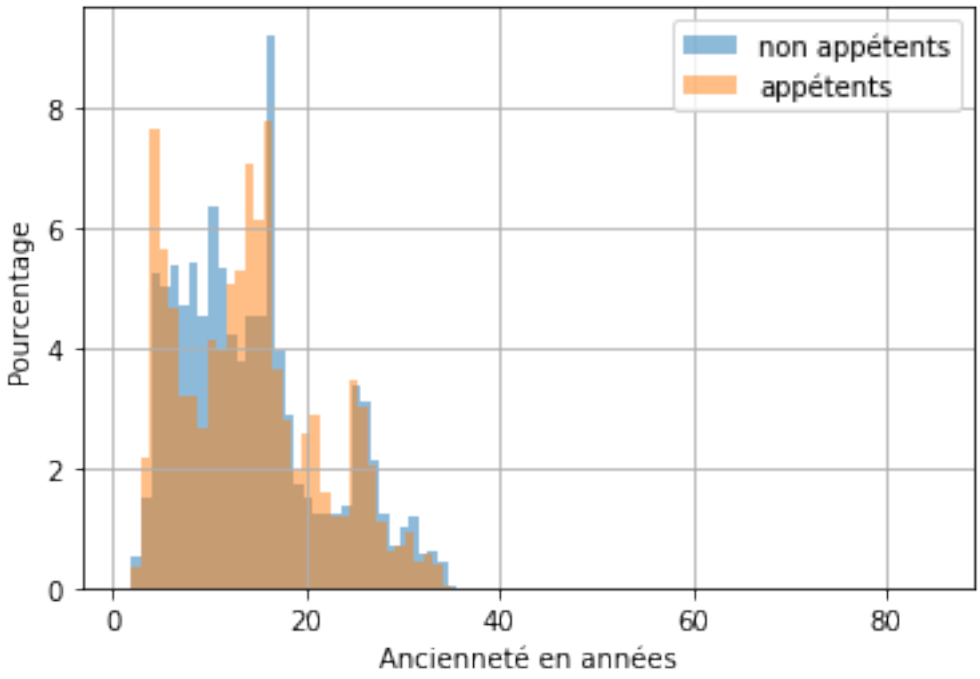

FIGURE 9 – Répartition de l'ancienneté des assurés selon qu'ils soient appétents ou non

On peut s'apercevoir que cette répartition est pratiquement identique entre les assurés appétents et non-appétents. Cependant, on constate quelques différences entre les 2 répartitions : on remarque un creux entre 7 et 12 ans d'ancienneté chez les appétents qui est immédiatement suivi d'un pic entre 12 et 15 ans d'ancienneté. De plus, on aperçoit un second pic entre 19 et 22 ans d'ancienneté. Le creux entre les 7 et 12 ans d'ancienneté peut être expliqué par l'approche de l'échéance fiscale sur les rachats. En effet, plus les clients ayant décidés d'effectuer un rachat se rapprochent de l'échéance fiscale, moins ils souhaiteront prendre des risques car, en cas de pertes, ils auront moins de temps pour les regagner. Une fois l'échéance fiscale passée, les assurés ayant eu cette stratégie quittent le portefeuille, ce qui explique une hausse de la part des appétents.

Concernant les professions de nos clients, celles-ci sont réparties en 8 catégories suivant la nomenclature des professions et catégories socioprofessionnelles de 2003

(PCS-2003) définie par l'INSEE [12]. La part de clients appétents à la diversification dans chaque catégorie est la suivante :

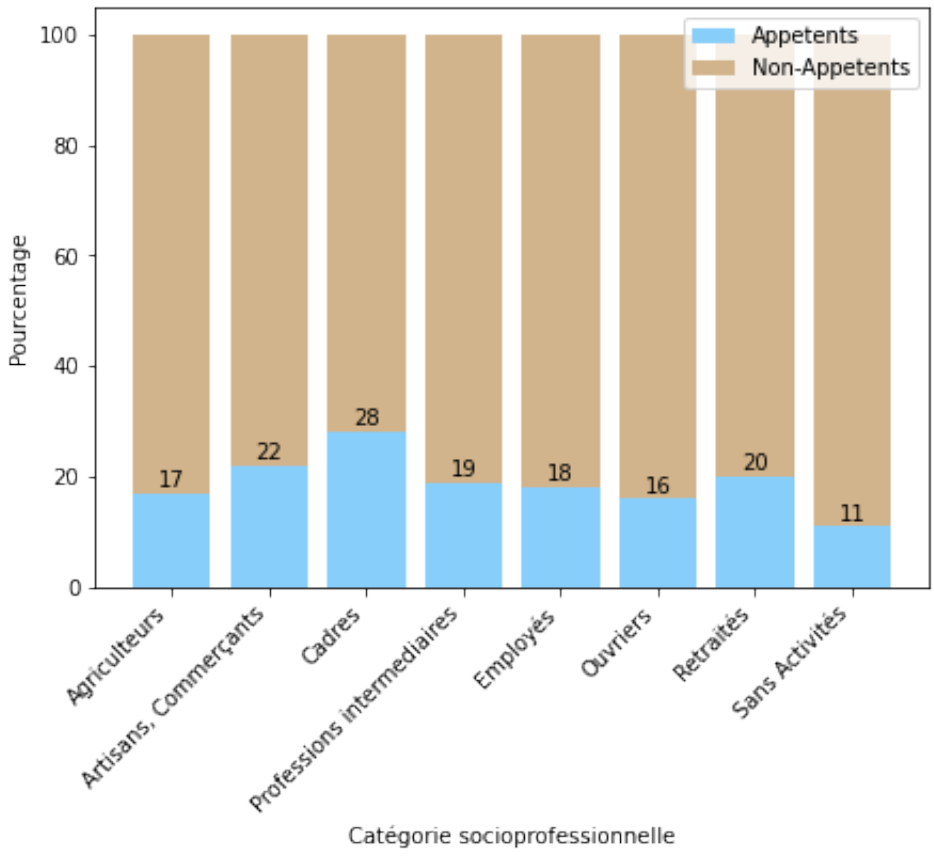

FIGURE 10 – Part d'appétents dans chaque catégorie socioprofessionnelle

La catégorie des cadres et professions intellectuelles supérieures est celle dont la part d'appétents en son sein est la plus élevée avec 28% des cadres qui sont appétents à la diversification selon notre critère. Vient ensuite la catégorie des artisans, commerçants et chefs d'entreprise avec 22% d'appétents.

Parmi les clients possédant de l'UC, les types de supports préférés sont les supports diversifiés suivis des supports action puis immobilier. La répartition moyenne sur les types de supports UC est à la suivante :

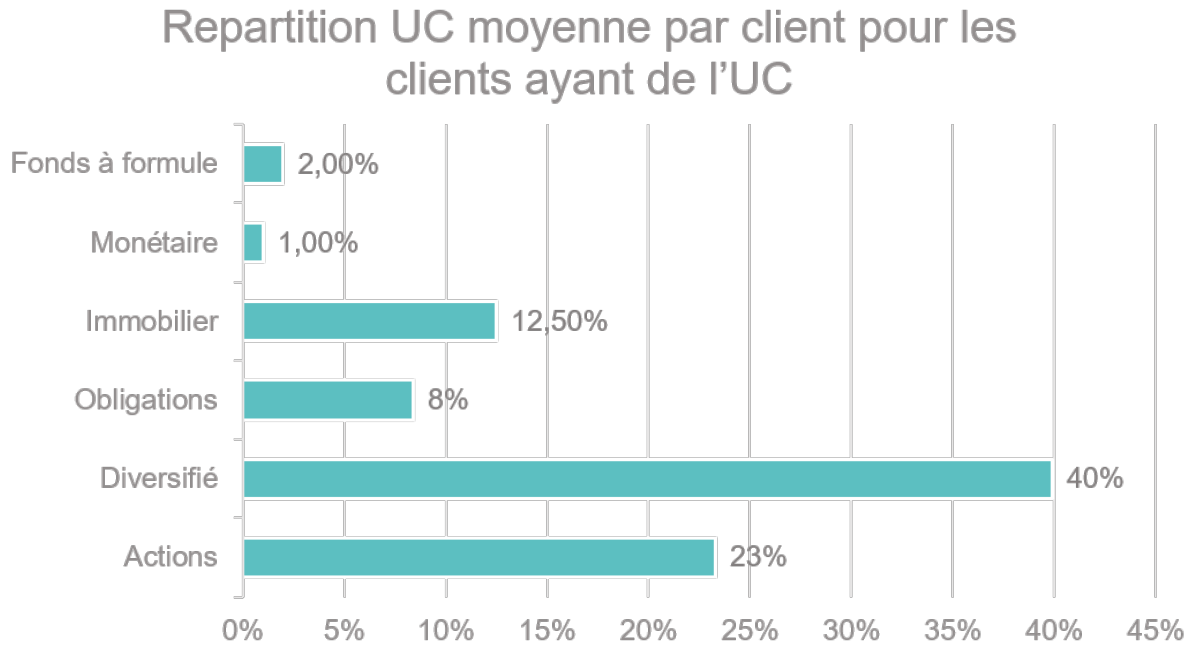

FIGURE 11 – Répartition moyenne selon les types de supports UC pour les clients ayant de l'UC

On observe quelques différences concernant la répartition moyenne selon les types de supports UC pour les clients appétents. En effet, le pourcentage moyen de l'encours UC mis sur des supports diversifiés n'est que de 33% pour les clients appétents contre 40% pour l'ensemble des clients ayant de l'UC. Les clients appétents ont aussi un taux moyen sur les supports obligations plus élevé que le client moyen ayant de l'UC de notre base, 12% et 8% respectivement.

Sur le graphique ci-dessous, on constate que pour les clients ayant moins 20% d'UC, les non-appétents vont en moyenne mettre plus d'actions parmi leurs UC que les clients appétents. Les supports actions ayant un taux de rendement plus élevé, en moyenne, que les autres types de supports en UC, on suppose que ces clients sont prêts à prendre plus de risques mais seulement pour une partie limitée de leur encours. On ne peut expliquer les pics constatés autour de 25% et 60% d'UC, où le taux moyen d'actions parmi les UC atteint 47% et 44% respectivement. La part d'actions parmi les UC est stable, aux alentours de 10%, chez les non-appétents ayant plus de 20% d'UC. Chez les appétents, après une décroissance de la part d'actions parmi les UC pour les appétents ayant moins

de 20% d'UC, on constate 2 niveaux sur la courbe : entre 25% et 50% d'UC, la part d'actions est de 14% en moyenne et pour plus de 50% d'UC, le taux d'actions est de 28% en moyenne.

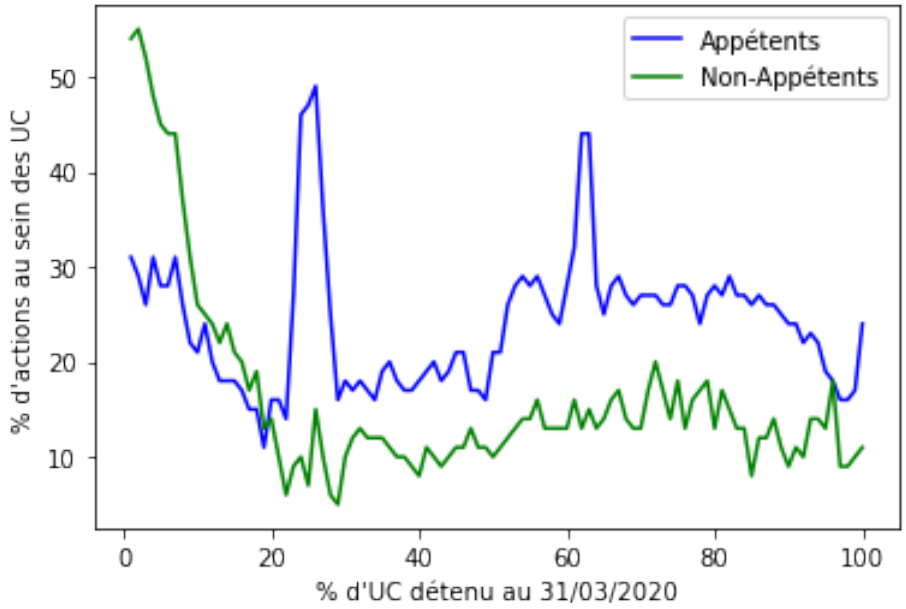

FIGURE 12 – Évolution du taux moyen d'actions parmi les UC selon le taux d'UC au 31/03/2020 selon l'appétence des clients

Parmi les clients ayant de l'UC, il existe une différence de comportement entre les appétents et les non-appétents sur le nombre de fonds UC utilisés. 70% des clients non-appétents n'utilisent qu'un seul fonds UC contre 43% des appétents. Ils sont que 15% des non appétents à en utiliser 2 contre 22% des appétents.

En comparant l'évolution du revenu fiscal médian en fonction de l'encours selon l'appétence des clients, on obtient le graphique ci-dessous :

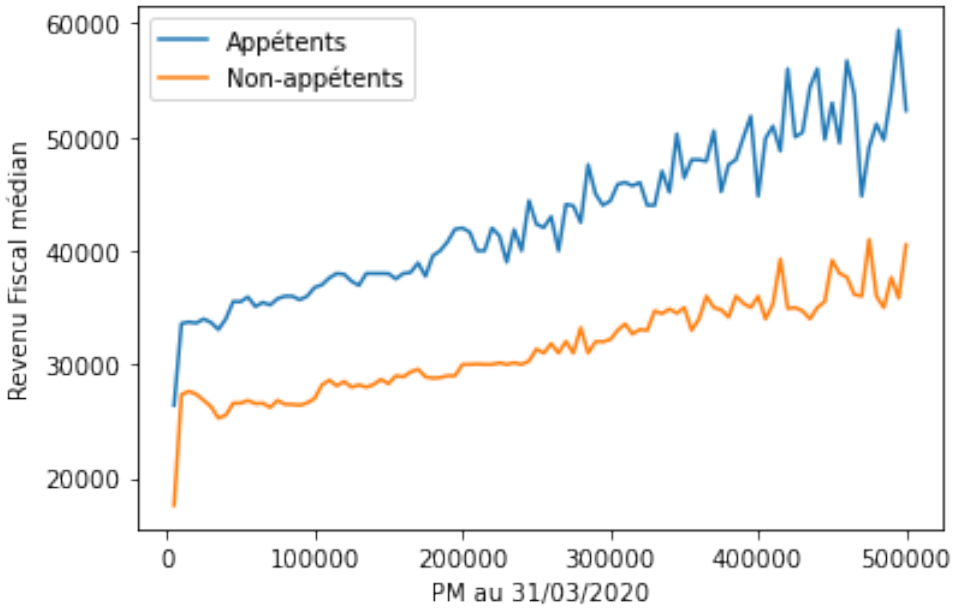

FIGURE 13 – Évolution du revenu fiscal médian en fonction de l'encours des clients selon l'appétence des clients

Le revenu fiscal médian est systématiquement plus élevé pour les appétents quel que soit le niveau de l'encours des clients. Cet écart est à peu près stable, variant entre 7 000€ et 9 000€ selon le niveau d'encours choisi. On peut constater que le taux de croissance du revenu fiscal médian en fonction de l'encours du client est assez faible. Ce taux de croissance est comparable pour les clients appétents et non-appétents.

Les clients appétents vont aussi se distinguer des clients non-appétents sur le montant des versement programmés qu'ils effectuent. Même si le pourcentage de la population appétente et non-appétente effectuant des versements programmés est le même, 43%, les appétents vont verser en moyenne 1 111€ contre 794€, en moyenne pour les non-appétents.

Ainsi les clients appétents de notre base se démarquent des autres clients de par le niveau de leur encours, leur taux d'UC, leur revenu fiscal et la répartition de leurs UC sur les différents supports. Ces caractéristiques que nous avons pu identifier avec les statistiques descriptives sont de possibles points que les modèles pourront utiliser pour prédire l'appétence de nos clients.

4 Modélisation

Afin d'aboutir à un score d'appétence à la diversification pour les clients épargne, nous avons dû passer par les étapes suivantes :

— Rééquilibrage de la base de données afin que les modèles de prédiction puissent être plus performants

— Prédiction des appétents

— Interprétation des résultats de la prédiction

— Création du score à partir des résultats de l'étape précédente

Dans cette partie, nous allons détailler le fonctionnement théorique des algorithmes et des différentes méthodes que nous avons utilisés durant les étapes de ce projet.

4.1 Prétraitement

Tout d'abord, un retraitement des variables est effectué avant de procéder à l'apprentissage des modèles afin que la forme des variables ait le moins d'incidence possible sur la modélisation. Pour les variables qualitatives, nous procédons à un encodage des variables, les différentes valeurs que peuvent prendre les variables qualitatives sont remplacées par des chiffres. Par exemple, pour une variable sur le sexe des individus, la catégorie "masculin" sera remplacé par un "1" et la catégorie "féminin" par un "2". Cette transformation est effectuée afin de permettre à notre modèle de prédiction de pouvoir traiter ce type de variables.

Pour les variables quantitatives, un retraitement est aussi effectué. Les variables sont normalisées à l'aide de la médiane et de l'écart inter-quartile. Cette méthode est préférée à la normalisation classique (réalisée à l'aide de la moyenne et de l'écart-type) afin d'être robuste aux valeurs extrêmes. Cette transformation permet d'éviter aux modèles de considérer la différence d'échelle entre les 2 variables comme un classement entre celles-ci, ce qui nuirait aux prédictions réalisées par le modèle. Cela permet aussi à certains modèles de réaliser plus rapidement leur apprentissage.

Pour la partie modélisation, la base de données initiale que nous avons construite a été divisée en 3 bases :

— La base d'entraînement, contenant 80% des lignes de la base initiale, qui servira à l'entraînement du modèle

— La base de validation comportant 10% des lignes, qui sera utilisée pour l'hyper-paramétrage et l'interprétation des résultats du modèle

— La base de test, contenant les 10% restants, permettra d'évaluer les performances du modèle

Chacune de ces bases est composée de la même proportion de clients appétents, la même que pour l'ensemble de la base, soit 18%.

4.2 Rééquilibrage de la base de données

Notre base de données est déséquilibrée du fait qu'elle ne contienne que 18% de clients appétents selon notre critère. Cela représente trop peu de lignes dans notre base pour que les modèles puissent correctement reconnaître les appétents.

Afin de rééquilibrer notre base de données, nous avons le choix entre le sur et le sous-échantillonage. Le sous-échantillonage a le désavantage de rejeter des données potentiellement intéressantes. Alors que le principal inconvénient du sur-échantillonage est qu'en créant des données synthétiques potentiellement identiques aux données "réelles", le modèle est plus à risque d'effectuer du sur-apprentissage. Le sur-échantillonage va aussi augmenter la taille des bases de données et donc allonger les temps de calcul des modèles. Toutefois, l'autre alternative au sur ou sous-échantillonage serait d'utiliser des modèles avec une pénalisation de la classe majoritaire. Cette possibilité n'est pas disponible dans tous les algorithmes de machine learning. De plus, il est difficile d'évaluer les coûts d'une mauvaise classification dans ces algorithmes [9].

Nous avons donc recours à l'algorithme *Synthetic Minority Oversampling Technique* abrégé en SMOTE qui a l'avantage de mélanger les techniques de sur et sous-échantillonage en effectuant un sur-échantillonage de la classe minoritaire et un sous-échantillonage de la classe majoritaire. Il a été développé par Chawla, Bowyer, Hall et Kegelmeyer en 2002 [6].

Le sur-échantillonage de la classe minoritaire a lieu via la création de données "synthétiques" sur la ligne liant les k-plus proches voisins de la classe minoritaire. Le sur-échantillonage de la classe minoritaire dans l'algorithme SMOTE peut être résumé de la manière suivante d'un point de vue algorithmique :

Sur-échantillonage de la classe minoritaire avec SMOTE

1) Initialisation : T : nombre d'échantillons de la classe minoritaire, k : nombre de plus proches voisins, N : le pourcentage de SMOTE voulu

2) si N<100 :

Randomisation des T échantillons de la classe minoritaire

$T = (N/100) * T$ et $N = 100$

3) *numattrs*=Nombre d'attributs, *Sample*[ ][ ] : matrice contenant les échantillons originaux de la classe minoritaire, *newindex* : index comptant le nombre d'échantillons synthétiques crées (initialisé à 0), *Synthetic*[ ][ ] : matrice contenant les échantillons synthétiques

Pour i allant de 1 à T :

Calcul des k plus proches voisins pour i, sauvegarde des indices dans la matrice *nnarray*

Tant que $N \neq 0$:

Choix d'un nombre aléatoire entre 0 et k appelé *nn*

Pour $attr$ allant de 1 à *numattrs* :

Calcul de $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$

Calcul de *gap*= nombre aléatoire entre 0 et 1

$Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$

$newindex++$ et $N = N - 1$

Le sous-échantillonage de la classe majoritaire a ensuite lieu en retirant aléatoirement des lignes de la classe majoritaire.

4.3 Gradient Boosting

Une fois l'algorithme SMOTE appliqué sur les données de la base d'entraînement des modèles, celle-ci n'est plus déséquilibrée et il est possible de passer à la partie modélisation qui va nous servir à prédire l'appétence de nos assurés. L'algorithme de prédiction utilisé durant ce projet est l'*eXtrem Gradient Boosting*, abrégé en XGBoost, appartenant à la famille des algorithmes de Gradient Boosting. Il est l'un des algorithmes le plus populaire au sein des compétitions de Data Science, de par ses performances. L'*eXtrem Gradient Boosting* a été développé par Chen et Guestrin dans leur article *XGBoost : A scalable Tree Boosting System* datant de Juin 2016 [5].

A l'inverse des forêts aléatoires, qui vont produire plusieurs arbres de décision, de manière indépendante, et les agréger afin, que dans le cas d'une classification, la catégorie la plus fréquente donnée par les arbres soit choisie, le boosting va combiner les uns après les autres les classifieurs faibles. Le boosting permet à chaque nouveau classifieur d'apprendre des erreurs commises auparavant.

Nous cherchons ici à expliquer la variable cible y à l'aide des variables $x = x1, x2, x3, ...$ tel que $f(x) = y$. Afin d'estimer cette fonction f, nous allons chercher à minimiser, à l'itération t, la fonction de coût suivante :

$$L^t = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

où y_i est la vraie valeur de y pour la i-ème ligne, $\hat{y_i}^{(t)}$ la prédiction pour la i-ème ligne à la t-ème itération, l une fonction de coût définie dans les paramètres du XGBoost, f_t la structure de l'arbre lors de la t-ème itération et Ω pénalise la complexité du modèle tel que $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ avec T, le nombre de feuilles de l'arbre, w, le poids des feuilles, λ, le terme de régularisation servant à réduire la sensibilité de la prédiction aux observations individuelles et γ représente la pénalité définie par l'utilisateur.

L'optimisation de cette fonction nécessite l'utilisation d'une approximation du second degré :

$$L^t \simeq \sum_{i=1}^{n} [l(y_i, \hat{y_i}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

où $g_i = \frac{\partial l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}}$ et $h_i = \frac{\partial^2 l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)^2}}$

En retirant la constante, nous obtenons la fonction objectif suivante à minimiser :

$$\tilde{L}^t :\simeq \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

En définissant $I_j = i|q(x_i) = j$ l'ensemble des cas de la feuille j, on trouve :

$$\tilde{L}^t \simeq \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

$$\simeq \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T$$

Pour une structure $q(x)$ fixe, nous obtenons comme valeur optimale des poids des feuilles :

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

Ainsi, la fonction objectif à minimiser devient :

$$\tilde{L}^t(q) \simeq -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T$$

Cette fonction peut être utilisée comme une fonction de scoring afin de mesurer la qualité de l'arbre q.

Cependant, il est impossible de minimiser cette fonction objectif pour tous les arbres existants q. En définissant I_L et I_R comme les ensembles des cas à gauche et à droite du noeud tel que $I = I_L \cup I_R$, la fonction objectif à minimiser à la division de chaque branche devient :

$$L_{split} = \frac{1}{2}[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}] - \gamma$$

Dès lors que cette fonction objectif devient négative et/ou inférieure au paramètre de gain minimum pour une séparation (*min_split_gain*) qui est renseigné par l'utilisateur, la division du noeud s'arrête.

Le XGBoost peut être résumé de façon algorithmique de la manière suivante :

Algorithme XGBoost

---

1) Initialisation : I : noeud actuel, d : dimension des données
$gain = 0$, $G = \sum_{i \in I} g_i$, $H = \sum_{i \in I} h_i$

2) Pour k allant de 1 à m :

G_L=0, H_L=0

pour j dans I (I classé par x_{jk}) :

$G_L = G_L + g_j$, $H_L = H_L + h_j$

$G_R = G - G_L$, $H_R = H - H_L$

$score = max(score, \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda})$

Sortie : Division du noeud avec le score maximum

Dans le cas de la classification binaire, la fonction de coût l utilisée est une fonction de perte logarithmique telle que :

$$l(y,p) = yln(p) + (1-y)ln(1-p)$$

avec $p = \frac{1}{(1+e^{-x})}$, la probabilité calculée en appliquant la fonction logistique à la sortie, x, de l'arbre boosté. On obtient donc :

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

$$h = p(p-1) = \frac{-e^{-x}}{(1+e^{-x})^2}$$

Le XGBoost est un algorithme dit *exact greedy* du fait que pour chaque noeud, il va lister l'ensemble des possibilités de division du noeud pour toutes les variables. Les algorithmes *exact greedy* demandent d'importantes ressources de calculs ce qui

oblige à avoir recours à des approximations. Le XGBoost va proposer des divisions de noeuds basées sur les centiles de chaque variable. L'algorithme laisse le choix à l'utilisateur entre 2 variantes, une méthode globale où les centiles sont calculés lors de la construction du premier arbre et la méthode locale qui va recalculer les centiles à chaque nouvelle étape.

L'algorithme XGBoost a l'avantage d'avoir un temps de calcul beaucoup plus faible que celui d'autres algorithmes de boosting du fait qu'il soit *sparcity aware*. Le XGBoost a un traitement particulier des valeurs manquantes dans les matrices creuses, les valeurs manquantes sont classées dans la direction par défaut au moment du calcul du score pour la division des noeuds. L'algorithme XGBoost recourt à de la parallélisation afin de réduire le temps de calcul.

Nous avons cherché à optimiser les paramètres suivants du XGBoost :

— *Taux d'apprentissage* : Coefficient appliqué aux nouveaux poids calculés à chaque étape afin de réduire le sur-apprentissage

— *Profondeur maximale* : La profondeur maximale de l'arbre, plus cette valeur est élevée plus l'arbre est complexe mais ceci peut mener à du sur-apprentissage

— *Sous-échantillon* : Ratio des données d'entraînement utilisées pour la création de l'arbre à chaque étape

— *Nombre d'estimateurs* : Nombre d'arbres à entraîner par le modèle

— *Lambda* : Terme de régularisation $L2$ servant à éviter le sur-apprentissage (λ dans les équations ci-dessus)

— *Alpha* : Terme de régularisation $L1$ servant à éviter le sur-apprentissage

4.4 Métriques d'évaluation

Une fois l'apprentissage des modèles effectué à l'aide de la base d'entraînement, il est nécessaire d'évaluer ses performances sur de nouvelles données. Pour estimer les capacités prédictives de nos modèles, nous nous baserons principalement sur des métriques d'aire sous la courbe (*Area Under the Curve* en anglais, abrégé en AUC).

Celles-ci sont particulièrement adaptées pour les problèmes de classification binaire car elles permettent d'évaluer la capacité d'un modèle à bien classer les points des différentes catégories. L'AUC va de 0 à 1 avec 1, la meilleure note possible représentant le fait que le modèle a parfaitement identifié tous les membres de chaque classe. Les calculs des différents AUC sont basés sur la matrice de confusion. Cette matrice permet de mesurer et de visualiser la qualité d'un système de classification. Les lignes indiquent les classes réelles et les colonnes les classes estimées par le modèle et prend la forme suivante :

Valeurs prédites		
	$\hat{y} = 0$	$\hat{y} = 1$
y=0	Vrais Négatifs (VN)	Faux Négatifs (FN)
y=1	Faux Positifs (FP)	Vrais Positifs (VP)

L'*Area Under the Curve Precision-Recall* soit l'aire sous la courbe Précision-Rappel en français, abrégée en AUC-PR, est particulièrement adaptée pour les cas de classification binaire à partir de base déséquilibrée où l'on s'intéresse surtout à la classe minoritaire [16].

On définit tout d'abord la précision et le rappel comme :

— $Précision = P(Y = 1/\hat{Y} = 1) = \frac{Nombre\ d'individus\ correctement\ attribues\ a\ la\ classe\ 1}{Nombre\ d'individus\ attribues\ a\ la\ classe\ 1} = \frac{VP}{FN+VP}$

— $Rappel = P(\hat{Y} = 1/Y = 1) = \frac{Nombre\ d'individus\ correctement\ attribues\ a\ la\ classe\ 1}{Nombre\ d'individus\ appartenant\ a\ la\ classe\ 1} = \frac{VP}{FP+VP}$

La précision peut être comprise comme une mesure de la qualité du modèle ou de son exactitude et le rappel comme une mesure de la quantité ou de l'exhaustivité du modèle.

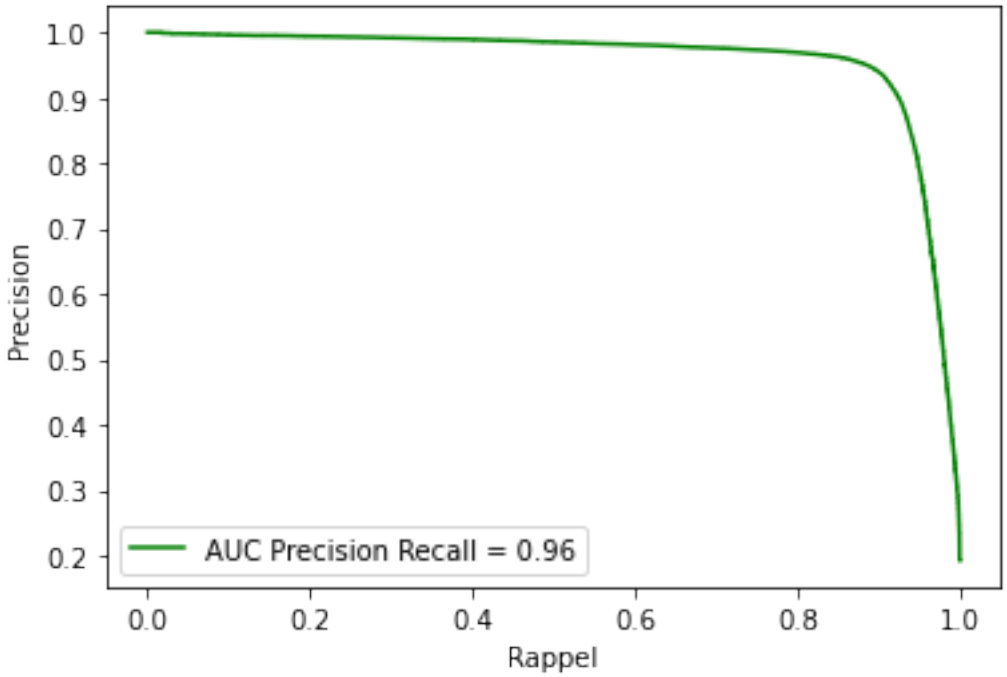

FIGURE 14 – Exemple de courbe Précision-Rappel

La courbre Precision-Rappel est construite en calculant pour chaque seuil la précision et le rappel. Plus la précision est haute, plus le rappel est bas. Cette courbe est une aide dans la résolution du dilemme précision-rappel où l'on doit choisir le seuil permettant d'avoir le meilleur modèle. L'AUC-PR sera notre métrique d'évaluation lors du calcul des hyperparamètres de nos modèles.

L'*Area Under the Curve Recever Operating Characteristic* soit l'aire sous la courbe de la fonction d'efficacité du récepteur en français, abrégée en AUC ROC est une autre métrique d'évaluation pour les classifieurs binaires. On définit, tout d'abord, le taux de faux positifs et le taux de vrais positifs comme :

— *Taux de faux positifs*= $\frac{FP}{FP+VN}$

— *Taux de vrais positifs*= $Rappel = \frac{VP}{FP+VP}$

La courbe ROC est construite en prenant pour chaque seuil le taux de faux positifs en abscisse et le taux de vrais positifs en ordonnées. Un AUC ROC de 0,5 correspond au niveau de performance d'un tirage aléatoire et il est commun de représenter la mé-

diatrice avec la courbe ROC afin de pouvoir apprécier visuellement les performances du modèle.

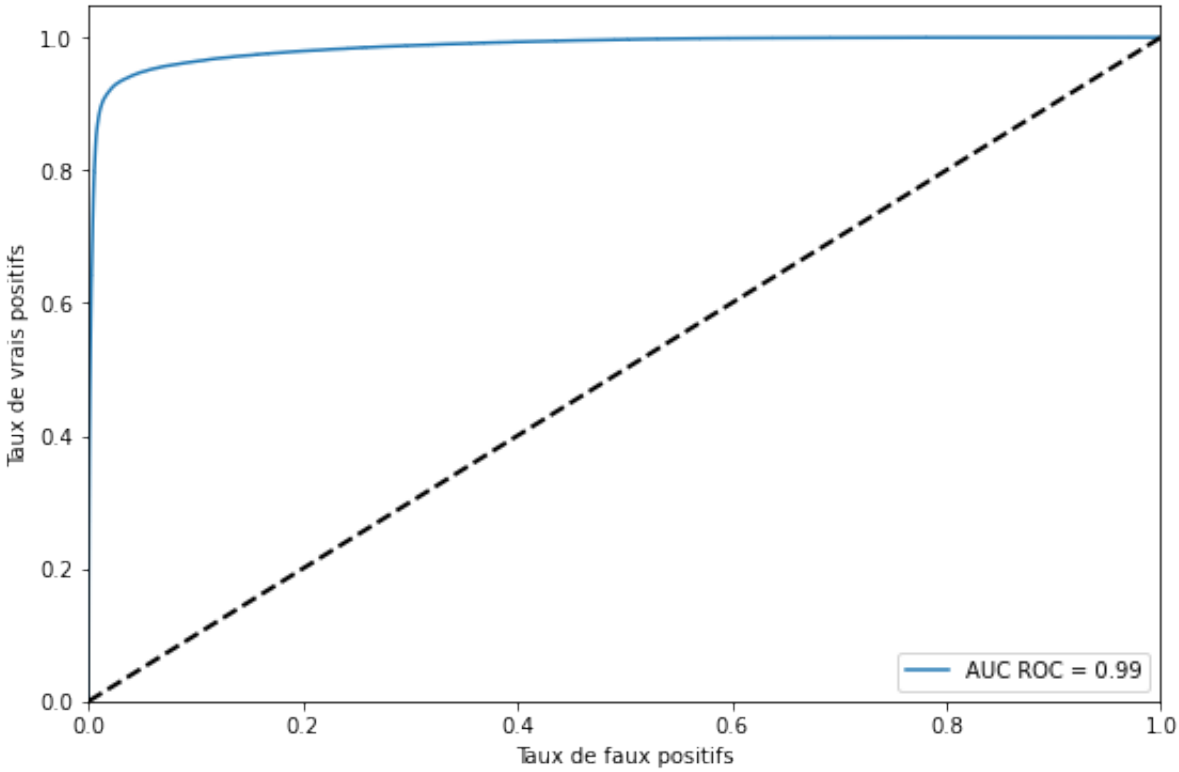

FIGURE 15 – Exemple de courbe ROC

Afin de définir le rappel et la précision optimale pour notre modèle, nous nous appuierons sur le *F1-Score* qui est défini de la façon suivante :

$$F1 - Score = 2 * \frac{precision * rappel}{precision + rappel}$$

4.5 Hyperparamétrage

Afin d'obtenir les meilleurs prédictions possibles sur l'appétence future de nos clients, nous devons choisir le paramétrage des algorithmes SMOTE et XGBoost. Le choix de ces paramètres a été effectué via hyperparamètrage qui va nous permettre d'utiliser les paramètres optimaux pour nos algorithmes. L'hypermaramétrage est réalisé à l'aide de l'optimisation bayésienne.

L'optimisation bayésienne construit un modèle probabiliste de la fonction qui va chercher la valeur des meilleurs hyperparamètres. Cette dernière est déterminée selon une fonction objectif calculée sur la base de validation que l'on cherche à optimiser. Contrairement à d'autres méthodes d'hyperparamétrage (*Grid search*, *Random Search* pour les plus connues), l'optimisation bayésienne est une méthode itérative, elle utilise les résultats de ses tests précédents afin de déterminer quels paramètres tester par la suite.

Durant l'hyperparamétrage, nous avons choisi comme fonction objectif à optimiser, l'AUC-PR pénalisé sur la base de validation, définie de la manière suivante :

$$-(AUC_{val} - ((AUC_{train} - AUC_{val}) * \frac{1}{100}))$$

Comme évoqué dans la partie précédente sur les métriques d'évaluation, nous choisissons l'AUC-PR pénalisé comme fonction objectif car l'AUC-PR est plus adaptée pour les cas de classification binaire avec une base déséquilibrée et nous pénalisons cette métrique afin d'éviter un sur-apprentissage du modèle.

L'optimisation bayésienne nécessite de définir un espace de recherche dans lequel la valeur des hyperparamètres va être recherchée. Toutefois, il est trop coûteux (en temps et en ressources) d'évaluer la fonction objectif en l'ensemble des points de l'espace de recherche. L'optimisation bayésienne va donc construire un modèle de substitution afin d'approximer notre vraie fonction objectif à l'aide d'une fonction de substitution. Le modèle de substitution que nous avons utilisé est le modèle des processus gaussiens. Il va chercher à calculer la probabilité de la valeur de la fonction objectif selon la valeur des hyperparamètres soit mathématiquement :

$$\mathbb{P}(valeur\ de\ la\ fonction\ objectif\ \setminus\ valeur\ des\ hyperparametres)$$

L'étape suivante de l'optimisation bayésienne consiste à évaluer la fonction objectif en 10 points choisis aléatoirement au sein de l'espace de recherche des hyperparamètres.

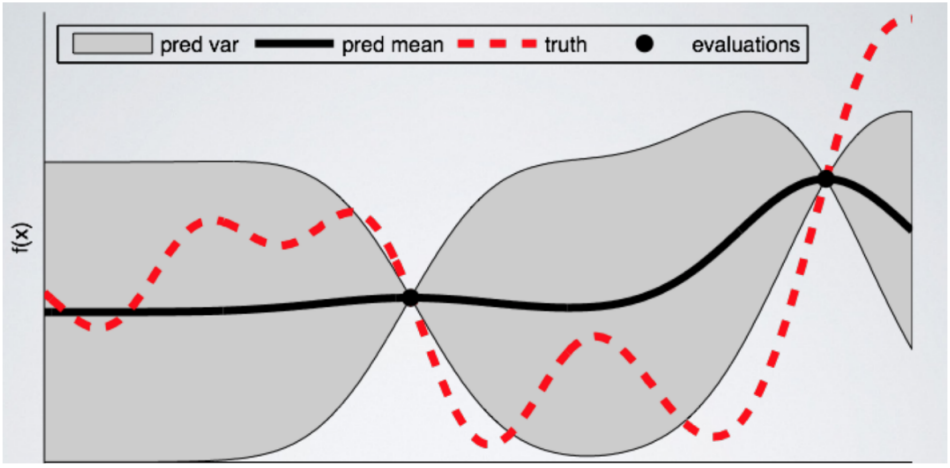

FIGURE 16 – Exemple avec 2 points choisis aléatoirement pour évaluer la fonction objectif [14]

Afin de déterminer quels paramètres tester pour les points suivants, l'optimisation bayésienne s'appuie sur une fonction d'acquisition. Comme fonction d'acquisition, nous utilisons la fonction d'amélioration attendue définie comme :

$$EI_x = \mathbb{E}[\ max(f(x) - f(x_+), 0)\ ]$$

avec f : la fonction objectif, x_+ : les hyperparamètres trouvés maximisant actuellement la fonction objectif et x : le nouveau point testé

Le point suivant pour lequel la fonction objectif sera évaluée est le point au sein du modèle de substitution qui maximise la fonction d'acquisition. Une fois la fonction objectif calculée en ce nouveau point, le modèle de substitution est mis à jour. Puis le nouveau point sur lequel estimer la fonction objectif est choisi, etc, jusqu'à ce qu'on atteigne le nombre d'itération maximum défini par l'utilisateur.

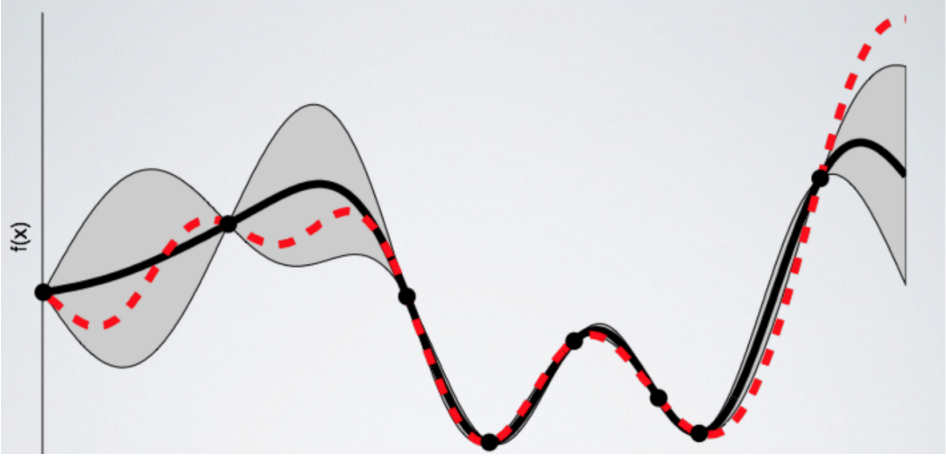

FIGURE 17 – Exemple avec la fonction objectif évaluée en 6 nouveaux points choisis selon la fonction d'acquisitionv[14]

De manière pseudo-algorithmique, cela donne :

Hyperparamétrage via l'optimisation Bayésienne [4]

---

1) Initialisation : H : historique d'observation de la paire (hyperparamètre, valeur de la fonction objectif, T : nombre maximum d'itération, f : fonction objectif,
M : fonction de substitution, S : fonction d'acquisition, $x*$: point suivant à évaluer

2) Pour t allant de 1 à T :

$x^* = argmax_x\ S(x, M_{t-1})$

calcul de $f(x^*)$

$H \longleftarrow \mathrm{H} \cup (\mathrm{x}^*, f(x^*))$

Calcul d'un nouveau modèle M_t selon H

Sortie : H

Une fois cela effectué, il ne reste plus qu'à choisir parmi les points évalués (H dans l'algorithme ci-dessus) celui qui maximise la fonction objectif.

4.6 Valeurs de Shapley

Outre les performances du modèle optimisé sur la base de validation, il nous est nécessaire de pouvoir comprendre et analyser les résultats et les prédictions faites afin de pouvoir construire un score. Le XGBoost est un algorithme dit "boite-noire", le fonctionnement interne du modèle est masqué. Afin de pouvoir interpréter les résultats des algorithmes, nous avons recours à une méthode d'intelligibilité locale des modèles, les valeurs de Shapley.

A la différence des méthodes d'intelligibilité globales qui vont donner les variables les plus importantes pour le modèle, les méthodes d'intelligibilité locales sont capables de donner pour chaque prédiction les variables les plus importantes, ce qui nous sera utile dans l'optique de créer notre score.

Les valeurs de Shapley sont issues de la théorie des jeux [20], elles permettent, dans le cas d'un jeux coopératif, d'estimer une répartition équitable des gains des joueurs selon la contribution de chaque joueur au groupe. Cette méthode a été adaptée aux modèles de machine learning par Scott M. Lundgerg et Su-In Lee [18].Les valeurs de Shapley font partie des méthodes additives d'attribution des caractéristiques de telle sorte que la fonction permettant d'expliquer le modèle soit une fonction linéaire des données du modèle :

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

Où g est l'explication du modèle, x' une version simplifiée de x, les données utilisées pour la prédiction du modèle, M le nombre de variables et $\phi_i \in \mathbb{R}$.

D'après le théorème de Shapley, une seule fonction appartenant à la famille des méthodes additives d'attribution des caractéristique vérifie les 3 propriétés suivantes :

— **Homogénéité :** Si le modèle change de manière à ce que la contribution de certaines données simplifiées augmente ou reste au même niveau, indépendemment des autres données, alors l'importance de la variable ne doit pas décroître. Ceci équivaut à ce que pour 2 modèles f et f', si : $f'(x') - f'(x'\backslash i) \geq f(x') - f(x'\backslash i)$, avec $x\backslash i$ désignant les cas $x'_i = 0$, alors $\phi_i(f', x) \geq \phi_i(f, x)$

— **Efficience :** Lors de l'approximation du modèle d'origine f pour des données x, la propriété d'efficience nécessite que le modèle d'explication g corresponde

au moins à la sortie de f pour les données simplifiées x'. Soit $f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i$

— **Joueur nul :** Des données manquantes ou nulle n'ont aucun effet sur l'explication du modèle soit si $x'_i = 0$ alors $\phi_i = 0$

La seule fonction vérifiant ces 3 propriétés est la suivante :

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f(z') - f(z' \setminus i)]$$

avec z' un sous-ensemble de x', $|z'|$ est le nombre d'entrées différentes de 0 dans z'.

Cette fonction va donc nous permettre de calculer les valeurs de Shapley de chaque variable pour toutes nos prédictions afin de pouvoir expliquer notre modèle. Si l'on souhaite représenter les valeurs de Shapley prises pour chaque valeur possible d'une variable, nous obtenons la courbe lissée suivante :

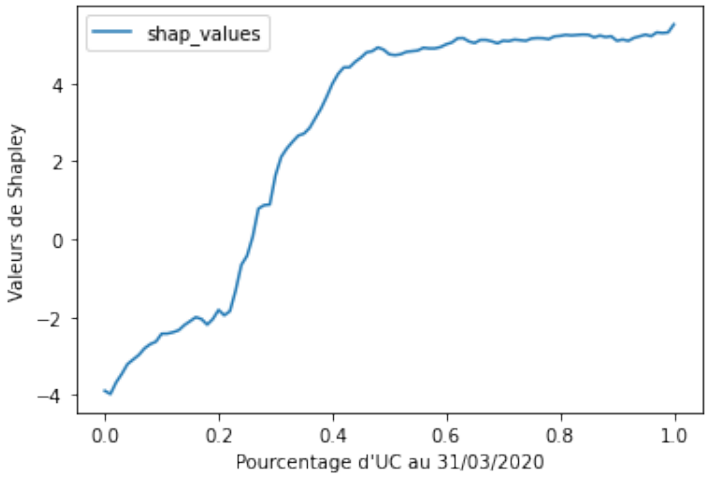

FIGURE 18 – Évolution des valeurs de Shapley moyennes pour la variable représentant le pourcentage de l'encours sur des fonds UC au 31/03/2020

4.7 BIRCH Clustering

L'interprétation du modèle utilisé s'est effectuée à l'aide des valeurs de Shapley, présentées dans la section précédente. Afin de passer des valeurs de Shapley, représentées sur le graphique ci-dessus, à un score, cela nécessite une étape de discrétisation des variables continues utilisées. Le choix a été fait de les discrétiser après la phase d'apprentissage du modèle afin d'obtenir les catégories les plus pertinentes possibles selon l'interprétation du modèle.

La discrétisation des variables continues a été effectuée grâce à 2 méthodes :

— les observations faites sur la courbe des valeurs de Shapley

— un algorithme de clustering

L'algorithme de clustering que nous avons utilisé est l'algorithme *Balanced Iterative Reducing and Clustering Using Hierarchies*, dit "BIRCH", développé en 1996 par Zhang, Ramakrishnan et Livny dans leur papier *"BIRCH : An Efficient Data Clustering Method for Very Large Databases"*[21]. Cet algorithme a été choisi car il a l'avantage :

— d'être un algorithme dit local du fait que les décisions de regroupement sont prises sans analyser tous les points ou tous les clusters existants.

— de ne pas considérer l'espace où sont représentées les données comme uniforme et donc d'attribuer une importance différente aux points selon qu'ils soient isolés ou dans une région dense permettant ainsi une gestion des valeurs aberrantes

— d'être optimisé pour les données de taille importante permettant d'avoir un temps d'exécution raisonnable

— de ne pas nécessiter dans ses paramètres la saisie du nombre de clusters voulus par l'utilisateur

L'algorithme peut se décomposer en 2 grandes étapes :

1. Construction d'un *Clustering Feature Tree* à l'aide d'un clustering local
2. Clustering global en appliquant un algorithme de clustering sur les feuilles du *Clustering Feature Tree*

Afin d'être optimisé pour le traitement de grands ensembles de données, BIRCH utilise les vecteurs *Cluster Features (CF)* afin de représenter un ensemble de données à l'aide de quelques statistiques. Elles sont définies de la manière suivante $CF = (N, \vec{LS}, SS)$ avec :

— N le nombre de points dans le cluster

— $\vec{LS} = \sum_{i=1}^{N} X_i$, la somme des points

— $SS = \sum_{i=1}^{N} X_i^2$

Par exemple, les CF permettent de calculer pour chaque cluster X :

— Le centroïde : $\vec{X_0} = \frac{1}{N} \sum_{i=1}^{N} X_i$

— Le rayon : $R = (\frac{1}{N} \sum_{i=1}^{N} \|X_i - X_0\|^2)^{\frac{1}{2}}$

— Le diamètre : $D = (\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \|X_i - X_j\|^2)^{\frac{1}{2}}$

Et pour 2 clusters, X et Y, il est possible de calculer à partir des CF :

— La distance euclidienne des centroïdes : $D_0 = \|X_0 - Y_0\|$

— La distance de Manhattan des centroïdes : $D_1 = |X_0 - Y_0|$

— La distance inter-cluster moyenne : $D_2 = (\frac{1}{N_X N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \|X_i - Y_j\|^2)^{\frac{1}{2}}$

— La distance inter et intra-cluster moyenne

La première étape de l'algorithme BIRCH consiste en la construction d'un arbre CF (*Clustering Feature Tree* en anglais). Cette arbre CF, comme les arbres de décision, se compose à sa base d'une racine. Cette dernière se scinde ensuite en une ou plusieurs partie, appelées branches. Au bout de chaque branche se trouvent des noeuds. Chaque noeud peut se diviser à nouveau en plusieurs branches donnant sur des noeuds. Si le noeud suivant n'est pas divisé en branches alors on aboutit aux feuilles. On peut donc distinguer 2 types de noeuds, ceux aboutissant sur des feuilles ou non. On appelle la profondeur de l'arbre, le nombre de noeud que possède un arbre entre sa racine et ses feuilles. Dans le cadre d'un arbre CF de l'algorithme de clustering BIRCH, chaque noeud est composé de clusters qui sont assez proches les uns des autres pour être regroupés ensemble.

La racine et chaque noeud non-composé de feuille sont identifiés par les couples $[CF_i, child_i]$: les *Cluster Features* des clusters composant le noeud et les branches partant de ce noeud, appelées *child*. Les noeuds non-composés de feuille contiennent au maximum B couples de la forme $[CF_i, child_i]$. B est appelé le facteur de ramification (*branching factor* en anglais), c'est l'un des paramètres de l'algortihme BIRCH. Les clusters de chaque noeud sont composés comme la somme des sous-clusters présents dans ses branches.

Les noeuds aboutissant sur des feuilles sont identifiés par les CF_i, les *Cluster Features* des clusters composant le noeud et par une flèche reliant les feuilles précédentes et suivantes. Ces noeuds ne doivent pas dépasser L feuilles et les feuilles doivent avoir un diamètre inférieur à T.

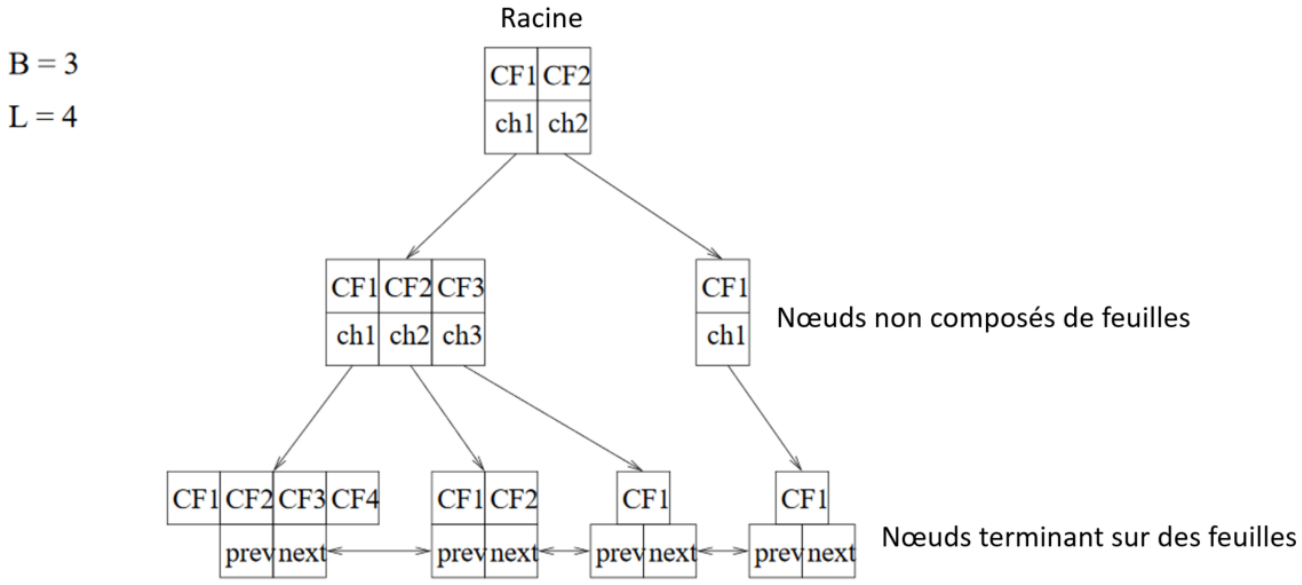

FIGURE 19 – Exemple d'arbre *CF* de l'algorithme BIRCH [17]

L'arbre *CF* est construit de manière dynamique en étant modifié à l'ajout de chaque nouvelle donnée. Cette opération s'effectue de la manière suivante :

1. Identification la feuille adéquate : En partant de la racine, on descend récursivement l'arbre *CF* en choisissant la branche (ou *child* avec la notation du paragraphe précédent) la plus proche selon la métrique choisie (distance Euclidienne/Manhattan des centroïdes, distance inter/intra-cluster moyenne ou autres)

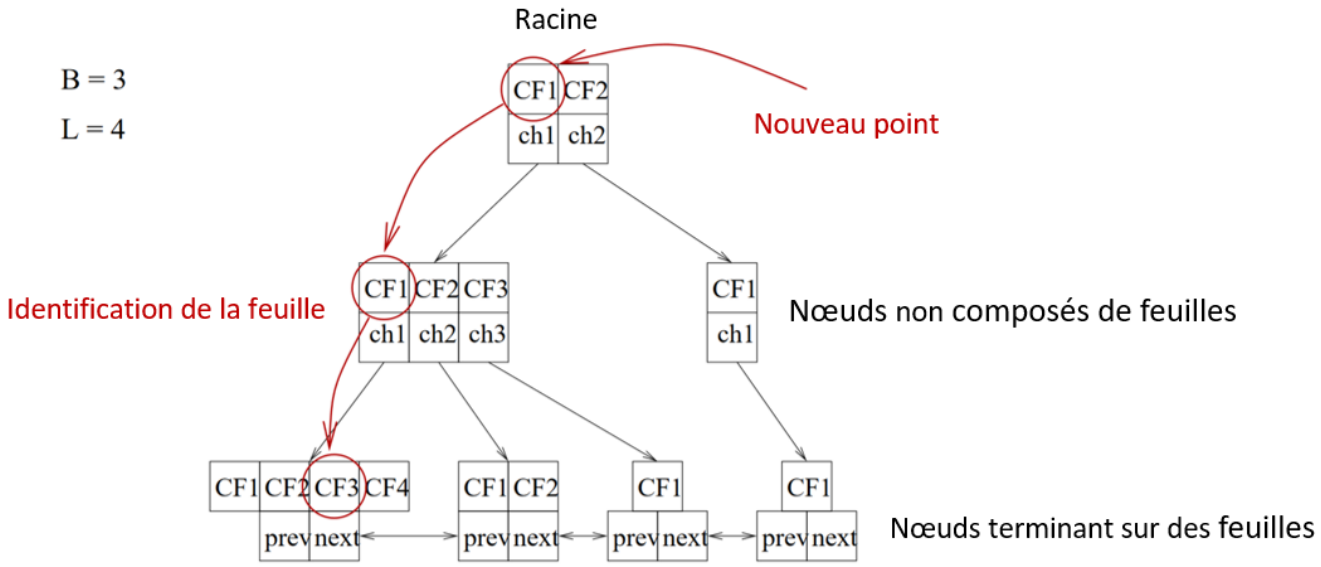

FIGURE 20 – Ajout d'un nouveau point dans l'arbre *CF* [17]

2. Modification de la feuille : si la feuille identifiée à l'étape précédente peut absorber la nouvelle donnée sans dépasser le diamètre T alors on ajoute une nouvelle entrée à cette feuille. Sinon, on divise la feuille, les entrées les plus éloignées (selon la métrique choisie) sont utilisées comme graines (*seeds* en anglais) et les données restantes réparties dans les feuilles selon le critère de la feuille la plus proche.

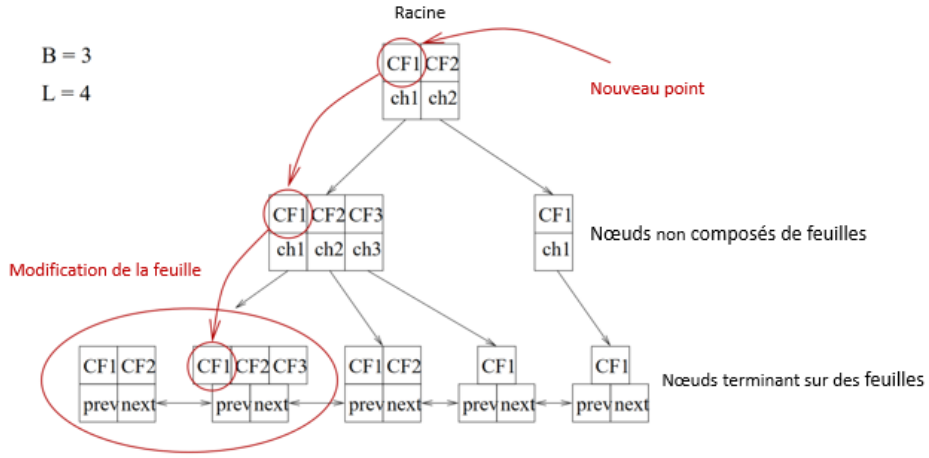

FIGURE 21 – Modification de la feuille suite à l'ajout d'un nouveau point dans l'arbre *CF* [17]

3. Modification du chemin jusqu'à la feuille : Si pas de division de la feuille, alors on met à jour les vecteurs *CF* sur le chemin pour atteindre la feuille. En cas de division de la feuille, on vérifie qu'on ne dépasse pas les *B* entrées pour ce noeud. Si ce critère n'est pas respecté, on procède à une nouvelle division

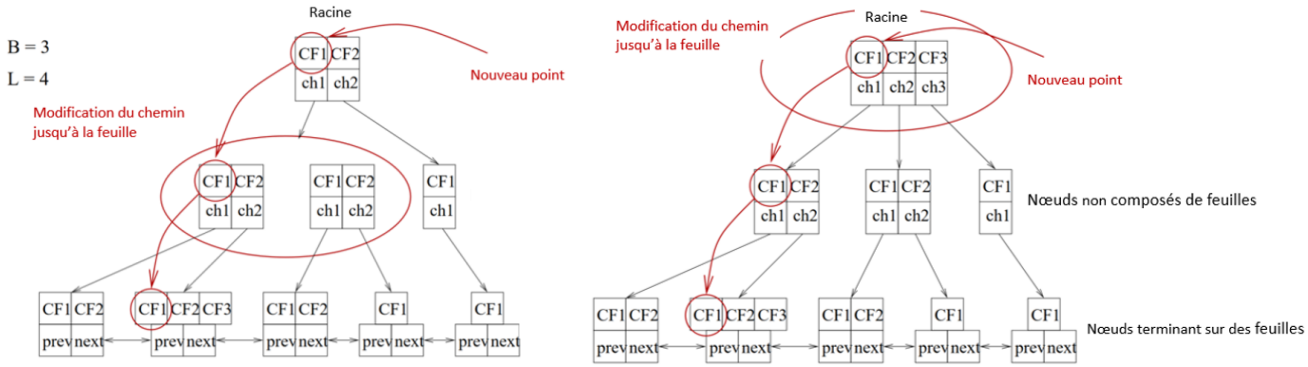

FIGURE 22 – Modification du chemin jusqu'à la feuille suite à l'ajout d'un nouveau point dans l'arbre *CF* [17]

4. Raffinement des noeuds : En cas de division remontant jusqu'à des noeuds non-composés de feuilles, on vérifie que la paire issue de la scission est la plus proche des entrées (toujours selon la métrique choisie). Sinon, on fusionne les entrées les plus proches et on redivise si nécessaire.

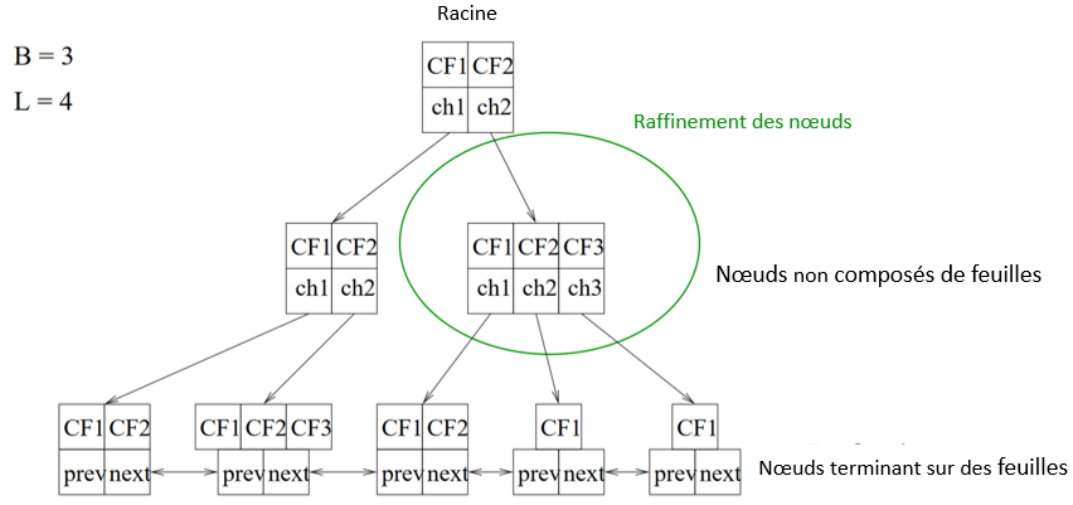

FIGURE 23 – Raffinement des noeuds suite à l'ajout d'un nouveau point dans l'arbre *CF* [17]

On désigne un algorithme de clustering comme global ou semi-global si pour chaque décision de rattacher un point à un cluster, l'algorithme regarde tous les autres points ou tous les clusters existants en les considérant de manière égale, peu importe la distance entre les points ou entre le point et le cluster. De plus, ces algorithmes utilisent des métriques de mesure globales nécessitant l'analyse de tous les points ou de tous les clusters existants.

La seconde étape de l'algorithme BIRCH consiste en l'application d'un algorithme de clustering global sur les sous-clusters issus de l'arbre CF. Cet algorithme de clustering global utilise soit les centroïdes pour représenter les sous-clusters, soit les N points du sous-cluster ou les informations contenues dans les vecteurs CF.

4.8 Méthode de création du score

La création du score à partir du travail précédemment effectué peut se décomposer de la manière suivante :

1. Pour chaque variable sur laquelle le modèle est évalué : celle-ci est discrétisée selon les valeurs de Shapley sur la base de validation à l'aide des observations sur la courbe et de l'algorithme de clustering BIRCH
2. Pour chaque catégorie créée, on lui attribue la valeur de Shapley moyenne des points qui composent la catégorie
3. Pour chaque variable, on corrige la valeur de Shapley de chaque catégorie de la valeur absolue de la valeur de Shapley minimale pour cette variable. Ainsi, la catégorie ayant la valeur de Shapley la plus basse est à 0
4. Calcul de la note maximale que l'on puisse obtenir en additionnant les points maximums attribués pour chaque variable
5. Calcul du facteur de correction égal à $\eta = \frac{1000}{Note\ max}$
6. Tous les points attribués par chaque modalité sont multipliés par le facteur de correction pour obtenir le score final sur 1000

Les étapes 3 à 6 peuvent être représentées de la manière suivante :

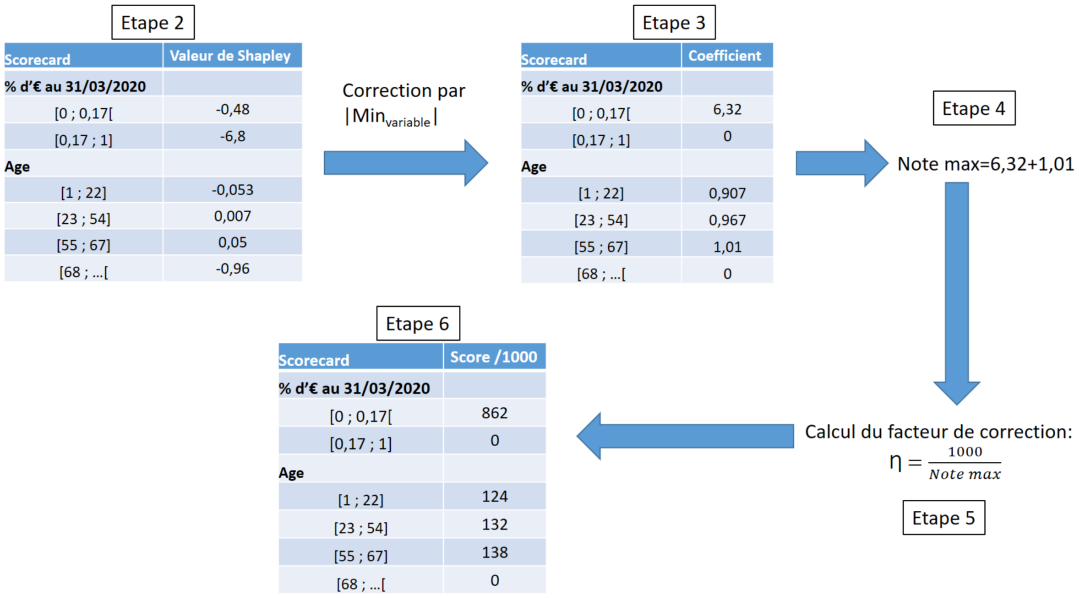

FIGURE 24 – Exemple de création du score pour 2 variables

5 Résultats

Dans cette partie, nous présentons et analysons les résultats obtenus aux différentes étapes de création du score d'appétence à la diversification pour les clients épargne dont la théorie a été précisée dans la partie précédente.

Avant d'effectuer l'apprentissage des modèles, nous avons tout d'abord procédé à une étape de sélection des variables. Un certain nombre de variables présentes dans notre base de données étaient fortement corrélées les unes aux autres. Nous avons, tout d'abord, comparé pour les variables fortement corrélées entre elles, l'importance de chacune pour le modèle selon les valeurs de Shapley. Nous avons ensuite testé l'effet sur les performances du modèle lorsque nous retirions progressivement ces différentes variables. Ainsi, l'ensemble des variables qui concernaient les polices prévoyance détenues par les clients de la base ont été retirées car elles n'avaient pas d'influence sur nos modèles. Au final, nous avons retenu les 10 variables qui avaient le plus d'importance pour la modélisations selon les valeurs de Shapley :

— Pourcentage d'UC au 31/03/2020
— Nombre de polices d'épargne multi-support détenues par le client
— PM totale au 31/03/2020
— Montant du revenu fiscal
— Âge
— Ancienneté
— Code Profession
— Montant des versements programmés entre le 31/03/20219 et le 31/03/2020
— Pourcentage d'action parmi les UC
— Pourcentage d'obligation parmi les UC

A l'aide de l'optimisation bayésienne, nous avons réalisé l'hyperparamétrage du XGBoost. Parmi les principaux paramètres évoqués dans la partie théorique des algorithmes utilisés, nous avons obtenu les résultats suivants :

— Pour SMOTE :

  — Un ratio de la classe minoritaire par rapport à la classe majoritaire (*ratio_minority_over_majority*) de 0,8 soit 45% d'appétents au sein de notre base d'entraînement

  — Un nombre de k plus proches voisins de 5 qui servira à la construction des points synthétiques par SMOTE. Ceci correspond au nombre appliqué par défaut par la fonction SMOTE du package *Imblearn* sous Python.

— Pour le XGBoost :

  — Un taux d'apprentissage de 0,01, ce qui pour ce paramètre est très bas, permet au modèle d'avoir un temps de calcul plus élevé et donc de potentiellement d'augmenter la performance du modèle puisqu'il évaluera plus de cas.

  — Une profondeur maximale des arbres de 10, ce qui est près de deux fois plus élevé que le paramètre par défaut mais permet de créer des arbres plus complexes et permettant posiblement une meilleure performance du modèle.

  — Un nombre d'estimateurs de 492, notre modèle va créer un grand nombre d'arbres (ce paramètre est de 100 par défaut pour le modèle) et donc d'étapes dans le processus de boosting.

  — Un lambda de 0,05, le paramètre de régularisation $L2$ est très peu élevé (il est 1 dans les paramètres par défaut du XGBoost). Notre modèle s'appuiera plus sur la régularisation $L1$ pour éviter un sur-apprentissage puisque le paramètre alpha utilisé est de 1 (contre 0 par défaut).

5.1 Performances du modèle

Pour notre modèle final, nous obtenons à partir des données de la base de test, la courbe Précision-Rappel ci-dessous avec un AUC-PR de 0,96. Un AUC-PR aussi élevé nous indique que notre modèle est capable de très bien identifier les clients appétents et nous permet de nous assurer que le score que nous créons par la suite sera précis.

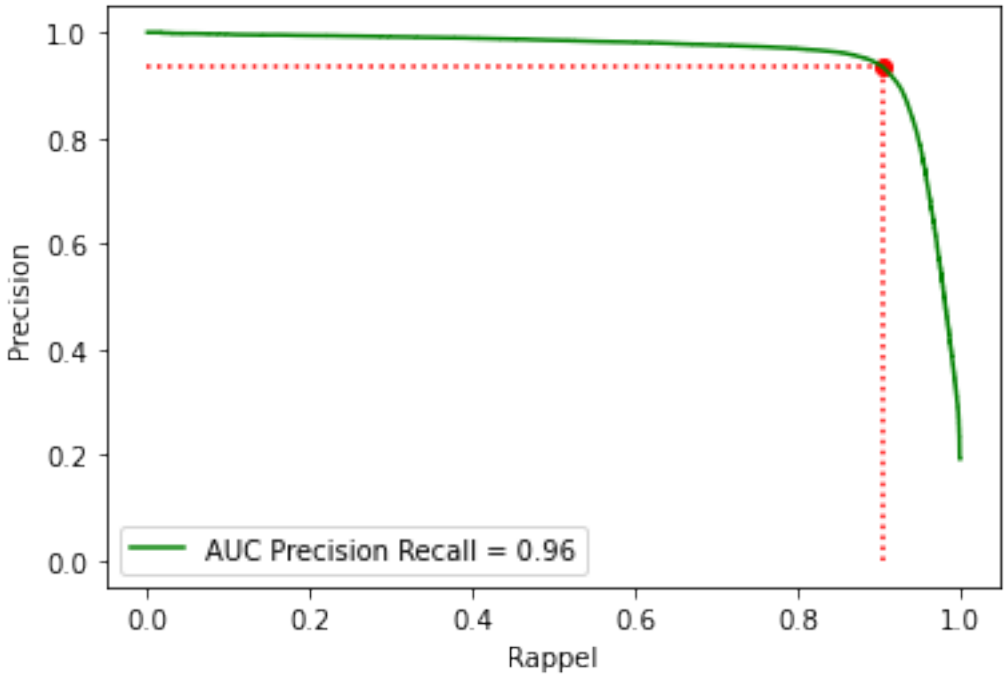

FIGURE 25 – Courbe Précision-Rappel pour le modèle final

Pour les prédictions effectuées à partir de ce modèle, nous considérons que le seuil optimal permettant d'assigner les individus à la classe des appétents est le seuil maximisant le F1-score, soit un seuil de 0,78. La matrice de confusion, sur la base de test, qui en découle est la suivante :

	Non appétents prédits	Appétents prédits
Non appétents réels	421 795	5 920
Appétents réels	8 854	84 695

Ainsi, la précision de notre modèle est donc de 0,905 et le rappel de 0,935. Ces 2 métriques sont représentées en rouge sur la courbe Précision-Rappel présente ci-dessus.

Concernant la courbe ROC et son AUC, nous obtenons un AUC de 0,99 sur la base de test et la courbe ci-dessous. Il est logique que nous obtenions un score si élevé pour l'AUC-ROC car le modèle final est très performant notamment sur les non-appétents.

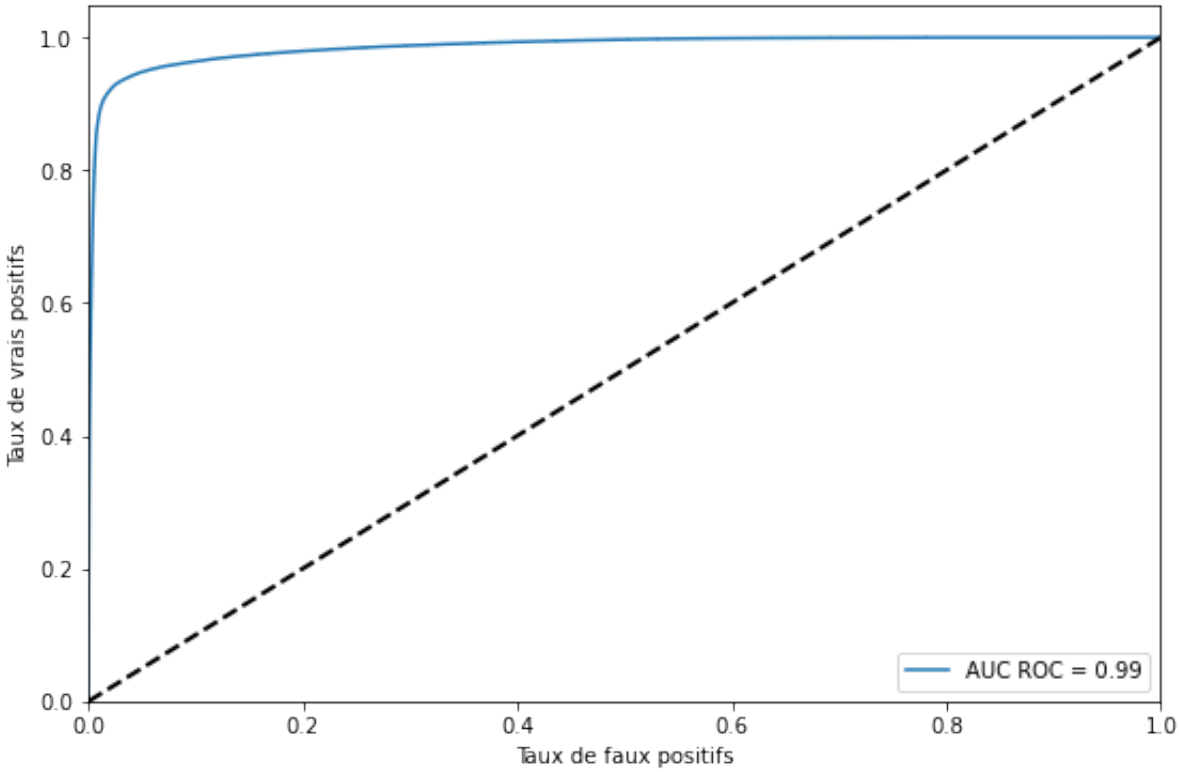

FIGURE 26 – Courbe ROC pour le modèle final

5.2 Importance des variables selon les valeurs de Shapley

L'importance des différentes variables pour le modèle avec les données de la base de validation selon les valeurs de Shapley sont les suivantes :

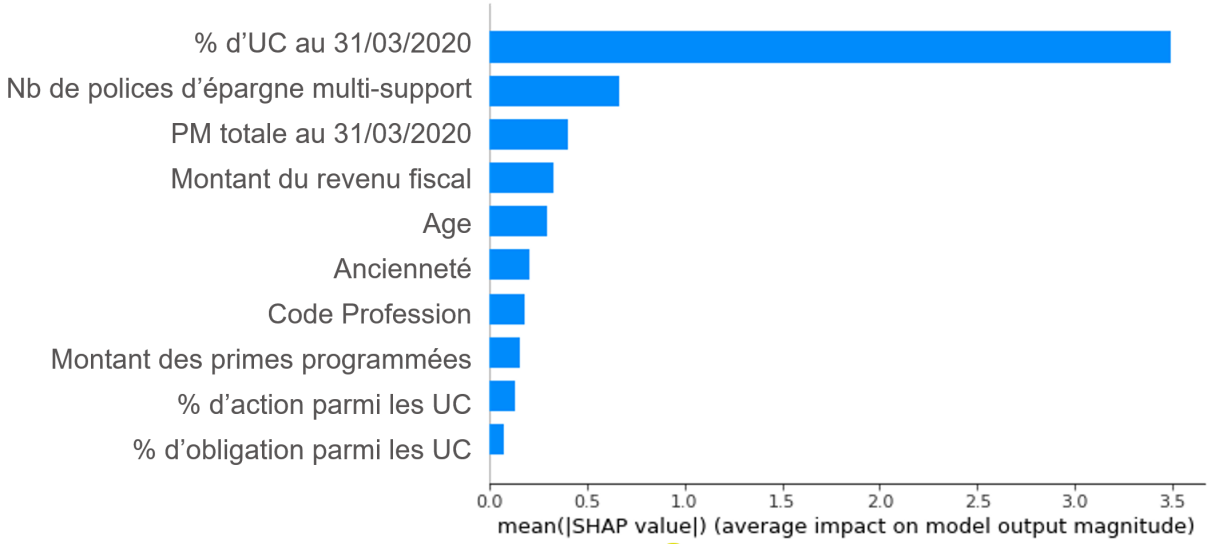

FIGURE 27 – Variables les plus importantes selon les valeurs de Shapley

La variable concernant le pourcentage d'UC au 31/03/2020 est de loin la variable la plus importante de notre modèle. Suivent ensuite, le nombre de polices multi-support détenues, le montant de l'encours total au 31/03/2020. Ces 3 variables concernent les caractéristiques des polices d'épargne du client. Suivent ensuite les variables ayant trait aux caractéristiques du client : le montant de son revenu fiscal, son âge, son ancienneté et sa profession. Les 3 variables restantes vont toucher aux choix de gestion de ses contrats effectuées par le client : le montant des versements programmés effectués par l'assuré sur l'année et comment il répartit son UC sur deux des types de support UC existant.

Pourcentage d'UC au 31/03/2020

Les valeurs de Shapley pour la variable concernant le pourcentage de l'encours du client mis sur des UC au 31/03/2020 se répartissent de la façon suivante :

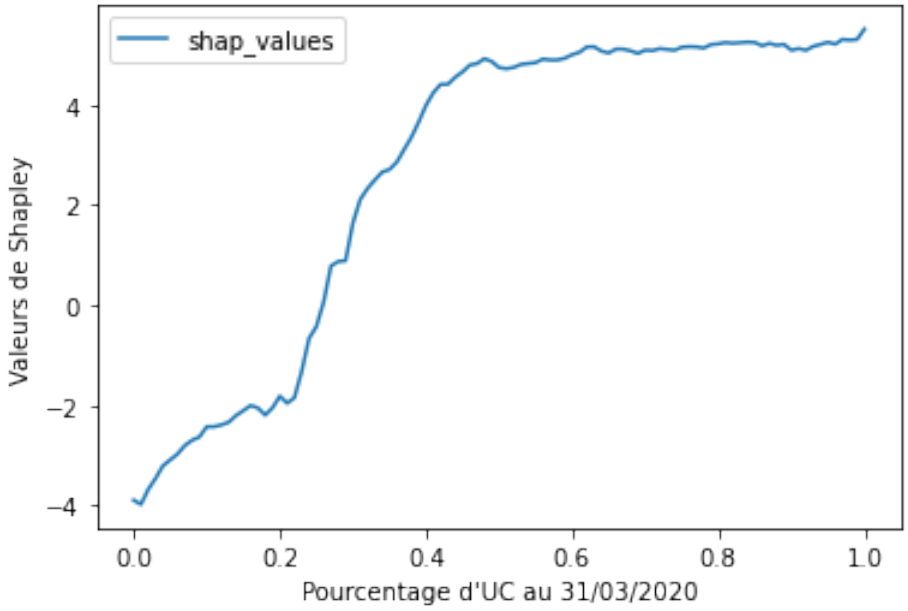

FIGURE 28 – Evolution des valeurs de Shapley moyennes pour la variable représentant le pourcentage de l'encours sur des fonds UC au 31/03/2020

On peut clairement distinguer 2 groupes sur la courbe : l'un entre 0 et 0,2 et l'autre entre 0,4 et 1. A l'aide de l'algorithme de clustering BIRCH, nous créons 2 autres catégories entre 0,2 et 0,4. L'appartenance à la catégorie "[0,4 ; 1]" est celle qui apporte le plus de points pour la variable sur le pourcentage d'UC détenu au 31/03/2020. Une fois que des clients ont dépassé le taux de 40% d'UC, il est rare de les voir repasser sous cette barre, ce qui explique le fait que cette catégorie soit celle qui donne le plus de points pour cette variable.

Nombre de polices d'épargne multi-support

Pour la variable dénombrant le nombre de polices multi-support, cette variable est découpée en 2 catégories, une pour les clients ne possédant pas de police multi-support et une pour ceux en possédant au moins une. Cette dernière catégorie étant, logiquement, celle qui rapporte le plus de points puisqu'il est nécessaire d'avoir une police multi-support afin de posséder des UC.

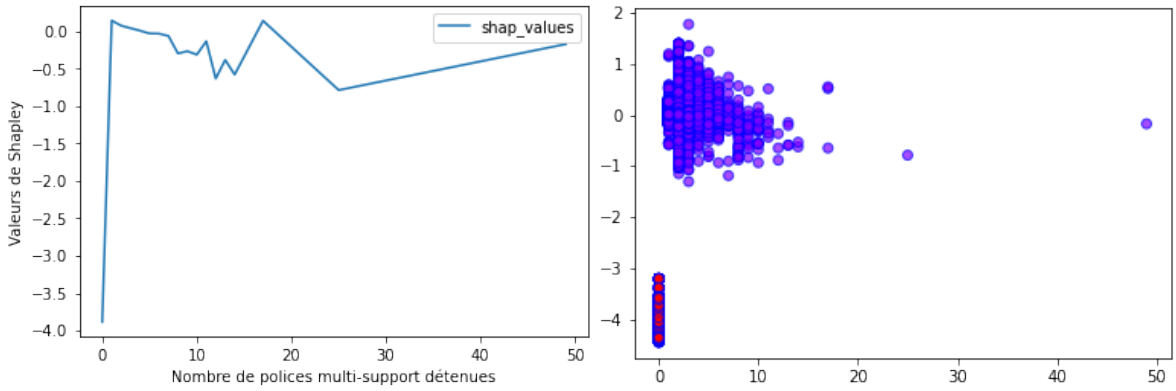

FIGURE 29 – Valeurs de Shapley pour la variable comptant le nombre de polices multi-support détenues et le clustering à l'aide de l'algorithme BIRCH

Encours total au 31/03/2020

La variable sur le montant total de l'encours du client au 31/03/2020 est découpée en 7 catégories. Le nombre de points attribué par le score est croissant jusqu'au niveau des 35 000€ d'encours avant de décroître pour les montants supérieurs. Pour les personnes ayant plus de 200 000€ d'encours, le score ne leur attribue aucun point sur cette variable. Cette répartition des points pour les encours très importants est issue d'une spécificité du portefeuille de Crédit Agricole Assurances. En effet, la politique de collecte décrite précédemment a été mise en place par Crédit Agricole Assurances que tardivement par rapport au reste des acteurs du marché de l'épargne. Ceci entraînant l'afflux au sein de son portefeuille de clients disposant d'un encours élevé et averses au risque donc ne souhaitant pas placer leur encours sur des UC.

Revenu fiscal

Nous avons divisé la variable concernant le montant du revenu fiscal en 4 catégories. La première, allant de 0 à 25 000€ de revenu fiscal, n'octroie aucun point dans notre score. Cette catégorie englobe pratiquement l'ensemble des clients se trouvant dans les 2 premières tranches d'impôt sur le revenu, celles-ci concernent, en 2021, les personnes ayant un revenu fiscal annuel inférieur à 25 710€ (la 1ère tranche va de 0 à 10 084€). On suppose que ces personnes ayant peu de revenus souhaitent sécuriser leur encours et donc ne pas prendre d'UC. La catégorie allouant le plus de points pour cette variable est celle des personnes ayant un revenu fiscal compris entre 95 000€ et 200 000€. Celle-ci attribue légèrement plus de points que celle des clients ayant un revenu fiscal de plus de 200 000€. Nous supposons que cela vient de l'affluence des clients ayant des montants d'encours élevés mais pas d'UC comme détaillé dans la partie sur le score pour la variable sur le montant total de l'encours du client au 31/03/2020.

Age

La variable concernant l'âge du client a été découpée en 7 catégories. Nous avons imposé 2 catégories au vu de leurs spécificités légales : celle des moins de 18 ans et des plus de 70 ans. Pour les moins de 18 ans, la gestion des polices d'épargne est assurée par ses représentants légaux et une fois le cap des 70 ans passé, la fiscalité appliquée sur les versements des contrats d'épargne change et passe à celle des droits de succession en cas de décès de l'assuré. La catégorie des 18-24 ans est celle qui rapporte le plus de points pour l'âge. On peut supposer qu'au vu des conditions de marché actuelles, les jeunes personnes acceptent de posséder plus d'UC car celles-ci permettent, en général, d'avoir un meilleur rendement. Notre modèle distingue un changement de comportement des personnes atteignant l'âge légal de départ à la retraite puisque l'une des catégories du score d'appétence à la diversification pour la variable âge concerne les personnes ayant entre 62 et 69 ans. Une fois l'âge de la retraite atteint, le nombre de points attribués par notre score diminuent, nos clients souhaitent réduire les risques pris avec leur encours et donc possèdent un taux d'UC moins élevé. Les plus de 70 ans est la classe qui, pour la variable âge, ne reçoit aucun point. A partir de cet âge, les contrats d'épargne sont

gérés dans l'optique de succession et les clients souhaitent prendre le moins de risque possible avec leur encours.

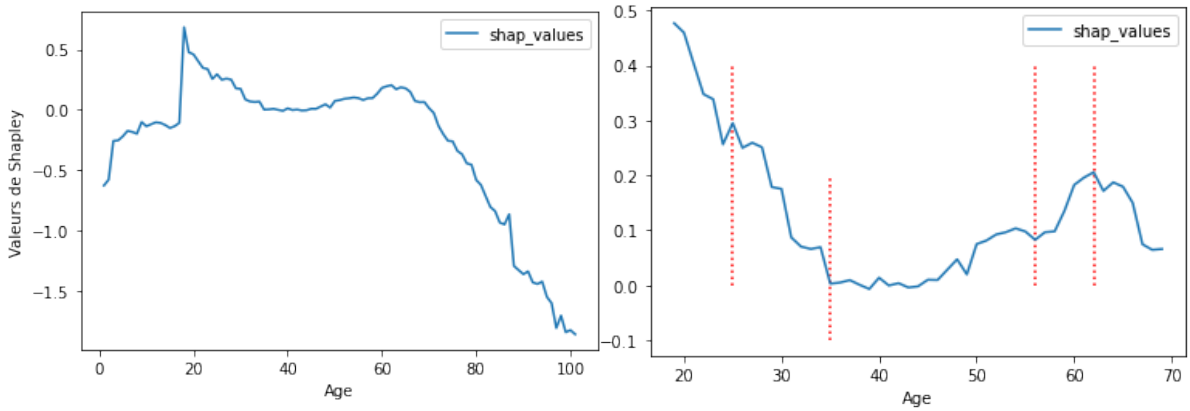

FIGURE 30 – Valeurs de Shapley pour l'âge et la discrétisation de la variable entre 18 et 70 ans

Ancienneté

Pour l'ancienneté de nos clients, la première catégorie que nous avons définie concerne les personnes ayant moins de 2 ans d'ancienneté, la seconde les clients ayant entre 2 et 12 ans d'ancienneté. On suppose qu'une fois le cap des 2 ans d'ancienneté passé, nos clients sont sûrs de vouloir garder leur police d'épargne et souhaitent donc obtenir un rendement plus élevé. Ce qui explique que cette seconde catégorie des 2-12 ans d'ancienneté soit celle qui rapporte le plus de points pour cette variable. Les points attribués pour les autres catégories sont décroissants avec l'ancienneté et aucun point n'est attribué pour les clients ayant plus de 30 ans d'ancienneté. Cela vient du fait que plus l'ancienneté, et donc l'âge, de nos clients augmente, plus ces derniers souhaitent réduire les risques pris sur leurs polices d'épargne et donc diminuent la part de leur encours placé sur des UC.

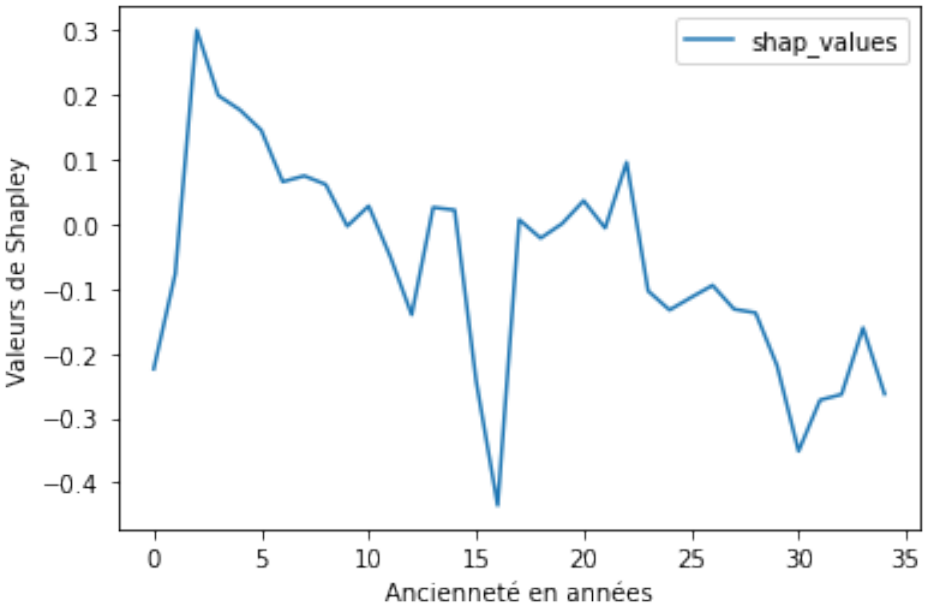

FIGURE 31 – Evolution des valeurs de Shapley moyennes pour la variable représentant l'ancienneté des clients en années

Cependant, les variables concernant l'âge et l'ancienneté des clients ont une corrélation importante, de 0,5. Nous avons vérifié si pour une même ancienneté, notre modèle traite différemment nos client selon leur âge. Sur le graphique ci-dessous, on ne constate pas de différences importantes entre les catégories d'âge pour une même ancienneté. Lorsque les catégories d'âge atteignent le maximum d'ancienneté possible (par exemple, les personnes de moins de 18 ans ne peuvent avoir plus de 18 ans d'ancienneté), il existe des divergences dues au faible nombre d'individus ayant une police d'épargne au Crédit Agricole Assurances depuis leur naissance. Pour la même raison, on observe des divergences entre les catégories d'âge pour les clients ayant moins d'un an d'ancienneté ce qui ne nous permet pas de tirer de conclusions sur les comportements de ces personnes.

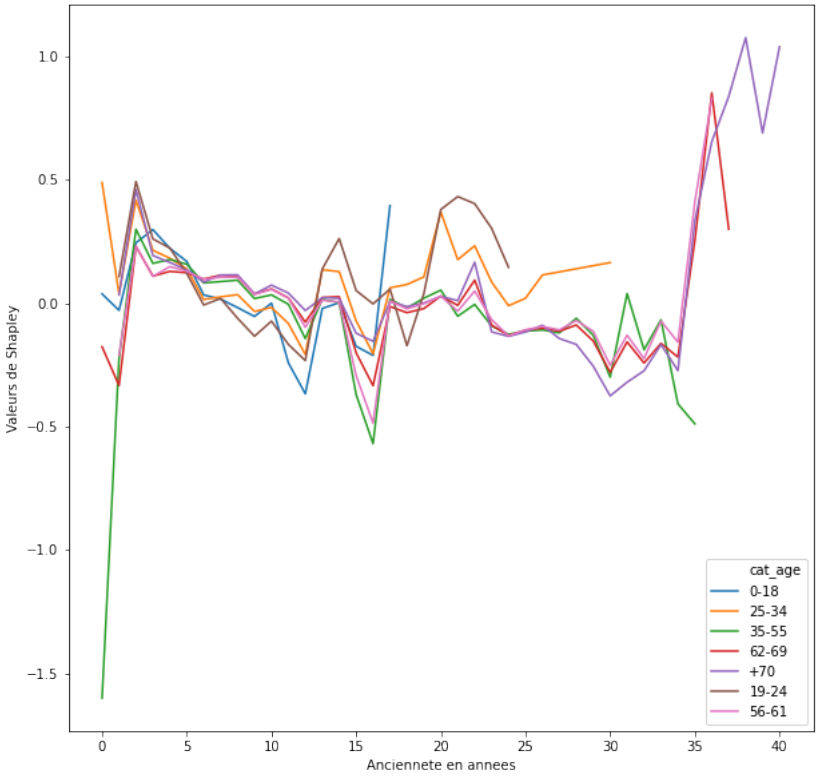

FIGURE 32 – Evolution des valeurs de Shapley selon l'ancienneté pour les différentes catégories d'âge

Profession

Les professions sont réparties en 8 catégories suivant la nomenclature des professions et catégories socioprofessionnelles de 2003 (PCS-2003) définie par l'INSEE [12].

Les 8 catégories sont :

— Agriculteurs exploitants

— Artisans, commerçants et chefs d'entreprise

— Cadres et professions intellectuelles supérieures

— Professions intermédiaires

— Employés

— Ouvriers

— Retraités

— Autres personnes sans activité professionnelle

Les agriculteurs sont la catégorie socioprofessionnelle attribuant le plus de points dans notre score. Ce groupe est la clientèle cible historiquement de Crédit Agricole et bénéficie de produits spécifiques, contrairement aux autres catégories professionnelles, ce qui peut expliquer une plus forte attache des clients aux suggestions de leurs conseillers. La seconde catégorie socioprofessionnelle se voyant attribué le plus de points dans notre score est celle des cadres et professions intellectuelles supérieures. On suppose que ces personnes ayant un niveau d'étude plus élevé que les autres catégories socioprofessionnelles, elles sont mieux à même d'évaluer les risques pris en placement un pourcentage important de leur encours sur de l'UC.

Montant des versements programmés

Pour la variable sur le montant des versements programmés effectués entre le 31/03/2019 et le 31/03/2020, nous divisons cette variable en 2 groupes, à l'aide de l'algorithme BIRCH, les clients ayant versé plus ou moins de 5 000€ sur cette période. Seuls 0,5%

de nos clients ont effectués plus de 5 000€ de versements programmés sur leurs polices d'épargne durant cette période. Toutefois, ces clients sont distingués dans notre score et sont ceux qui reçoivent le plus de points pour cette variable. Le versement programmé de montants importants sur les polices d'épargne peut indiquer la volonté de faire fructifier ces montants et donc d'avoir un taux d'UC plus élevé permettant un taux de rendement plus élevé.

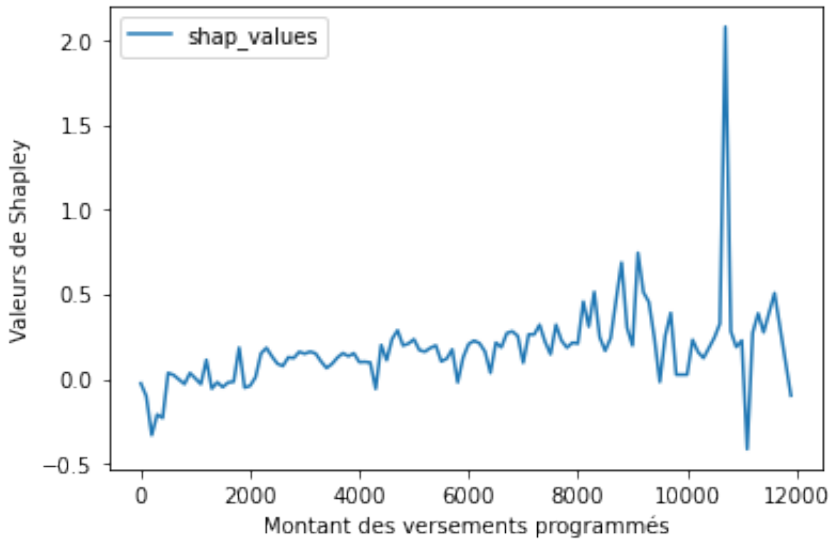

FIGURE 33 – Évolution des valeurs de Shapley selon le montant des versements programmés

Pourcentage d'actions parmi les UC

Pour la part des UC placés sur des actions, les points attribués par notre score sont croissants avec ce taux jusqu'à atteindre le seuil des 85% de l'UC placé sur des actions. Une fois ce seuil des 85% d'actions parmi les UC dépassé, notre score octroie beaucoup moins de points. Les actions étant les UC les plus risquées, certains des clients rentrant dans cette dernière catégorie vont prendre des risques importants mais seulement pour une partie limitée de leur encours. Ainsi, ils auront un taux d'UC plutôt faible. Nous n'allouons aucun point pour les personnes ayant moins de 30% d'actions parmi leur UC.

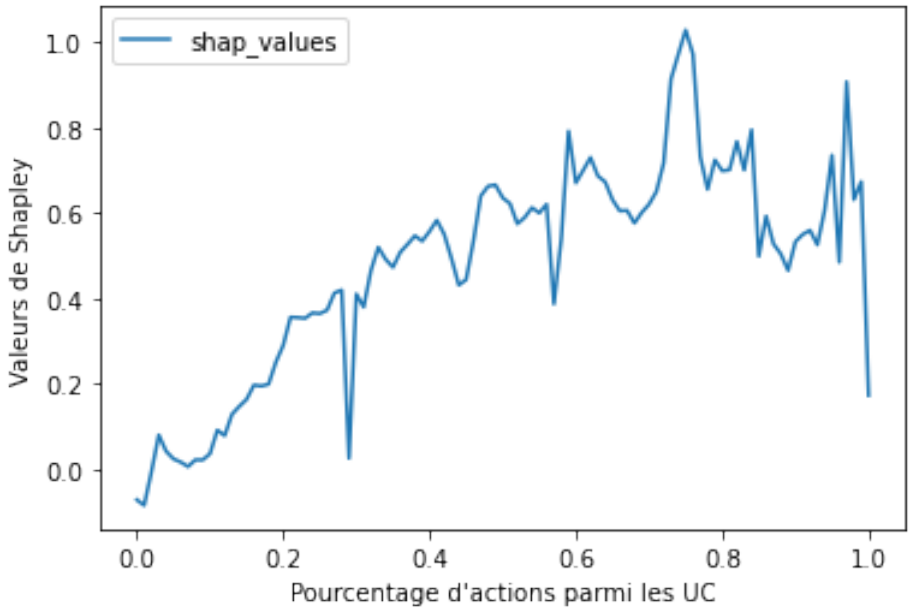

FIGURE 34 – Évolution des valeurs de Shapley selon la part des UC sur des supports actions

Pourcentage d'obligations parmi les UC

Pour la part des UC placés sur des obligations, nous avons découpé cette variable en 5 catégories. Les personnes ayant moins de 22% d'obligations parmi leurs UC ne reçoivent aucun. Les personnes ayant plus de 55% de leurs UC sur des obligations se repartissent en 2 catégories qui se voient attribuer un nombre de points proches représentant le maximum de points qu'un client puisse obtenir pour cette variable. A l'inverse des comportements constatés sur les personnes ayant plus de 85% de leurs UC sur des actions, ici certaines personnes vont avoir un taux d'UC élevé tout en investissant fortement sur des obligations qui sont beaucoup moins risquées.

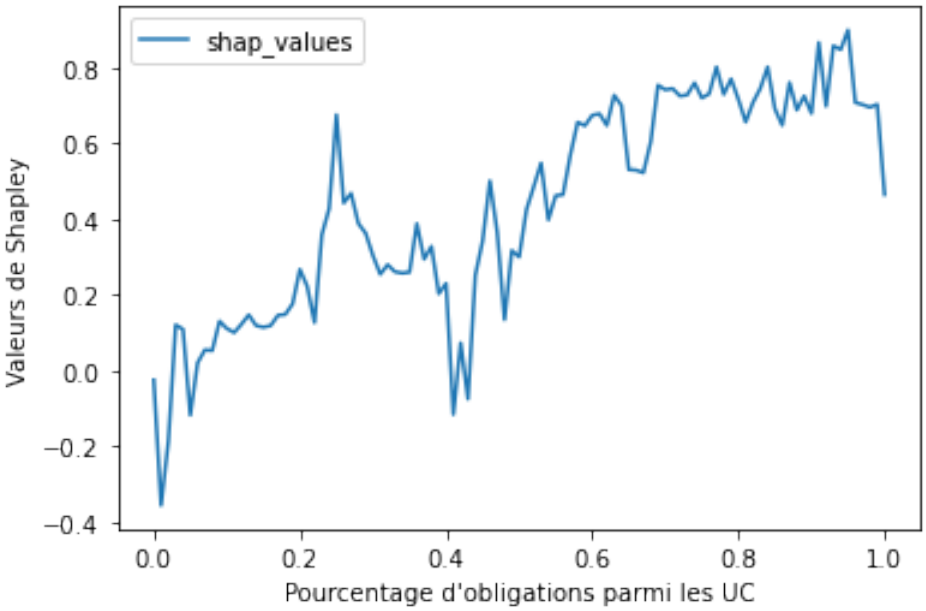

FIGURE 35 – Évolution des valeurs de Shapley selon la part des UC sur des supports obligations

Du fait de ces 2 comportements clients différents entre les personnes ayant un taux élevé d'obligations ou d'action au sein de leurs UC, il n'est pas possible pour un client d'obtenir un score de 1 000 points sur 1 000.

5.3 Adéquation du score créé

En appliquant, notre score d'appétence à la diversification sur 1 000 points à l'ensemble des clients présents dans nos données, les répartitions suivantes sont obtenues pour les assurés appétents et pour les assurés non-appétents :

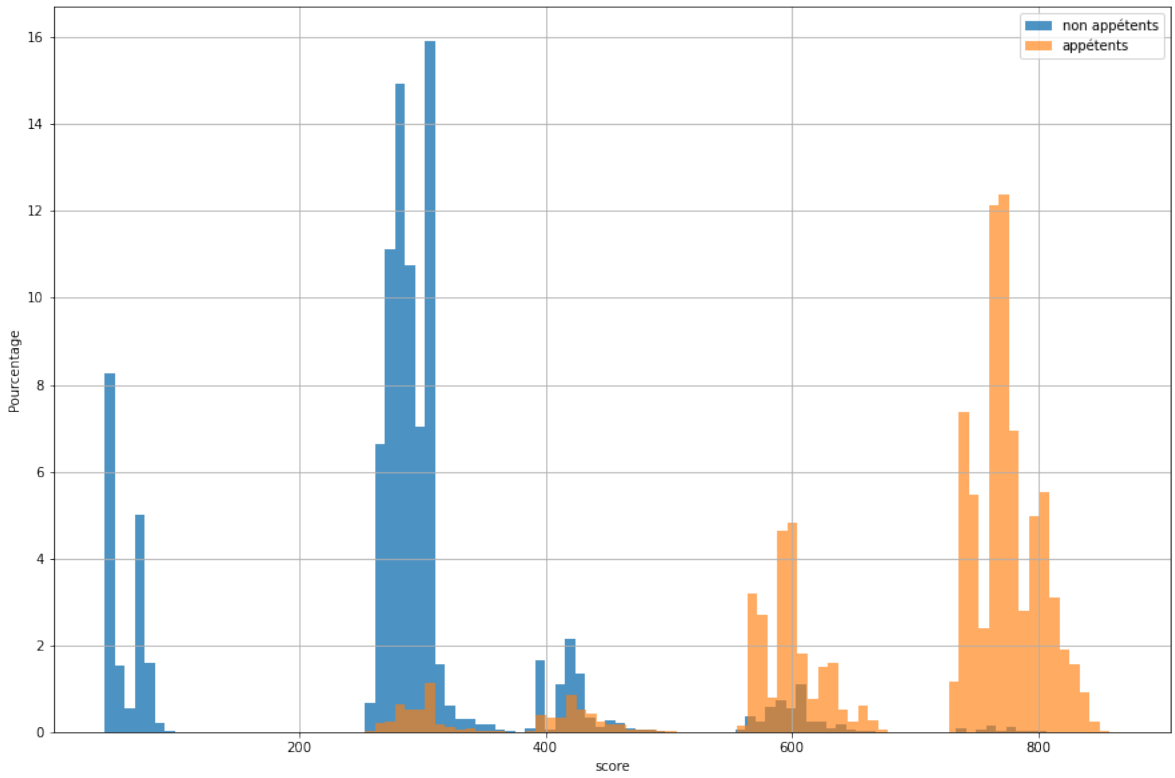

FIGURE 36 – Répartition des scores selon l'appétence des clients

Notre score est capable de clairement distinguer les clients appétents des clients non-appétents. On peut remarquer sur le graphique ci-dessus l'existence de 5 pics, nous qualifierons ces pics de classes d'appétence. Elles se répartissent de la façon suivante selon le score :

— Entre 40 et 100

— Entre 250 et 330

— Entre 380 et 450

— Entre 550 et 680

— Entre 720 et 850

Groupe des retraités : score de 40 à 100

La première classe d'appétence identifiable sur le graphique ci-dessus concerne les clients ayant un score entre 40 et 100. Ces derniers représentent 14% des clients de notre base de données. Nous surnommons cette classe, le groupe des retraités. En effet, cette classe d'appétence est composée à près de 50% de retraités. Ce groupe est donc beaucoup plus âgé que la moyenne de l'ensemble de la base, 70 ans pour les clients membres de ce groupe contre 55 ans pour l'ensemble des clients de la base de données, comme on peut le voir sur le graphique ci-dessous. Ils ont aussi, logiquement, une ancienneté élevée, de 21 ans en moyenne. Ces personnes ne possèdent pas de police d'épargne multi-support et ont 0% d'UC. L'encours total médian de ce groupe est près de 2 fois plus élevé que pour l'ensemble de nos clients, 16 255€ contre 8 554€ pour l'ensemble de la base. Toutefois, l'encours moyen est à peu près du même niveau que pour l'ensemble de la base, aux alentours de 40 000€. Le revenu fiscal médian de ce groupe est de 21 386€, ce qui équivaut pratiquement au niveau de vie médian annuel des retraités en 2018 qui était de 22 200€ [7]. Il semble très peu probable de voir ce groupe accepter de placer une partie de leur encours sur des UC.

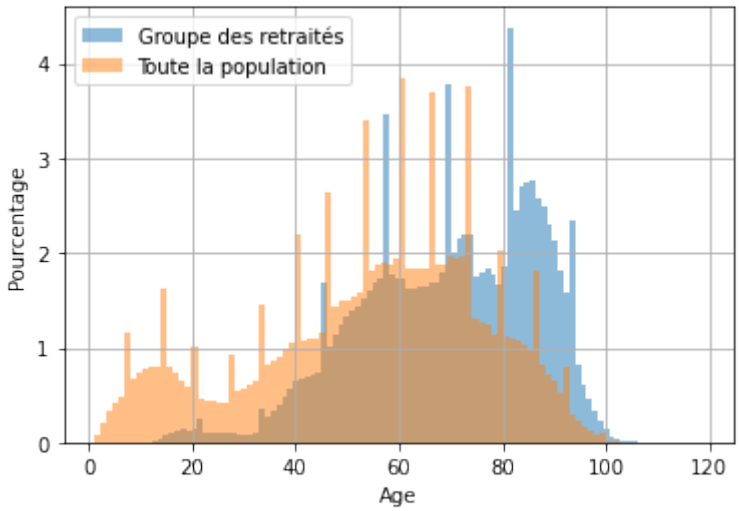

FIGURE 37 – Répartition de l'âge pour les personnes ayant score entre 40 et 100 et pour l'ensemble des clients de la base

Groupe des petits épargnants : score de 250 à 330

Le groupe suivant est composé des personnes ayant un score allant de 250 à 330. Ce groupe rassemble près de 58% des clients de la base et est surnommé le groupe des petits épargnants. En effet, l'encours total médian de ce groupe est de 4 648€ contre près de deux fois plus pour l'ensemble de la base. Les clients de ce groupe possèdent au moins une police d'épargne multi-support mais peu voire pas d'UC. En effet, 80% des membres de ce groupe possèdent moins de 5% d'UC. Le revenu fiscal médian de ce groupe est de 22 667€, ce qui est très proche du revenu fiscal médian des français qui s'élevait en 2017 à 21 120€ [11]. Ce groupe est plus jeune que l'ensemble de la base avec un âge moyen de 49 ans contre 55 ans pour l'ensemble de nos données. De même, l'ancienneté moyenne est plus basse avec 11 ans d'ancienneté en moyenne contre 14 ans pour l'ensemble de la base. Au sein de ce groupe, il existe une sur-représentation de certaines catégories socioprofessionnelles : les personnes sans activité, les ouvriers et les employés.

Groupe des clients moyens : score de 380 à 450

Le groupe des personnes ayant un score compris entre 380 et 450 contient 6% des clients de notre base. Les caractéristiques des clients de cette classe d'appétence sont très proches de ceux du client moyen. Ainsi, nous surnommons cette classe d'appétence, le groupe du clients moyen. L'âge moyen de ce groupe est de 56 ans et l'ancienneté moyenne de 14 ans, le client moyen de la base à 55 ans et la même ancienneté. De même, l'encours moyen est de près de 40 000€. Toutefois, les clients de ce groupe possèdent un peu plus d'UC que le client moyen, 23% contre 15% pour le client moyen. Les clients de ce groupe mettent en moyenne 9% de leurs UC sur des actions et 3% sur des obligations.

Le groupe du client moyen contient 9% d'appétents. Les appétents et les non-appétents de cette classe ont à peu près le même taux d'UC en moyenne. Toutefois, les appétents ont des encours moins élevés que les non-appétents en moyenne, 25 630€ pour les appétents, 42 120€ pour les non-appétents. Les appétents vont en moyenne mettre 38% de leurs UC sur des supports actions contre 7% pour les non-appétents. De plus, les appétents sont plus jeunes avec un âge moyen de 51 ans contre 57 pour les non-appétents.

Sur le graphique ci-dessous, on peut constater un pic sur les moins de 30 ans chez les appétents à la diversification. A l'inverse, on constate un plus grand nombre de clients de plus de 65 ans chez les non-appétents.

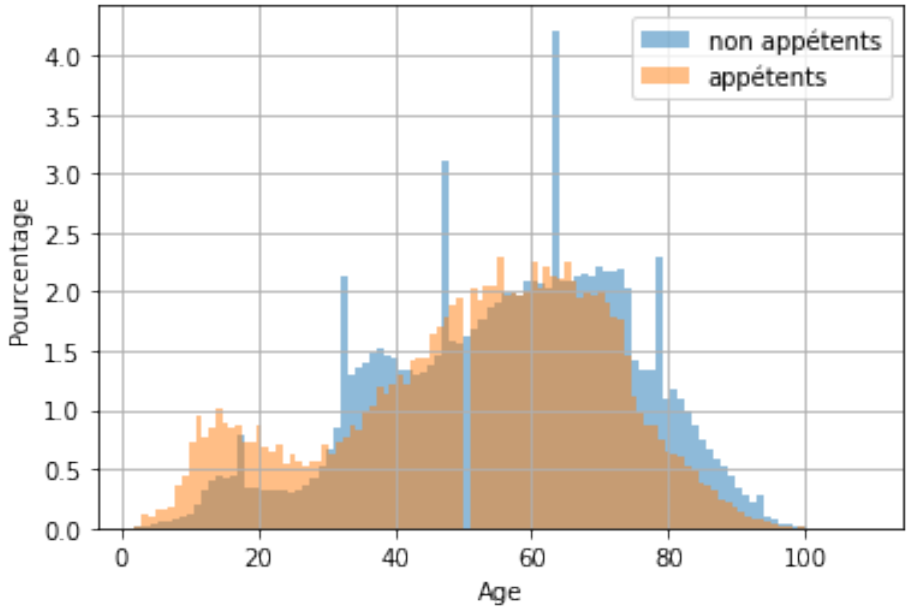

FIGURE 38 – Répartition de l'âge pour les personnes ayant score entre 380 et 450 selon leur appétence

Groupe des petits appétents : score de 550 à 680

Les clients ayant un score compris entre 550 et 680 appartiennent au groupe dit des petits appétents. 8% des clients de notre base appartiennent à ce groupe. Ce groupe a été nommé ainsi car le taux moyen d'UC de ces clients est de 32%. Ces clients possèdent un encours moyen plus élevé, de près de 50%, que la moyenne de l'ensemble des clients avec un encours moyen de 62 000€. Le revenu fiscal moyen de ces clients est de près 60 000€, ce qui les place au niveau du 9e décile dans la distribution du niveau de vie des français dont le niveau de vie moyen en 2019 est de 60 170€ [10]. Ces clients sont en moyenne un peu plus âgé que l'ensemble de la base avec un âge moyen de 59 ans. L'ancienneté moyenne de cette classe d'appétence est de 13 ans contre 14 ans pour l'ensemble de la base. Toutefois, la répartition de l'ancienneté diffère par rapport à celle de l'ensemble de la base pour les personnes ayant moins de 12 ans d'ancienneté. En

effet, sur le graphique ci-dessous, on constate un pic entre 2 et 4 ans d'ancienneté suivi d'un creux. Les petits appétents placent en moyenne, 14% de leurs UC sur des actions et 11% sur des obligations. Les clients appartenant à ce groupe sont de potentiels clients à cibler afin de les amener à augmenter leur taux d'UC.

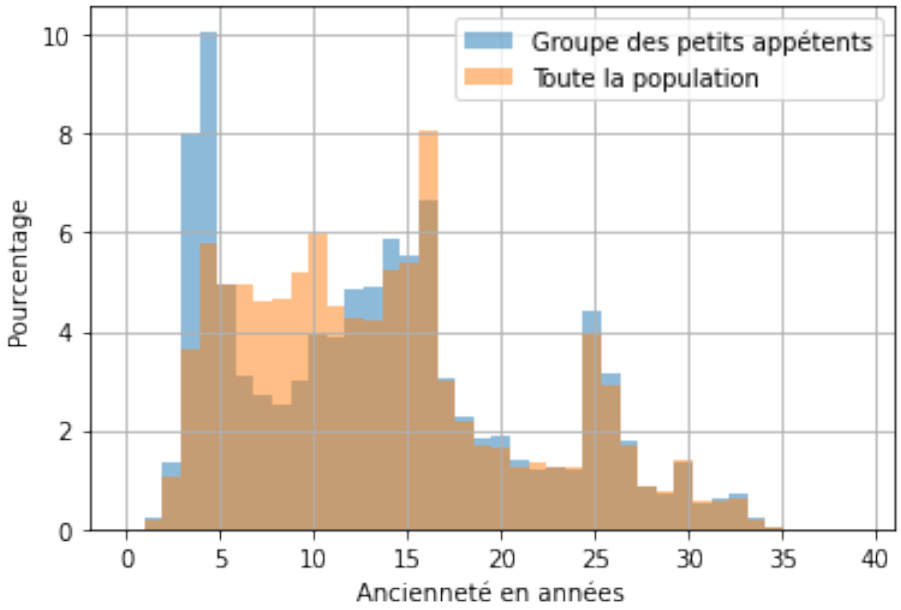

FIGURE 39 – Répartition de l'ancienneté pour les personnes ayant score entre 550 et 680 et pour l'ensemble des clients de la base

Le groupe des petits appétents contient 54% d'appétents selon notre critère d'appétence à la diversification. Les appétents et les non-appétents ont en moyenne le même taux d'UC. Toutefois, ils diffèrent largement sur le montant de leurs encours, l'encours médian pour les appétents est de 11 780€ contre plus de 4 fois plus, à 54 000€ pour les non-appétents. Les non-appétents de ce groupe sont aussi beaucoup plus âgés, avec un âge moyen de 62 ans contre 56 ans pour les appétents. Les appétents de ce groupe vont placer près de 19% de leurs UC sur des actions alors que les non-appétents vont seulement en placer 8%.

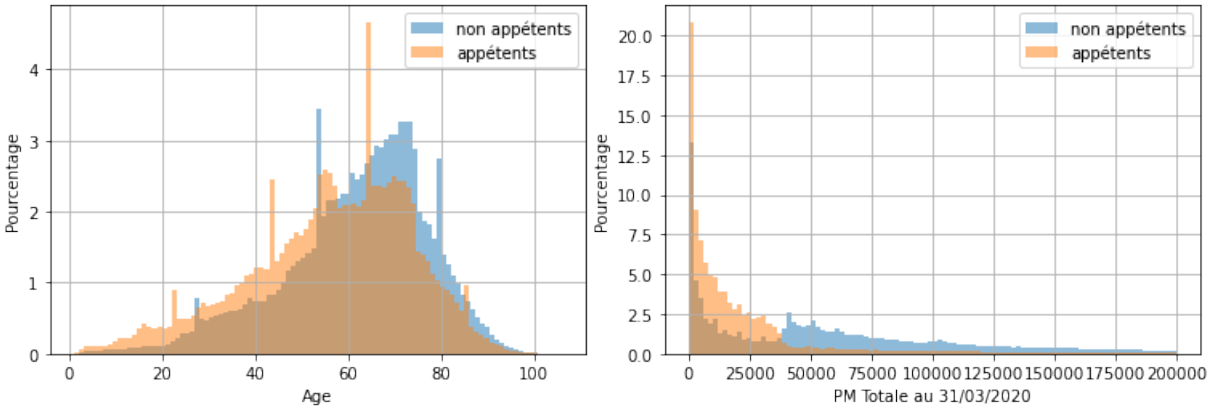

FIGURE 40 – Répartition de l'âge et de l'encours des personnes ayant un score entre 550 et 680 selon leur appétence

Groupe des grands appétents : score de 720 à 850

Le dernier groupe identifiable sur la répartition du score sur l'ensemble des clients de notre base est le groupe des personnes ayant un score entre 720 et 850 qui regroupe 13% des clients de la base. Ce groupe est surnommé celui des grands appétents du fait de son taux d'UC moyen par client soit élevé. En effet, ce dernier est de 66% en moyenne. Ces grands appétents ont un encours moyen très élevé, ce dernier est de 74 240€, ce qui est près de 75% plus que l'encours moyen pour l'ensemble de la base. De la même manière que le groupe précédent, ces clients appartiennent au 9ème décile avec un revenu fiscal moyen de près de 71 000€. En moyenne, les grands appétents investissent 24% de leurs UC sur des actions et 12% sur des obligations. Dans ce groupe, on constate une sur-représentation des cadres et professions intellectuelles supérieures. Les grands appétents sont des clients que l'on peut possiblement amener à augmenter leur taux d'UC via des actions marketing ciblées.

6 Limites et améliorations possibles

Le score que nous avons construit comporte quelques limites. Tout d'abord, certaines variables utilisées, la profession et le revenu fiscal, ne sont pas mis à jour de manière régulière et nous ne connaissons pas la régularité de cette mise à jour. On peut supposer qu'elles sont au moins mises à jour lors de la souscription d'un nouveau contrat chez Crédit Agricole Assurances ou au sein du réseau bancaire (LCL et Crédit Agricole) pour les personnes également client bancaire.

Durant la construction du score, nous avons essayé d'intégrer des données venant d'autres types de produits commercialisés par Crédit Agricole Assurances : les assurances IARD et les assurances emprunteurs. Toutefois, ces informations n'ont pas permis d'améliorer les performances de notre modèle du fait d'un taux de multi-équipement trop peu élevé chez les clients de notre base de données (un client multi-équipé est un client qui possède au moins une police dans 2 catégories d'assurance parmi l'épargne, l'emprunteur, la prévoyance et le IARD). Il pourrait être intéressant d'ajouter des données externes à notre base, notamment des données relatives aux marchés financiers et aux performances des fonds UC.

Il a été choisi de créer ce score pour l'ensemble des clients épargne de Crédit Agricole Assurances, nous obligeant ainsi à agréger nos données à la maille client. Ce niveau d'agrégation a entraîné une perte d'informations contenues au niveau des polices détenues par les assurés. Ainsi, il serait peut être intéressant de créer un score d'appétence à la diversification pour chacun des plus gros produits du portefeuille de Crédit Agricole Assurances afin de mieux prendre en compte la spécificité des polices. Dans la même idée, il pourrait être intéressant de différencier le score selon l'appartenance du client au réseau LCL ou aux caisses régionales du Crédit Agricole.

Ce score a pour objectif d'accompagner les équipes marketing sur le ciblage client dans le but d'augmenter la part d'UC au sein des polices d'épargne des clients de Crédit Agricole Assurances. Il sert donc à cibler les clients qui possèdent déjà des UC et ne peut être utilisé pour cibler les clients ne possédant pas d'UC à en prendre. Cet objectif nécessiterai la réalisation d'une autre étude.

Afin de suivre l'évolution des comportements clients, sur lesquels, on peut déjà

constater les effets de la crise de la Covid-19, ce score d'appétence à la diversification devra être mis à jour régulièrement. Il est actuellement prévu de le mettre à jour de manière annuelle. De plus, le score sera recalculé mensuellement pour chaque client épargne de Crédit Agricole Assurances.

7 Conclusion

La création du score d'appétence à la diversification décrit dans ce mémoire a tout d'abord nécessité l'utilisation de l'algorithme SMOTE sur nos données car celles-ci étaient déséquilibrées. La phase de modélisation s'est effectuée à l'aide du modèle XGBoost, un algorithme de boosting. Les résultats du modèle et les prédictions ont été interprétées à l'aide des valeurs de Shapley, nous permettant ainsi de connaître l'importance des différentes variables pour le modèle. Une fois l'importance des différentes variables dans notre modèle connue, nous avons dû discrétiser nos variables continues en partie à l'aide de l'algorithme de clustering BIRCH, ceci afin de pouvoir construire notre score d'appétence à la diversification.

Le score a été construit à partir de 10 variables : 3 concernent les caractéristiques des polices épargne du client, 4 les caractéristiques du client et 3 aux choix de gestion de ses polices épargne effectuées par le client. Le modèle de machine learning utilisé a obtenu de très bons résultats dans la prédiction du critère d'appétence à la diversification, nous permettant ainsi de construire un score précis vis-à-vis de notre objectif. Le score que nous avons obtenu nous a permis de bien séparer les appétents des non-appétents et d'identifier 5 classes d'appétences qui ont toutes des caractéristiques propres. Parmi ces 5 classes d'appétence, 2 semblent pouvoir être aisément ciblées par les équipes marketing.

Outre l'aide au ciblage des clients par les équipes marketing, ce score servira aussi notamment au calcul d'une participation au bénéfice préférentielle récompensant les clients les plus appétents vers les UC. Suite à l'industrialisation du score suivra une période d'évaluation de son utilisation réelle ainsi que de la performance du ciblage marketing à l'aide cet outil. Il pourrait être intéressant par la suite de coupler ce score avec les autres outils de ciblage existants ou qui seront développés dans le futur afin d'affiner le ciblage selon les objectifs choisis et de faciliter leur emploi par les utilisateurs.

8 Annexes

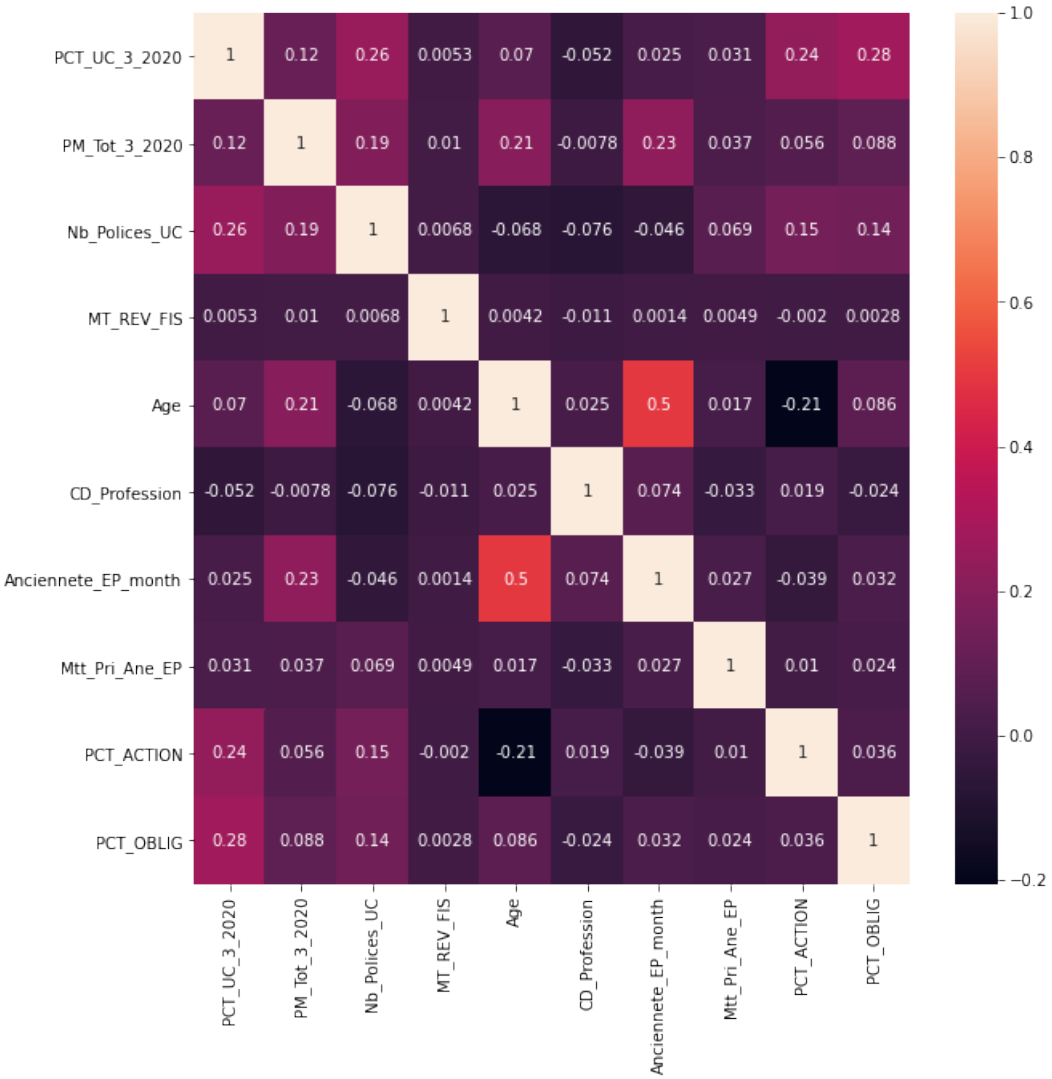

FIGURE 41 – Tableau de corrélations entre les variables utilisées dans le modèle

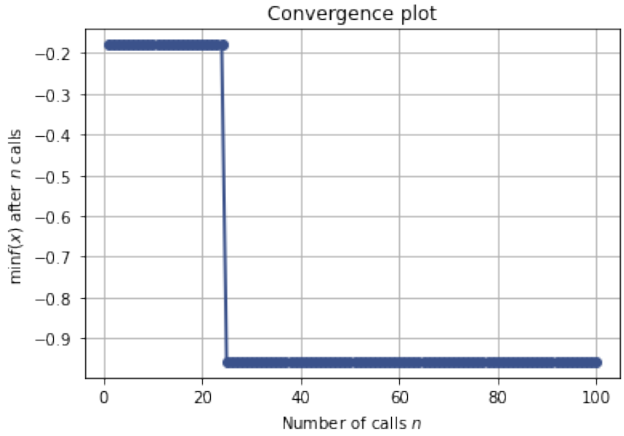

FIGURE 42 – Amélioration de la fonction de minimisation de l'hyperparamètrage bayésien au fil des itérations

Importance des variables pour la modélisation selon Total Gain :

Variable	Total Gain
Pourcentage d'UC au 31/03/2020	13962913,6
PM Totale au 31/03/2020	457254,0
Pourcentage d'actions parmi les UC	156806,4
Revenu Fiscal	83925,5
Ancienneté	77607,3
Âge	65126,6
Nombre de polices multi-support	60316,5
Montant des versements programmés	46331,2
Pourcentage d'obligations parmi les UC	40146,3
Code Profession	11991,3

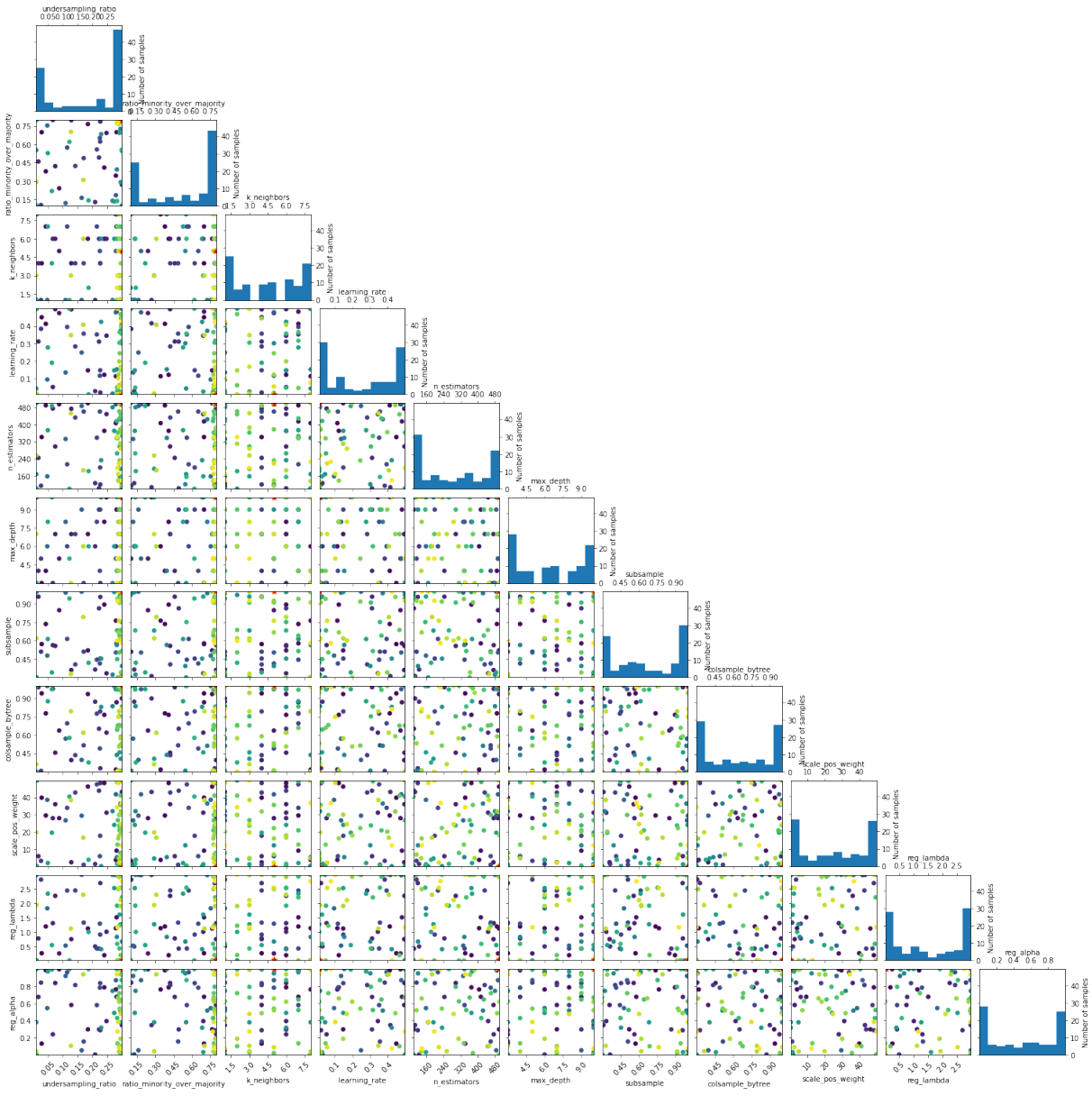

FIGURE 43 – Zone de recherche par hyperparamétrage au fil des itérations

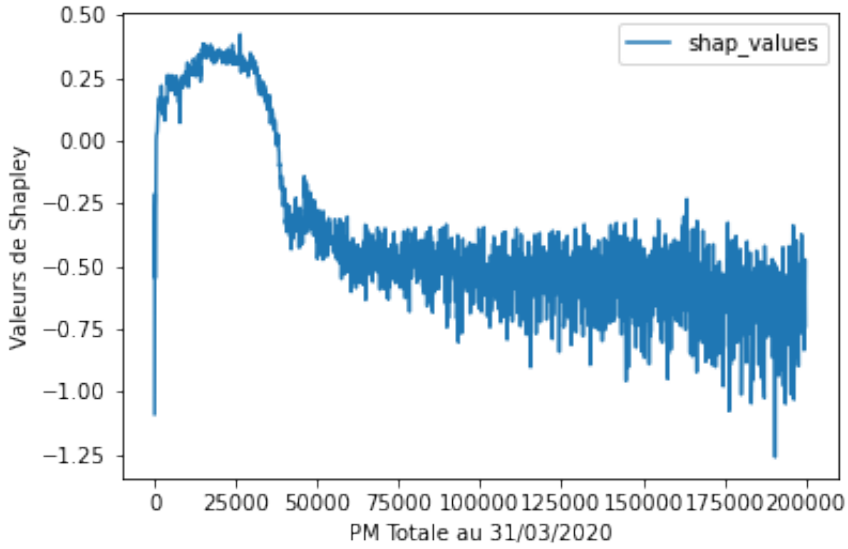

FIGURE 44 – Valeurs de Shapley pour la PM Totale au 31/03/2020

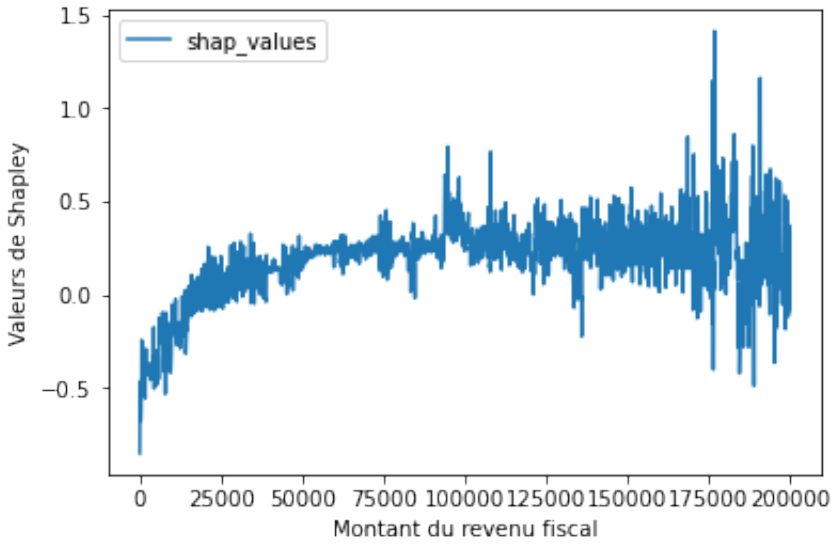

FIGURE 45 – Valeurs de Shapley pour le montant du revenu fiscal

9 Bibliographie

Références

[1] AGENCE FRANCE TRÉSOR. *LE PROGRAMME DE FINANCEMENT 2021 S'INSCRIT DANS LA CONTINUITÉ DE L'ANNÉE 2020, MARQUÉE PAR UNE FORTE HAUSSE DES BESOINS DE FINANCEMENT FACE À LA CRISE DE LA COVID-19*. 367. Déc. 2020. URL : https://www.aft.gouv.fr/files/medias-aft/7_Publications/7.2_BM/367_Bulletin%5C%20mensuel%5C%20d%5C%C3%5C%A9cembre%5C%202020.pdf.

[2] ARIAS et al. *Les placements des assureurs résistent à la crise malgré les tensions*. 232. Déc. 2020. URL : https://publications.banque-france.fr/les-placements-des-assureurs-resistent-la-crise-malgre-les-tensions.

[3] Guillaume BENOIT. "La France s'est financée à taux négatif en 2020". In : *Les Echos* (1er juill. 2021). URL : https://www.lesechos.fr/finance-marches/marches-financiers/la-france-sest-financee-a-taux-negatif-en-2020-1272515.

[4] BERGSTRA JAMES et al. "Algorithms for hyper-parameter optimization". In : *NIPS'11 : Proceedings of the 24th International Conference on Neural Information Processing Systems* (2011), p. 2546-2554. URL : https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

[5] CARLOS GUESTRIN et TIANQI CHEN. "XGBoost : A Scalable Tree Boosting System". In : Conference : the 22nd ACM SIGKDD International Conference. Gand, Belgique : ACM SIGKDD, août 2016.

[6] N. V. CHAWLA et al. "SMOTE : Synthetic Minority Over-sampling Technique". In : *Journal of Artificial Intelligence Research* 16 (2002), p. 321-357. URL : https://www.researchgate.net/publication/220543125_SMOTE_Synthetic_Minority_Over-sampling_Technique.

[7] DREES, DIRECTION DE LA RECHERCHE, DES ÉTUDES, DE L'ÉVALUATION ET DES STATISTIQUES. *Le niveau de vie des retraités*. 2021. URL : https://drees.solidarites-sante.gouv.fr/sites/default/files/2021-05/Fiche%5C%2009%5C%20-%5C%20Le%5C%20niveau%5C%20de%5C%20vie%5C%20des%5C%20retrait%5C%C3%5C%A9s.pdf.

[8] FFA. *Les chiffres de l'assurance en 2020*. 24 mars 2021. URL : https://www.ffa-assurance.fr/etudes-et-chiffres-cles/les-chiffres-de-assurance-en-2020.

[9] GARY M. WEISS, KATE MCCARTHY et BIBI ZABAR. "Cost-Sensitive Learning vs. Sampling : Which is Best for Handling Unbalanced Classes with Unequal Error Costs ?" In : Conference : Proceedings of the 2007 International Conference on Data Mining. DBLP, juin 2007, p. 35-41.

[10] INSEE. *Niveau de vie moyen par décile | Insee*. URL : https://www.insee.fr/fr/statistiques/2417897#tableau-figure1 (visité le 02/11/2021).

[11] INSEE. *Structure et distribution des revenus, inégalité des niveaux de vie en 2017 | Insee*. 23 jan. 2020. URL : https://www.insee.fr/fr/statistiques/4291712 (visité le 02/11/2021).

[12] INSEE, INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES. *Professions et catégories socioprofessionnelles PCS 2003 | Insee*. URL : https://www.insee.fr/fr/information/2400059 (visité le 02/11/2021).

[13] *Je suis bénéficiaire d'une assurance-vie, comment sont imposées les primes ?* 5 août 2020. URL : https://www.impots.gouv.fr/portail/international-particulier/questions/je-suis-beneficiaire-dune-assurance-vie-comment-sont-imposees (visité le 29/11/2021).

[14] Divakar KAPIL. *Hyperparameter Search : Bayesian Optimization - Analytics Vidhya*. Anglais. 15 nov. 2019. URL : https://medium.com/analytics-vidhya/hyperparameter-search-bayesian-optimization-14be6fbb0e09 (visité le 08/11/2021).

[15] *L'assurance-vie et le PEA*. 29 avr. 2021. URL : https://www.impots.gouv.fr/portail/particulier/lassurance-vie-et-le-pea-0 (visité le 25/11/2021).

[16] MARC REHMSMEIER et TAKAYA SAITO. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets”. In : *PLoS One* (2015). URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800/.

[17] SAMI VIRPIOJA. *BIRCH : Balanced Iterative Reducing and Clustering using Hierarchies*. 23 avr. 2008. URL : https://pdfs.semanticscholar.org/798b/34f582f7e9a78c2e18f08b4e3a55b0c84a62.pdf.

[18] SCOTT M. LUNDBERG, SU-IN LEE et GABRIEL G. ERION. “Consistent Individualized Feature Attribution for Tree Ensembles”. In : *ArXiv* (2018). URL : https://www.semanticscholar.org/paper/Consistent-Individualized-Feature-Attribution-for-Lundberg-Erion/861aaf3e9c8af9e23f1990d20815f7602d6646

[19] SCOTT M.LUNDBERG et SU-IN LEE. “A Unified Approach to Interpreting Model Predictions”. In : *NIPS* (2017). URL : https://www.researchgate.net/publication/317062430_A_Unified_Approach_to_Interpreting_Model_Predictions.

[20] L. S. SHAPLEY. “17. A Value for n-Person Games”. In : *Contributions to the Theory of Games (AM-28), Volume II* (1953), p. 307-318. DOI : 10.1515/9781400881970-018.

[21] Tian ZHANG, Raghu RAMAKRISHNAN et Miron LIVNY. “BIRCH”. In : *ACM SIGMOD Record* 25.2 (1996), p. 103-114. URL : https://dl.acm.org/doi/10.1145/235968.233324.