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

le 7 Septembre 2022

Par : Ahmed Amine IBN HADJ FARHAT
Titre : Modélisation de l'évolution d'un portefeuille en assurance santé individuelle

Confidentialité : Oui - (Durée: 2 ans)

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

Membre présent du jury de l'Institut des Actuaires :
Julien BOUDOT
Mathieu LIMOUZIN
Signatures :

Entreprise :
Garcia Rochette & Associés
Signature :

Membres présents du jury de l'EURIA :
Jean-Marc DERRIEN

Directeur de mémoire en entreprise :
Pierre COCHAIN
Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

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

Résumé

Entre la mise en place du 100% santé, la crise sanitaire et la mise en œuvre de la résiliation infra-annuelle des contrats santé, les complémentaires santé ont de plus en plus de mal à estimer leurs résultats avec les méthodes de tarification usuellement employées dans un milieu qui subit différentes évolutions. Dans ce contexte, ce mémoire vise à concevoir un modèle dynamique permettant de prédire l'évolution d'un portefeuille en assurance santé individuelle, en terme d'effectifs et de consommation, à l'horizon d'un an à partir des données historiques du portefeuille et de paramètres en entrée tels que la variation des cotisations, l'arrivée d'une vague épidémique ou bien la transformation du niveau des garanties. Pour ce faire, nous allons estimer pour chaque assuré du portefeuille la probabilité de résiliation à horizon un an, la durée de couverture restante dans le cas d'une résiliation probable et la consommation probable pendant la durée de couverture restante (un an dans le cas de non résiliation). Un tel modèle dynamique, prenant en compte les événements qu'à connu dernièrement l'assurance santé, permettra aux assureurs non seulement d'optimiser leurs résultats en fonction de l'indexation tarifaire mais aussi de prévoir l'évolution de la composition de leurs portefeuilles en terme de profils de risque.

Mots clefs: Assurance santé, Résiliation infra-annuelle, 100% santé, Élasticité-prix, Apprentissage automatique, XGBoost, GLMs, SHAP.

Abstract

Between the establishment of the 100% health insurance scheme, the health crisis and the implementation of the infra-annual cancellation of health contracts, complementary health insurance companies are finding it increasingly difficult to estimate their profits using the pricing methods usually employed in an environment that is undergoing various changes. In this context, this thesis aims at designing a dynamic model that allows to predict the evolution of an individual health insurance portfolio, in terms of membership and consumption, over a one-year horizon on the basis of historical portfolio data and input parameters such as the variation of contributions, the arrival of an epidemic wave or the transformation of the level of coverage. To do this, we will estimate for each insured in the portfolio the probability of cancellation at the one-year horizon, the remaining duration of coverage in the case of probable cancellation and the probable consumption during the remaining duration of coverage (one year in the case of non-cancellation). Such a dynamic model, taking into account recent events in the health insurance industry, will allow insurers not only to optimise their results in function of tariff indexation but also to predict the evolution of the composition of their portfolios in terms of risk profiles.

Keywords: Health insurance, Infra-annual cancellation, 100% health insurance scheme, Price-elasticity, Machine learning, XGBoost, GLMs, SHAP.

Synthesis Note

In recent years, health insurance has undergone several regulatory changes that have directly impacted the activities of supplementary health insurance providers. Among these changes, we mainly find the introduction of the 100% Health initiative and the enactment of infra-annual cancellation.

With a portfolio in individual health insurance, we aim to model the evolution of the outgoing contract flows from the portfolio as well as the subsequent consumption of beneficiaries linked to these contracts. This evolution must take into account the impact of recent regulations.

Since the implementation of infra-annual cancellation, the total exposure of the portfolio over a year is no longer maintained, a large part of the portfolio related to contracts with less than one year of seniority may cancel during the year, presenting a risk of non-coverage of the insurer's expenses.

The seasonality of reimbursements presents, among other things, a risk of underestimating pure premiums since the enactment of infra-annual cancellation. Indeed, let's start by looking at the monthly reimbursements of the available portfolio.

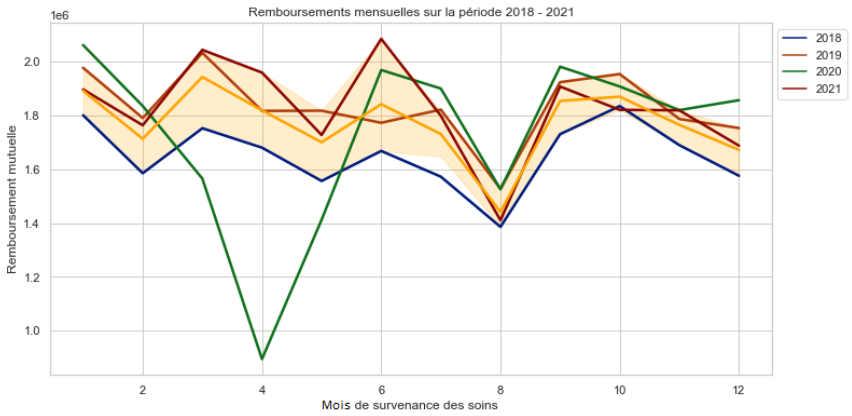


FIGURE 1 – Mutual reimbursements per month of occurrence by exercise

With infra-annual cancellation, seasonality can have a direct impact on the estimation of the insured's pure premium.

Often calculated over the whole year, the estimation of the pure premium can vary for a calculation over a shorter period of the year.

To estimate this variation, we place ourselves at the beginning of each exercise and calculate the annual average reimbursements per beneficiary (pure premiums) based on a coverage period from January 1 to a date t that evolves between February 1 and the end of the exercise.
For an exercise N and t the number of days difference between 1/1/N and the j/m/N, the annual pure premium based on the period from 1/1/N to j/m/N is written as :

$$Pure\,premium(t) = \frac{\sum_{s=0}^{t} Mutual\,reimbursements(s)}{\sum_{s=0}^{t} Total\,exposure(s)}$$

With,

— Mutual reimbursements(s) : Sum of mutual reimbursements referring to the day of occurrence s.

— Total exposure(s) : Sum of the exposure of beneficiaries on day s (in years).

The representation of these pure premiums for the years 2018, 2019 and 2021 is as follows :

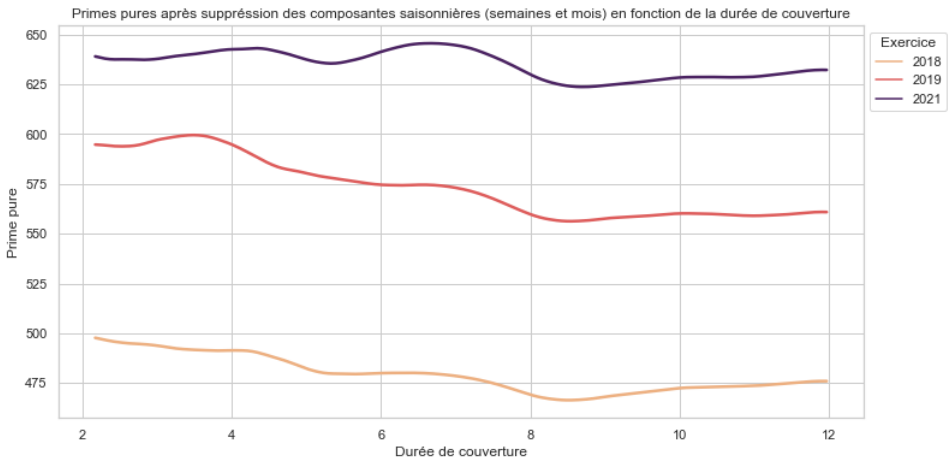


FIGURE 2 – Smoothed pure premiums according to the coverage duration of the year

The annual pure premiums follow a similar seasonality for these three exercises.
By shifting the end of the curves to zero (by vertical translation) and averaging over the three exercises, we obtain the average variation of the annual pure premium as a function of the coverage duration compared to a full year's coverage :

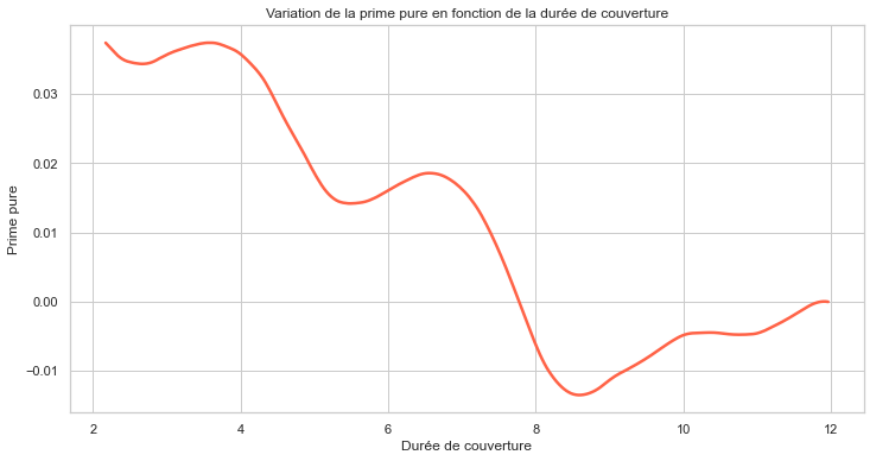
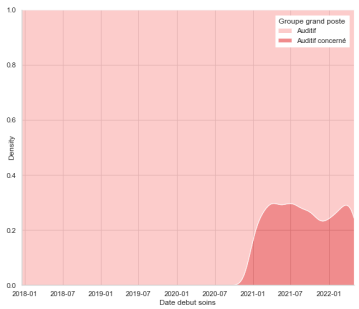


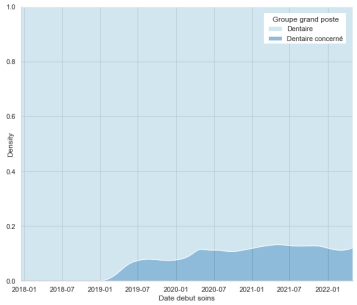
FIGURE 3 – Variation of the annual pure premium according to the coverage duration

Assuming we have estimated the pure premiums based on the whole year of coverage. For an insured person who cancels in April, their pure premium would therefore be underestimated by more than 3% on average. Similarly, an insured person who cancels in September, their pure premium is overestimated on average by 1%.

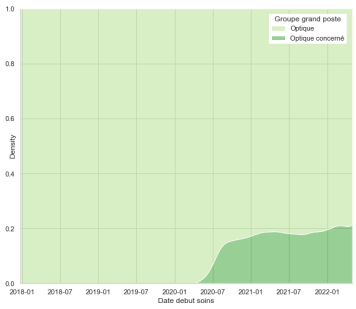
In addition to seasonality impacting pure premiums, the implementation of the 100% Health coverage could impact the consumption of certain acts by the insured.
Indeed, acts with zero remaining charge are increasingly requested by beneficiaries. To look at the impact of the implementation of the 100% Health coverage on the consumption of acts concerned by the regulation, we began by looking at the evolution of the proportion of act codes covered by the 100% Health coverage over time for each major concerned post.



(a) Proportion of acts concerned by the 100% Health coverage in Audiology



(b) Proportion of acts concerned by the 100% Health coverage in Dental



(c) Proportion of acts concerned by the 100% Health coverage in Optics

The superimposition of these distributions for the three posts is as follows :

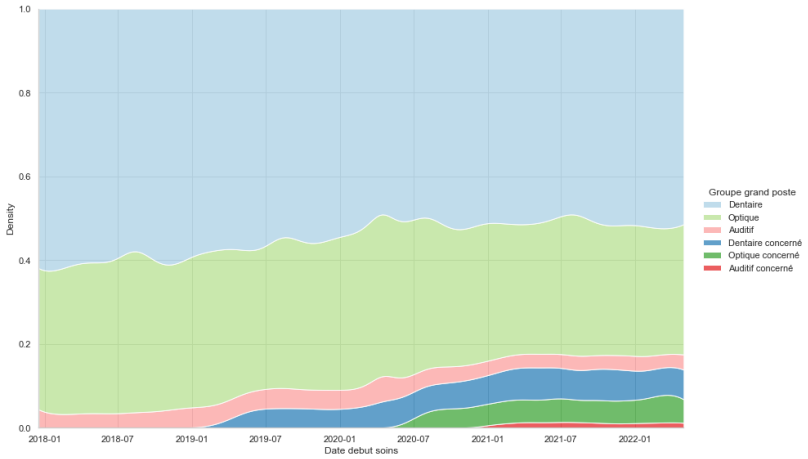


FIGURE 5 – Evolution of the consumption of acts covered by the 100% Health coverage over time

The three distributions in the lower part of the graph represent the proportion (in number of acts) of acts concerned by zero remaining charge for the three posts concerned compared to the rest of the acts.

The proportion of acts concerned remained increasing until the full implementation of 100% Health baskets. Subsequently, policyholders have more recourse to types of acts covered by the 100% Health coverage after the regulation came into force. These acts now represent 15% of the acts in the portfolio.

The evolution of the average cost of an act per month depending on whether it is concerned by the regulation for the last 3 posts is as follows :

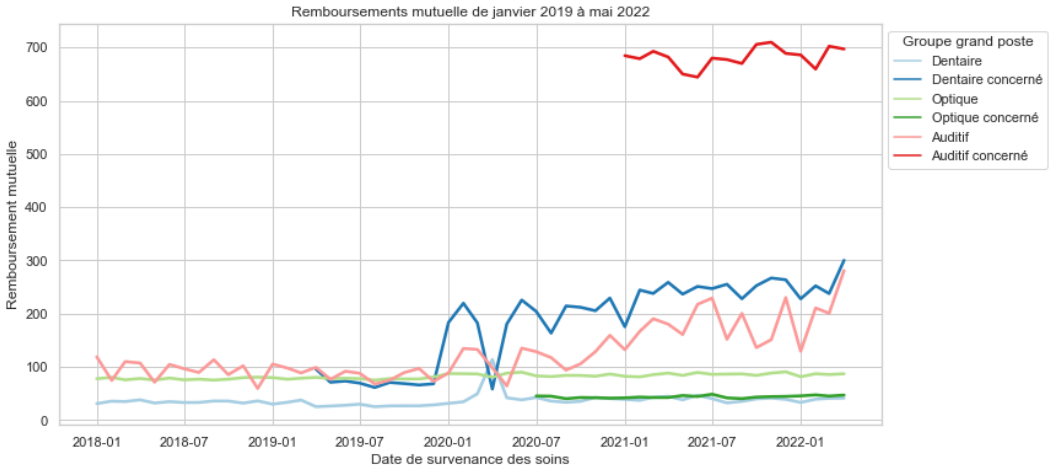


FIGURE 6 – Evolution of the average cost of acts concerned by 100% Health sheme per month

the average cost of acts concerned by the 100% Health coverage is always higher than the rest of the acts for the audiology and dental posts.

Finally, the increase in the proportion of acts covered by the 100% Health coverage and their higher average costs than the rest of the acts will lead to an increase in consumption in terms of average cost and frequency.

In this context, modeling the evolution of the portfolio required the establishment of a two-step process as follows :

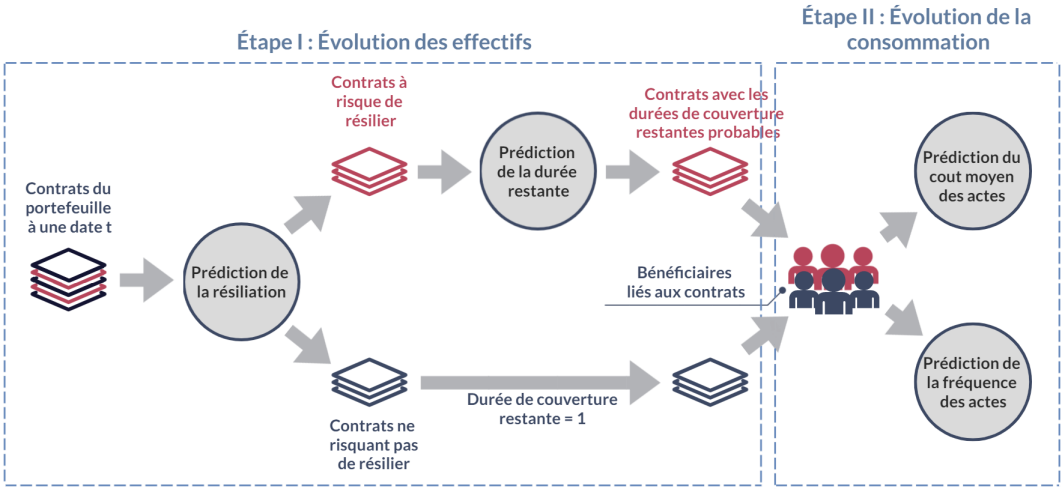


FIGURE 7 – Modeling process diagram

The first step concerns the evolution of the portfolio's headcount. Observing the portfolio at a date t, we want to know if an insured person is likely to cancel in the following year and their remaining coverage duration in the event of a probable cancellation. In this part, predictions are made by contract since the cancellation of an insured person mostly concerns all the beneficiaries of the same contract.

Once the remaining coverage duration, at a one-year horizon, is known for each contract, we move on to data related to each beneficiary to predict their consumption during their remaining coverage period by predicting the average cost of their acts and their consumption frequency. This model allows the detection of a behavioral change related to a probable cancellation.

The prediction of the variables to be explained, at a given observation date of the portfolio, requires linking the events prior to this date to the values taken by these variables over a subsequent period.

The construction of databases is therefore carried out by concatenating data from a succession of monthly observations of the portfolio over the studied activity period. Observations are made on the first of each month. The explanatory data, respectively to be explained, concern a period of one year before, respectively one year after, the

observation date, as shown in the graph below :

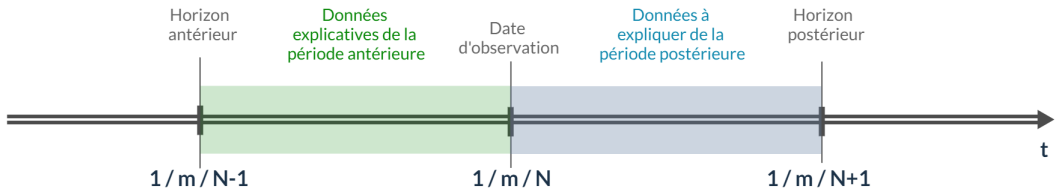
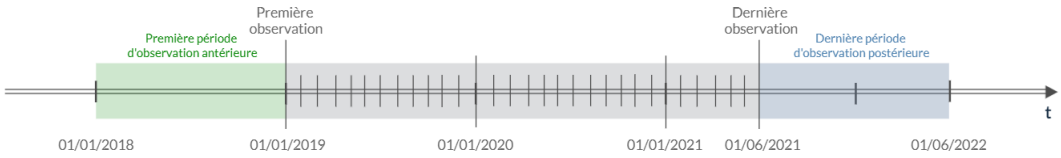


FIGURE 8 – Portfolio observation periods

The data transmitted by the mutual corresponds to the period from **January 1, 2018** to **May 6, 2022**. Given the constraint on the previous and subsequent horizons, the monthly observations of the portfolio will be from **January 1, 2019** to **May 1, 2021**.



The variables describing the prior activity of an insured person relative to a given observation date are broken down as follows :

— Variables related to the **profiles of insured persons** : Age, gender, type, postal code, etc.
— Variables related to the **contracts** : Range, guarantee, contributions, taxes, etc.
— Variables related to **consumption** : Mutual reimbursement by act post, number of acts, etc. (summed over a year prior to the observation date.)

In order to provide the models with all the necessary information for predictions, it was necessary to create some additional variables from the available data. Among these variables, we find the previous and subsequent indexations of tariffs, and the duration before the eligibility of contracts to cancel.

For an observation date on 1/m/N, the indexing variables are as follows :

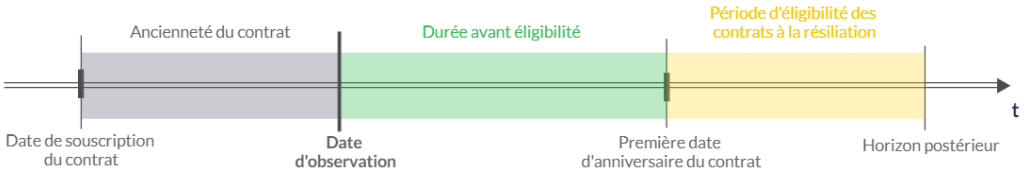
$$\textbf{Previous indexing} = \frac{Annualcontribution(N) - Annualcontribution(N-1)}{Annualcontribution(N-1)}$$

$$\textbf{Subsequent indexing} = \frac{Annualcontribution(N+1) - Annualcontribution(N)}{Annualcontribution(N)}$$

The variable **Subsequent indexing** allows for a certain dynamism of the models. Indeed, predictions can be oriented according to the subsequent indexing, through simulations, to analyze the impact of the indexing value on the probable outgoing contract flow.

In order to provide the models with information on the eligibility of members to cancel, we create a variable called "Duration before eligibility" which gives the duration before which the member can cancel his contract according to the observation date.
This variable is calculated in years and depends on the observation date. For example, the calculation of this variable after the entry into force of infra-annual cancellation is as follows :

For a given observation date, this variable is worth 0 for contracts with seniority greater than 1 year (thus eligible to cancel at any time).
In the case of seniority less than 1 year at the observation date, the calculation of this variable is as follows :



$$\begin{cases} DAE = 0 & if\ AC \geqslant 1 \\ DAE = 1 - (AC - \lfloor AC \rfloor) & if\ AC < 1 \end{cases}$$

With,

DAE : Duration before eligibility
AC : Contract seniority

$\lfloor AC \rfloor$ denotes the lower whole part of the contract's seniority.

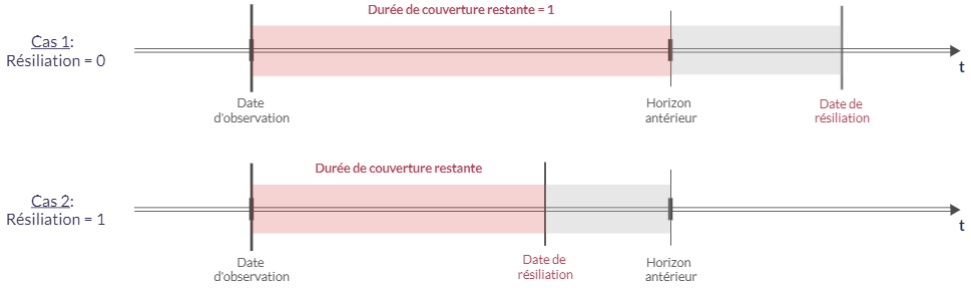
On the side of the variables to be explained, the variables to be predicted for stage I of the global model are cancellation and remaining coverage duration. These variables are assigned to the bases by contract and defined as follows :

$$\textbf{Cancellation} = 1_{\{Cancellationdate < Horizondate\}}$$

$$\textbf{Remaining coverage duration} = \begin{cases} 1 & if \quad Cancellation = 0 \\ DR - DO & if \quad Cancellation = 1 \end{cases}$$

With,

DR : Cancellation date
DO : Observation date



The prediction of posterior consumption in part II is done by beneficiary, the variables to be predicted for a given act post i are :

— **Posterior average cost (i)** = Posterior reimbursements (i) / Number of posterior acts (i)
— **Posterior frequency (i)** = Number of posterior acts (i) / Remaining coverage duration

These variables are calculated over the coverage period following observation and with a one-year horizon.

Finally, in order to exploit all available information, observation dates after May 1, 2021, were considered for cases of termination before May 6, 2022. Indeed, in the case of a termination, the horizon is limited to the date of termination.

When comparing the distributions of the explanatory variables with the subsequent contract terminations, we find a significant difference for the following variables :

Ages of members

The distribution of the members' ages according to the contract termination at a one-year horizon is represented as follows :

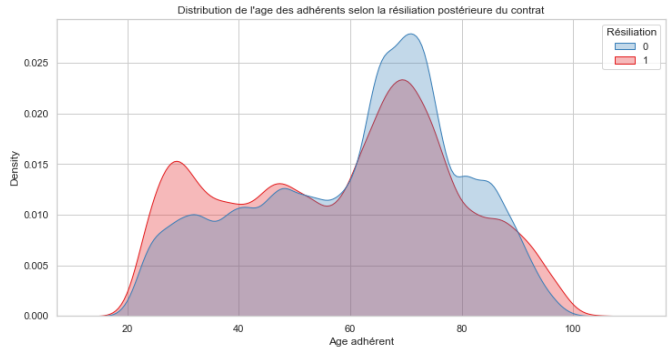


FIGURE 9 – Distribution of members' ages according to contract termination

We note that terminations are more represented for contracts whose members are under 50 years old. This gap is especially present for members under 35 years old. On the other hand, termination is also better represented for contracts of the oldest members, these terminations are probably due to more frequent deaths in this age group.

Time before eligibility

The distribution of this variable, over the entire coverage duration of the portfolio, according to contract terminations is represented as follows :

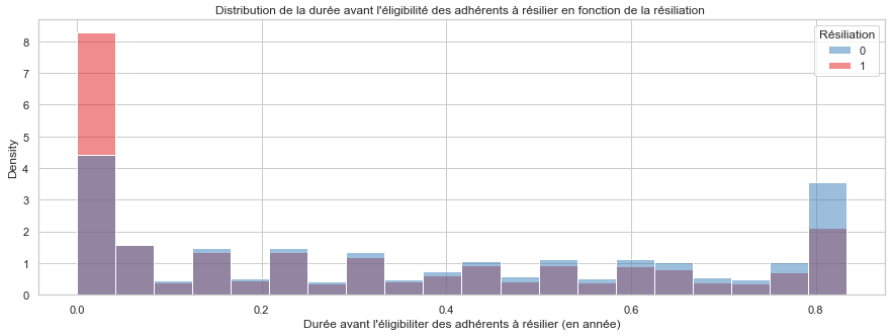
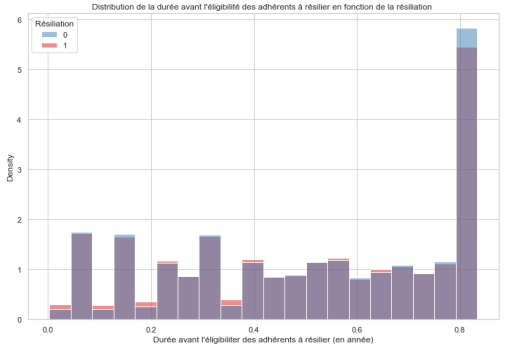
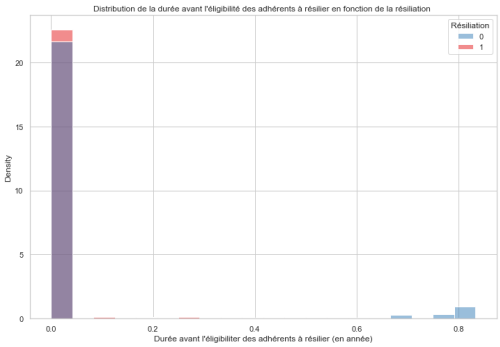


FIGURE 10 – Distribution of time before eligibility according to contract termination

This duration experienced a before and after the entry into force of the intra-annual termination at the end of 2020. The distributions of this variable related to observations from 2019 and 2021 are presented as follows :



(a) Distribution of time before eligibility in 2019

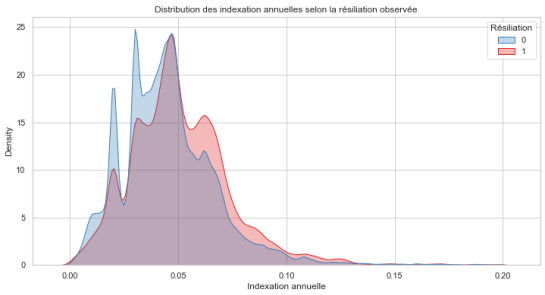


(b) Distribution of time before eligibility in 2021

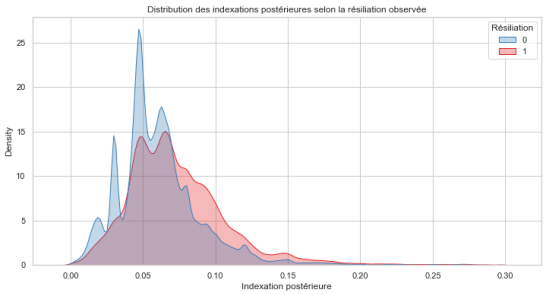
Unlike the year 2019, the majority of contracts are eligible for termination in 2021 thanks to the new regulation.

Previous and subsequent indexation

The distributions of previous and subsequent indexation according to termination are presented as follows :



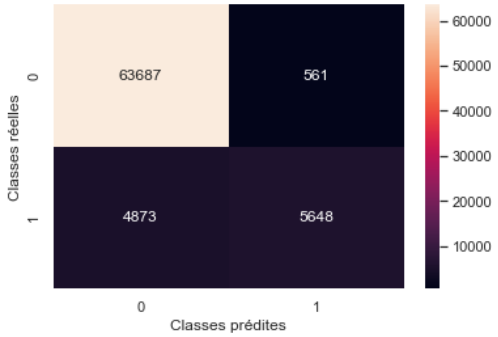
(a) Distribution of previous indexation



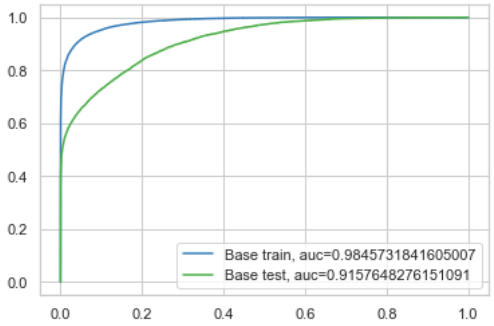
(b) Distribution of subsequent indexation

High-level indexations, whether previous or subsequent, favor the risk of contract termination according to these two graphs.

For the prediction of terminations, we opted for an XGBoost model. To improve the performance of the model, the contracts used for training are related to terminations after the entry into force of the intra-annual termination, and the results of the classification obtained are as follows :



(a) Confusion matrix for the test base



(b) ROC Curve

	precision	recall	F1-score	Support
Non-termination	0.93	0.99	0.96	64,248
Termination	0.91	0.54	0.68	10,521

The sensitivity of the model reached 54% with a precision of 91%. Considering the contracts present at the entry into force of the intra-annual termination allowed us to detect more terminations with better precision.

The interpretation of the contribution of variables to the predictions of the model by the SHAP method is presented as follows :

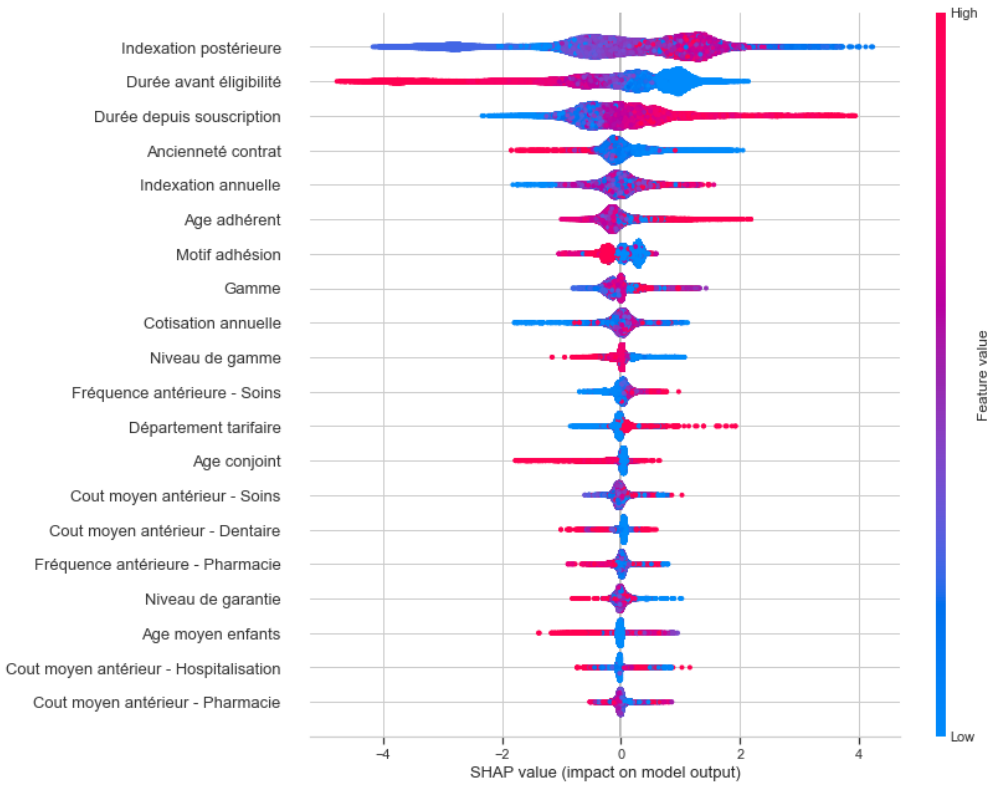


FIGURE 14

According to the Shapley values, the most influential variable on prediction is **subsequent indexation**. The representation shows that the majority of high subsequent indexations have increased the probability of termination for the contracts concerned.

Time before eligibility represents the second most influential variable on the predictions of the model. We see that a high time before eligibility decreases the probability of contract termination. Indeed, the greater the value this variable takes, the shorter the period during which a contract can be terminated in the subsequent horizon. Similarly, low values of this variable representing contracts that are eligible for termination throughout the year are represented with positive Shapley values that increase the probability of termination.

The third variable is the **duration since subscription**. We can see that this variable increases according to the value of Shapley. Indeed, as we saw in the preceding chapter, the majority of contracts change their range or level of guarantee after 5 years, this variable refers to the date of subscription to the current range and level of guarantee. Thus, a high value of this duration increases the probability of termination.

On the other hand, the **seniority of contracts** is decreasing according to the value of Shapley. This reflects the loyalty of older members.

To interpret the influence of the **member's age** on contract termination, we trace the Shapley value according to the member's age (taking into account the presence of a spouse or not in the contract for a better interpretation) :

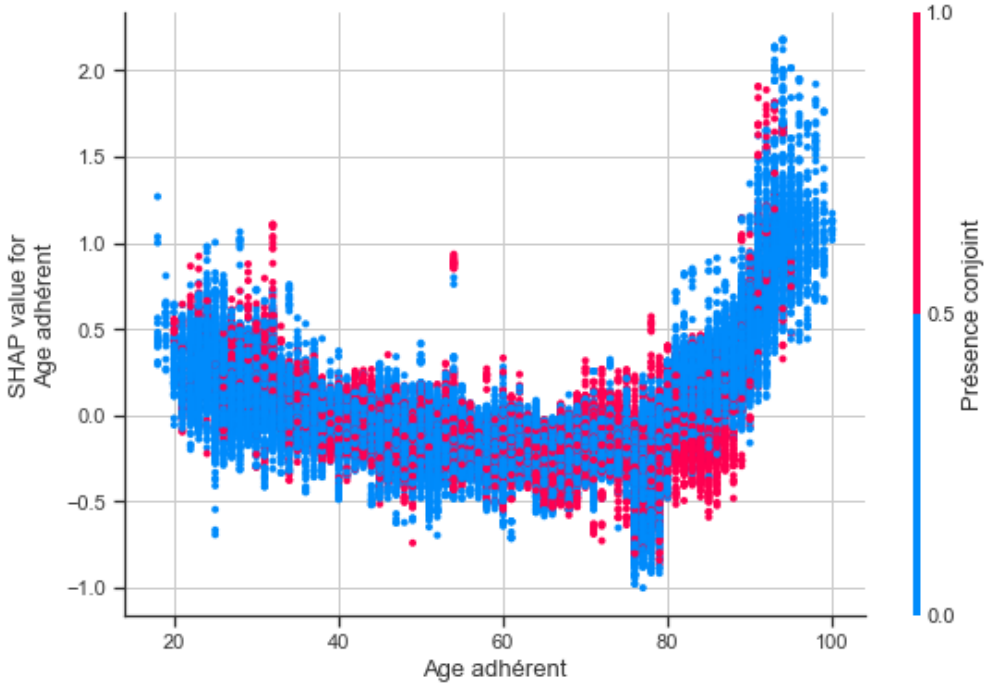


FIGURE 15 – Shapley value according to members' ages

As we noted when comparing the distribution of members' ages according to the termination variable, terminations concern especially the younger members under 40 years and the older ones (probably due to deaths). The Shapley values confirm this through this representation, with Shapley values decreasing for the younger and increasing for the older. In particular, we find that the presence of a spouse for members over 80 years old lowers the probability of termination.

In terms of the remaining coverage duration, we will compare its distribution for two observation months according to the observation year.

The distribution of the remaining coverage duration for an observation in the month of March is as follows :

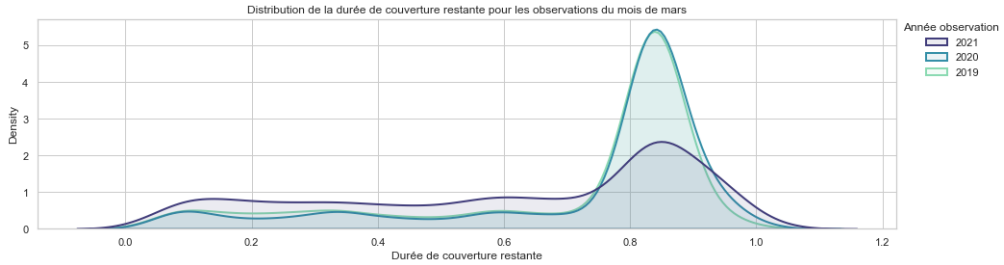


FIGURE 16 – Distribution of remaining coverage duration for March observations

The observed peaks on the curve are related to the month of January. The distribution of the remaining coverage duration is almost identical for the observations of 2019 and 2020. Indeed, the intra-annual cancellation does not come into effect until December 2020, so there is not a big difference in terms of eligibility to cancel for an observation in March.

However, the distribution is better distributed in 2021 thanks to the intra-annual cancellation. The cancellations observed between March and the following January (duration between 0 and 8.33) have almost doubled compared to the previous two years.

The distribution of the remaining coverage duration for an observation in the month of October is as follows :

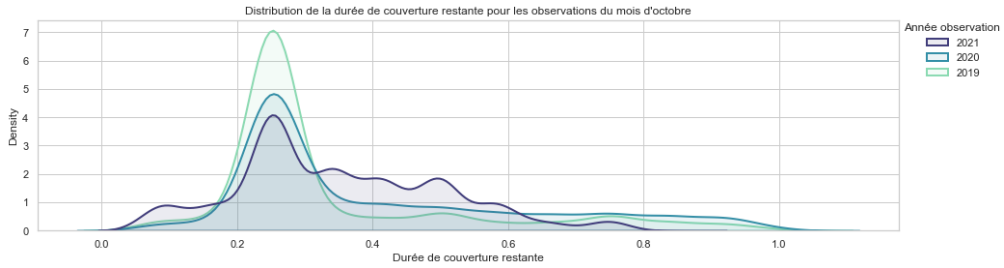
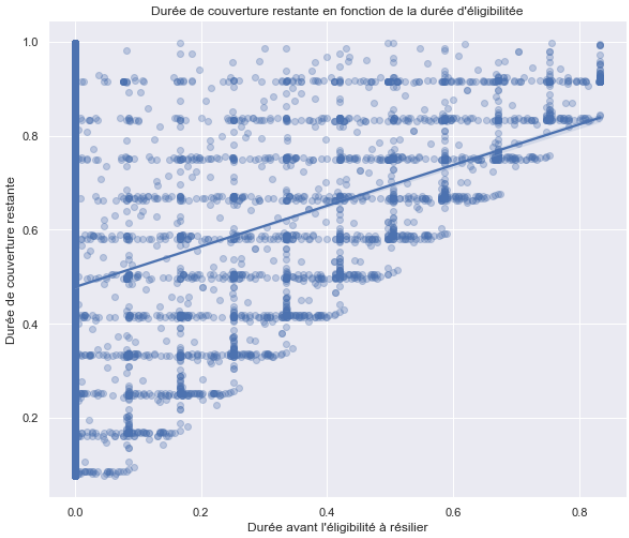


FIGURE 17 – Distribution of remaining coverage duration for October observations

For the observations of 2020, the impact of the intra-annual cancellation is visible this time. Indeed, the regulation came into force in December, only two months later. Subsequently, cancellations are less concentrated in January and more distributed over the following period.

In 2021, the distribution only takes into account cancellations dated before May 6, 2022 (our posterior data horizon). This distribution is more spread out over the entire year following the observation.

The remaining coverage duration is framed by the time before eligibility. Indeed, contracts can only cancel after the time before which they will be eligible to cancel has passed. The remaining coverage duration according to the time before eligibility is shown below :



The remaining coverage duration is always greater than the time before eligibility. This variable will therefore be of great importance for prediction.

The XGBoost regression model used is based solely on cases of cancellation, the number of contracts canceled after 2021 is 6,523. Following the training of the model, the test results obtained are as follows :

Training Score	MAE	RMSE
0.78	0.16	0.21

For the second part of the model, modeling the evolution of the portfolio's subsequent consumption is done by predicting the average cost and frequency for each beneficiary during their remaining coverage period and for each major type of act.

The distribution of posterior and anterior average costs and frequencies to observation dates by type of act is presented as follows :

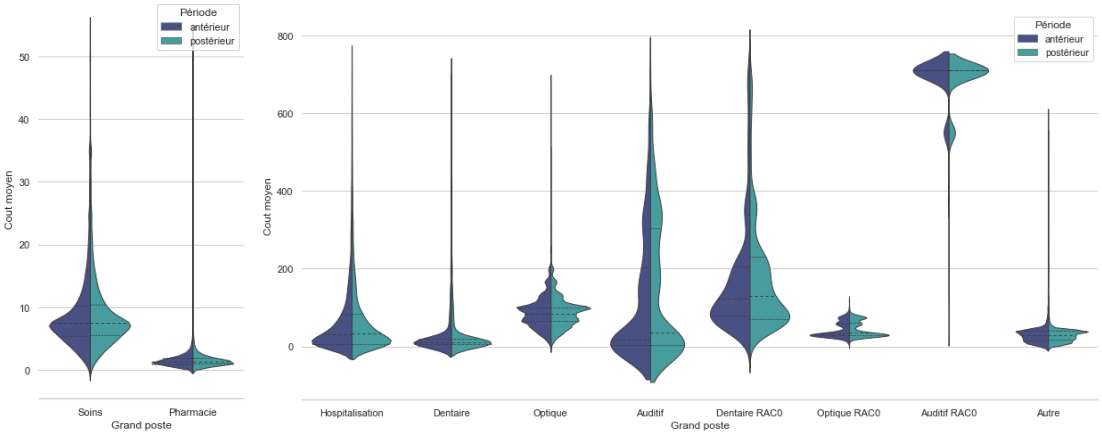


FIGURE 18 – Posterior and anterior average costs by type of act

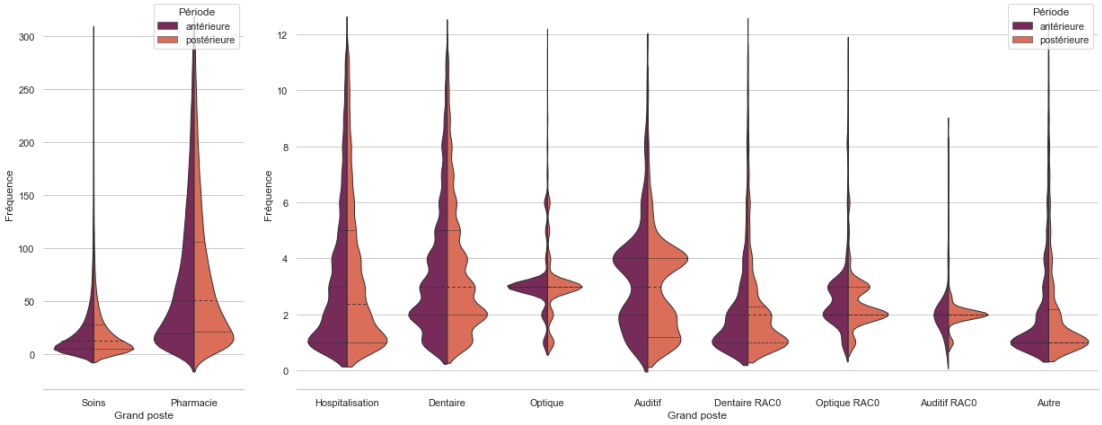


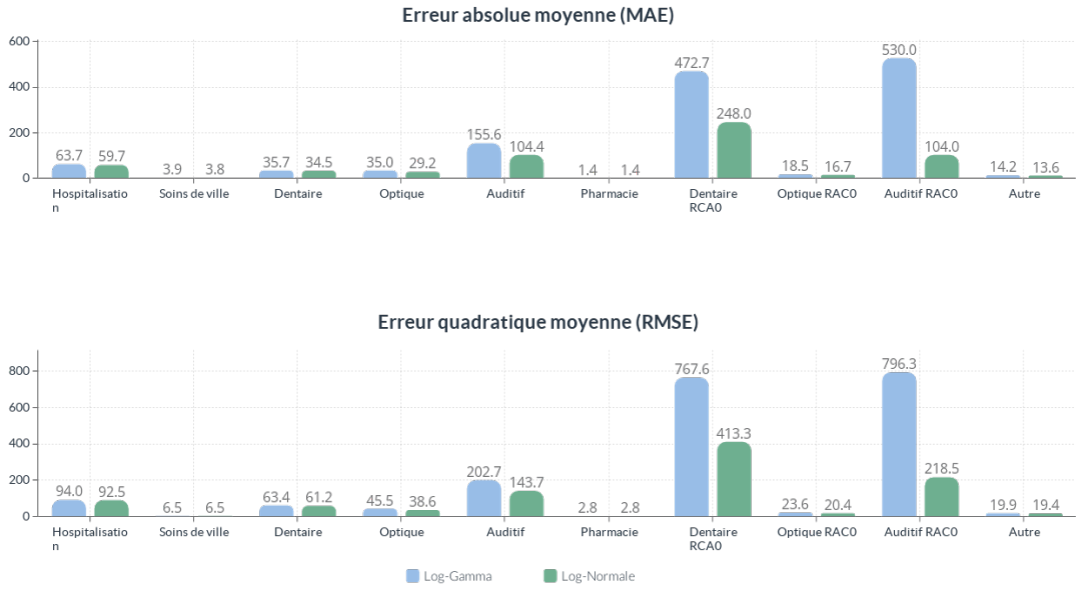
FIGURE 19 – Posterior and anterior frequencies by type of act

To avoid double counting of average costs and frequencies anterior and posterior due to successive observations and to take into account 100% health sheme on all concerned posts, this graph is from a single observation in April 2021.

The distributions of the posterior average costs and frequencies to this observation date and by type of act remain very close to those observed in the year prior to the observation date.

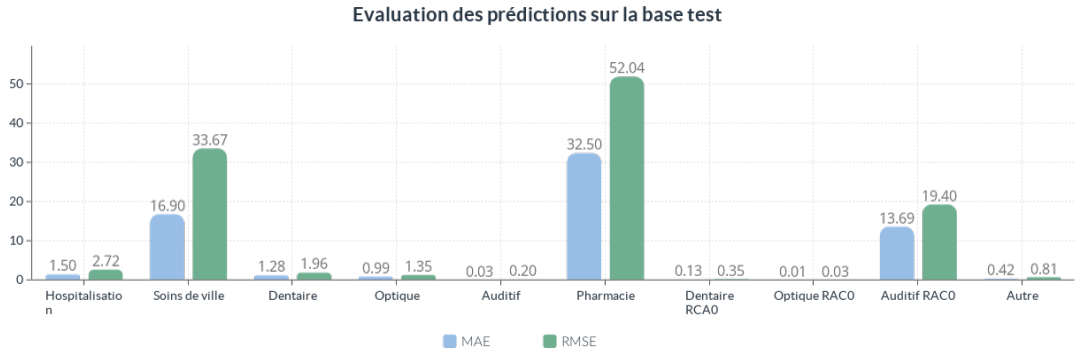
For the prediction of posterior average cost for each type of act, we will choose between a GLM-LogGamma and a GLM-LogNormal model depending on their performances for each type of act.

By type of act, the results of the models obtained are presented as follows :



Across all the obtained models, the Log-Normal models present better results. Indeed, the RMSE and MAE values for this model are always lower than those of the Log-Gamma models. The Log-Normal model is therefore better suited to our distributions. Thus, we retain this model for predicting the posterior average costs of acts.

For the prediction of frequency, we use a GLM Poisson. By type of act, the results of the models obtained are presented as follows :



The precision of the models obtained for each type of act is acceptable. The obtained models seem efficient.

Finally, we proceed with the implementation of the contract cancellation model in a case application on a portion of contracts withdrawn from the portfolio at the beginning of the modeling. This portion concerns 5,965 contracts all present on May 1, 2021.

The case application consists of predicting the evolution of the number of contracts present at a one-year horizon from an observation made on May 1, 2021, for the contracts in question. The evolution of the age pyramid through the implementation of the model is as follows :

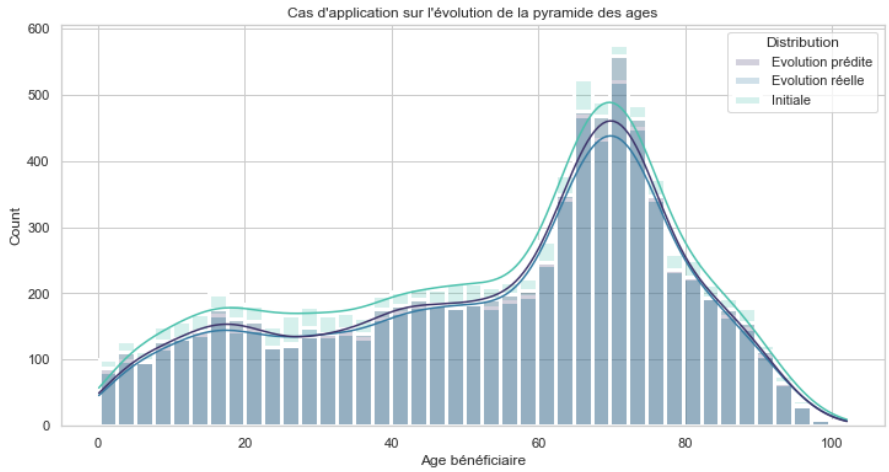


FIGURE 20 – Prediction of the evolution of the age pyramid

The model's prediction of the evolution of the age pyramid is very close to reality. However, it remains slightly higher than the actual observation of the pyramid after one year, so the model has predicted fewer cancellations for certain ages.

To get an idea of the impact of the evolution of the composition of the workforce on the portfolio's consumption, we plot the obtained age distributions.

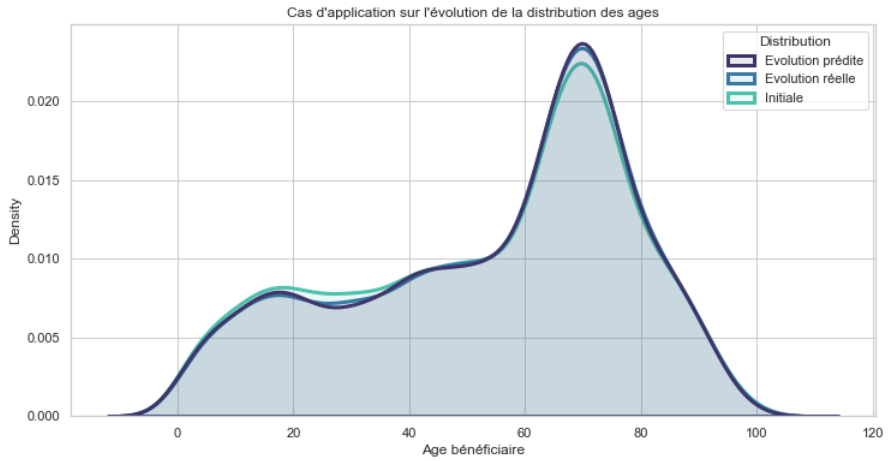


FIGURE 21 – Prediction of the evolution of age distribution

The obtained age distribution is very close to reality. The XGBoost has particularly managed to detect cancellations concerning the young aged 20 to 40 years, which allowed predicting the trough observed in this age range.

This evolution of distribution allows for a better estimate of the portfolio's consumption as the age of beneficiaries is the main indicator of their consumption.

The application of the models for predicting average costs and frequencies by type of act allows us to obtain the pure premiums for the year following the observation, which we plot according to the age of the insured as follows :

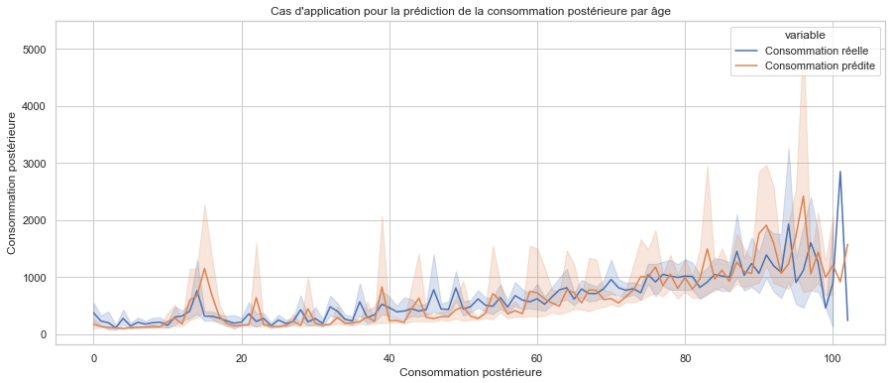


FIGURE 22 – Prediction of the evolution of pure premiums by age

Although the GLMs were not excellent in terms of precision, given the irregularity of the distributions due to the different packages and reimbursement systems defined in the mutual's guarantee grids, we managed to approach the real pure premiums by age with quite good caution. Modeling consumption by major type of act allowed us to detect the peak in consumption for young people aged 17 to 19 years due to orthodontic acts as seen previously.

Note de synthèse

Depuis quelques années, l'assurance santé a connu plusieurs évolutions réglementaires qui ont directement impacté l'activité des complémentaires santé. Parmi ces évolutions, nous retrouvons principalement l'arrivée du 100% santé, et l'entrée en vigueur de la résiliation infra-annuelle.

Disposant d'un portefeuille en assurance santé individuelle, nous souhaitons modéliser l'évolution des flux de contrats sortants du portefeuille ainsi que la consommation postérieure des bénéficiaires liées à ces contrats. Cette évolution doit tenir compte de l'impact des réglementations récentes.

Depuis l'entrée en vigueur de la résiliation infra-annuelle, l'exposition totale du portefeuille sur une année n'est plus maintenue, une grande partie du portefeuille relative aux contrats ayant moins d'un an d'ancienneté peut résilier au cours de l'année et cela présente un risque de non-couverture des frais de l'assureur.

La saisonnalité des remboursement présente entre autre un risque de sous estimation des primes pures depuis l'entrée en vigueur de la résiliation infra-annuelle. En effet, commençons par regarder les remboursements mensuels du portefeuille à disposition.

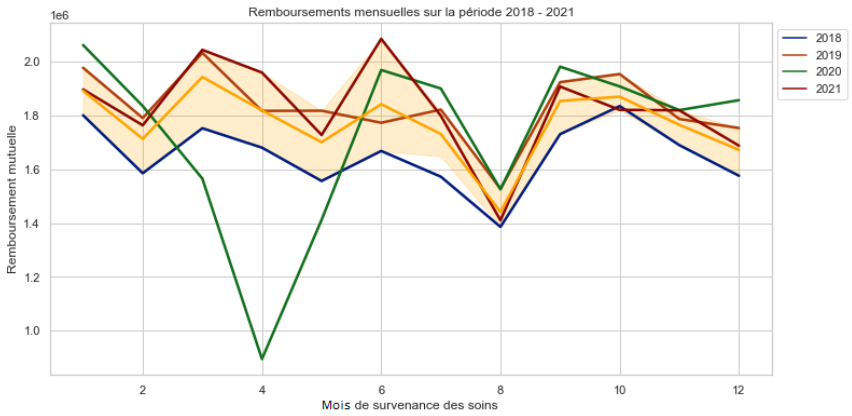


FIGURE 23 – Remboursements mutuelle par mois de survenance par exercice

Avec la résiliation infra-annuelle, la saisonnalité peut avoir un impact direct sur l'estimation de la prime pure des assurés.

Souvent calculée sur toute l'année, l'estimation de la prime pure peut varier pour un calcul sur une période plus courte de l'année.

Afin d'estimer cette variation, nous nous plaçons au début de chaque exercice et calculons les remboursements annuels moyens par bénéficiaire (primes pures) en se basant sur une durée de couverture allant du 1 janvier à une date t qui évolue entre le 1 février et la fin de l'exercice.
Pour un exercice N et t l'écart en nombre de jours entre 1/1/N et le j/m/N, la prime pure annuelle basée sur la période du 1/1/N au j/m/N s'écrit :

$$Prime\,pure(t) = \frac{\sum_{s=0}^{t} Remboursements\,mutuelle(s)}{\sum_{s=0}^{t} Exposition\,totale(s)}$$

Avec,

— Remboursements mutuelle(s) : Somme des remboursements mutuelle référant au jour de survenance s.

— Exposition totale(s) : Somme de l'exposition des bénéficiaires le jour s (en année).

La représentation de ces primes pures pour les années 2018, 2019 et 2021 se présente comme suit :

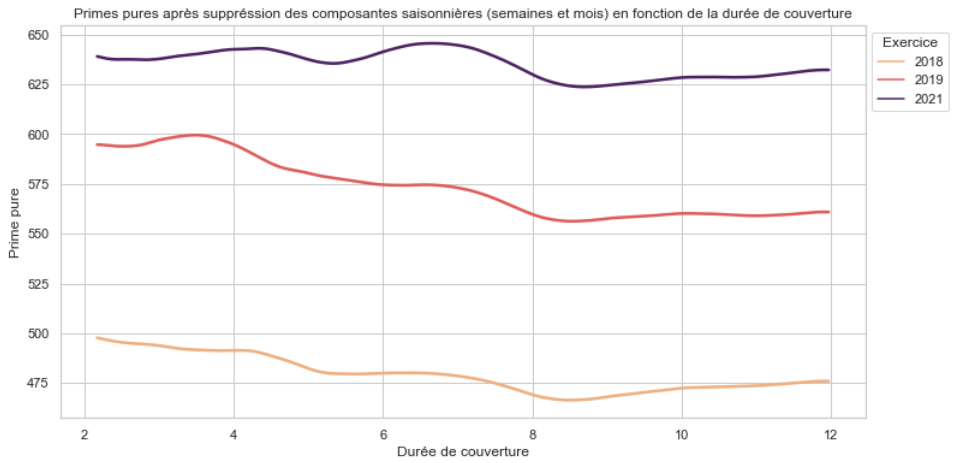


FIGURE 24 – Primes pures lissées en fonction de la durée de couverture de l'année

Les primes pures annuelles suivent une saisonnalité similaire pour ces trois exercices. En rapportant la fin des courbes à zéro (par une translation verticale) et en moyennant sur les trois exercices, nous obtenons la variation moyenne de la prime pure annuelle en fonction de la durée de couverture par rapport à une couverture d'une année complète :

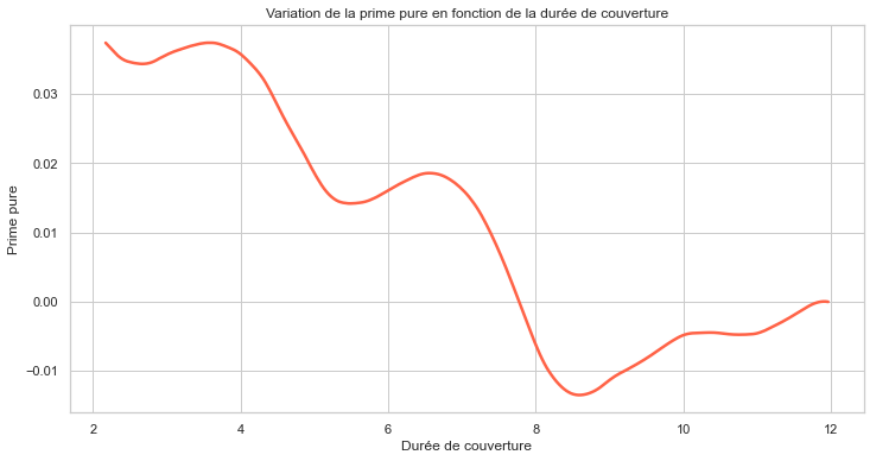
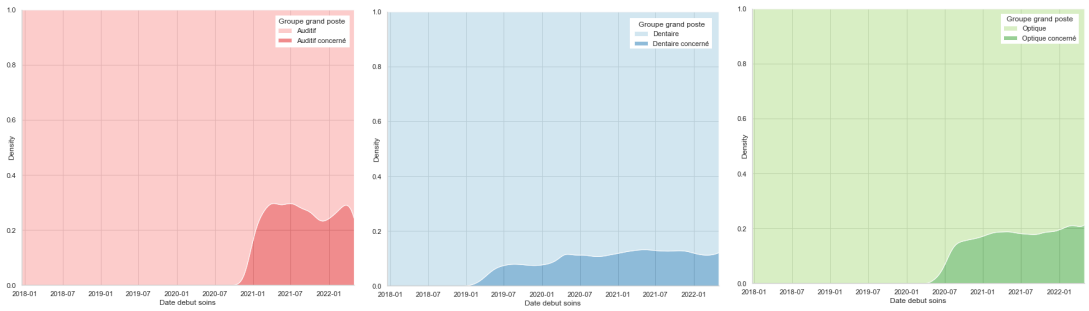


FIGURE 25 – Variation de la prime pure annuelle en fonction de la durée de couverture

Supposons que nous avons estimé les primes pures en se basant sur toute l'année de couverture. Pour un assuré qui résilie en avril, sa prime pure serait donc en moyenne sous-estimée de plus de 3%. De même, un assuré qui résilie en septembre, sa prime pure est en moyenne sur-estimée de 1%.

En plus de la saisonnalité qui impacte les primes pures, la mise en place du 100% santé pourrait impacter la consommation des assurés pour certains actes.
En effet, les actes à zéro reste à charge sont de plus en plus sollicités par les bénéficiaires. Afin de regarder l'impact de l'entrée en vigueur du 100% santé sur la consommation des actes concernés par la réglementation, nous avons commencé par regarder l'évolution de la proportion des codes actes pris en charge par le 100% santé dans le temps pour chaque grand poste concerné.



(a) Proportion des actes concernés par le 100% santé en Audiologie

(b) Proportion des actes concernés par le 100% santé en Dentaire

(c) Proportion des actes concernés par le 100% santé en Optique

La superposition de ces distributions pour les trois postes se présente comme suit :

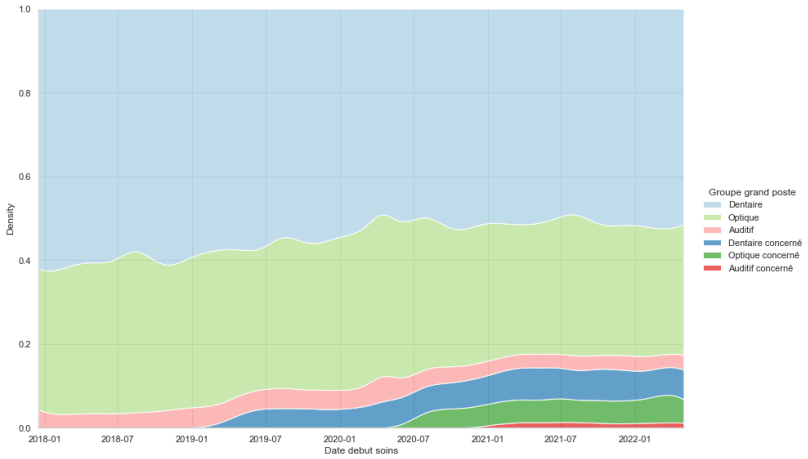


FIGURE 27 – Evolution de la consommation des actes pris en charge par le 100% au cours du temps

Les trois distributions dans la partie basse du graphe représentent la proportion (en nombre d'actes) des actes concernés par le reste à charge zéro pour les trois postes concernés par rapport au reste des actes.

La proportion des actes concernés est restée croissante jusqu'à la mise en place totale des paniers 100% santé. Les assurés ont par la suite plus recours aux types d'actes pris en charge par le 100% santé après l'entrée en vigueur de la réglementation. Ces actes représentent désormais 15% des actes du portefeuille.

L'évolution du coût moyen d'un acte par mois selon s'il est concerné par la réglementation pour les 3 derniers postes se présente comme suit :

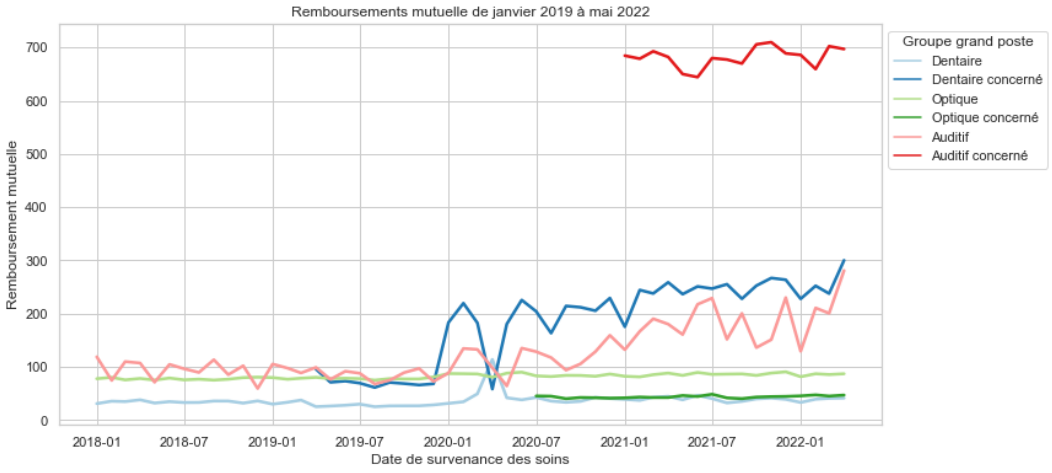


FIGURE 28 – Evolution du coût moyen des actes concernés par le 100% santé par mois

le coût moyen des actes concernés par le 100% santé est toujours supérieur au reste des actes pour les postes audiologie et dentaire.

Enfin l'augmentation de la proportion des actes pris en charge par le 100% santé et leurs coûts moyens plus élevés du reste des actes donneront une hausse de la consommation en termes de coût moyen et fréquence.

Dans ce contexte, la modélisation de l'évolution du portefeuille a nécessité la mise en place d'un processus à deux étapes principales qui se présente comme suit :

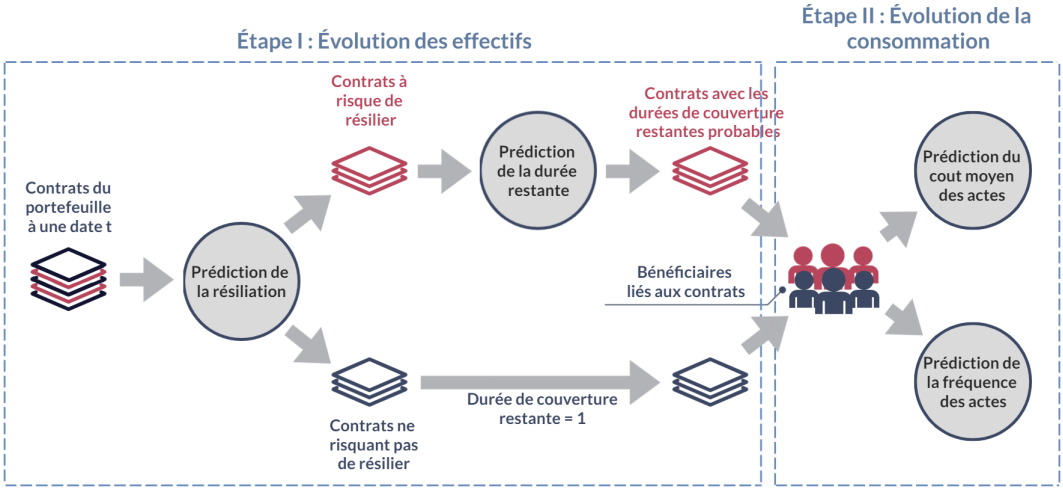


FIGURE 29 – Schéma du processus de modélisation

La première étape concerne l'évolution de l'effectif du portefeuille. En observant le portefeuille à une date t, on souhaite connaître si un assuré risque de résilier dans l'année qui suit et sa durée de couverture restante dans le cas d'une résiliation probable. Dans cette partie les prédictions sont faites par contrat vu que la résiliation d'un assuré concerne dans la majorité des cas tous les bénéficiaires du même contrat.

Une fois que la durée de couverture restante, à un horizon d'un an, est connue pour chaque contrat, on passe aux données relatives à chaque bénéficiaire pour prédire sa consommation durant sa période de couverture restante par la prédiction du coût moyen de ses actes et de sa fréquence de consommation. Ce modèle permet la détection d'un changement comportemental relatif à une résiliation probable.

La prédiction des variables à expliquer, à une date donnée d'observation du portefeuille, nécessite la liaison des évènements antérieurs à cette date aux valeurs prises par ces variables sur une durée postérieure.

La construction des bases de données est donc réalisée par la concaténation de données provenant d'une succession d'observations mensuelles du portefeuille sur la période d'activité étudiée. Les observations sont faites le premier de chaque mois. Les données

explicatives, respectivement à expliquer, concernent une période d'un an avant, respectivement un an après, la date d'observation, comme le montre le graphe si dessous :

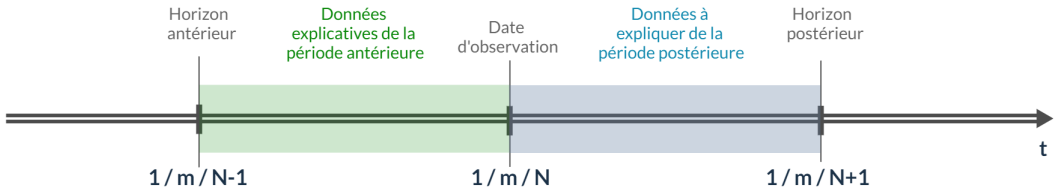
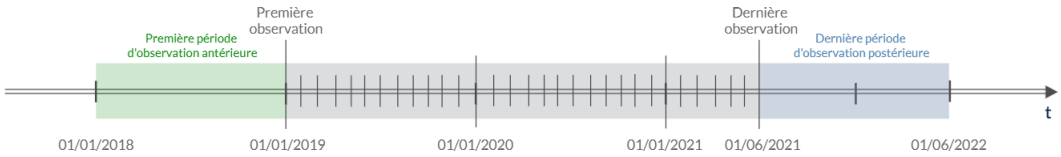


FIGURE 30 – Périodes d'observations du portefeuille

Les données transmises par la mutuelle correspondent à la période du **1 janvier 2018** au **6 mai 2022**. Vu la contrainte sur les horizons antérieurs et postérieurs, les observations mensuelles du portefeuille seront du **1 janvier 2019** au **1 mai 2021**.



Les variables décrivant l'activité antérieure d'un assuré par rapport à une date d'observation donnée se décomposent de la manière suivante :

— Variables relatives aux **profils des assurés** : Âge, sexe, type, code postal, ...
— Variables relatives aux **contrats** : Gamme, garantie, cotisations, taxes, ...
— Variables relatives à la **consommation** : Remboursement mutuelle par poste d'actes, nombre d'actes, ... (sommées sur une année antérieure à la date d'observation.)

Afin de fournir aux modèles toutes les informations nécessaires aux prédictions, il a été nécessaire de créer quelques variables complémentaires à partir des données à disposition. Parmi ces variables nous retrouvons les indexations antérieures et postérieures des tarifs, et la durée avant l'éligibilité des contrats à résilier.

Pour une date d'observation le 1/m/N, les variables d'indexation se présentent comme suit :

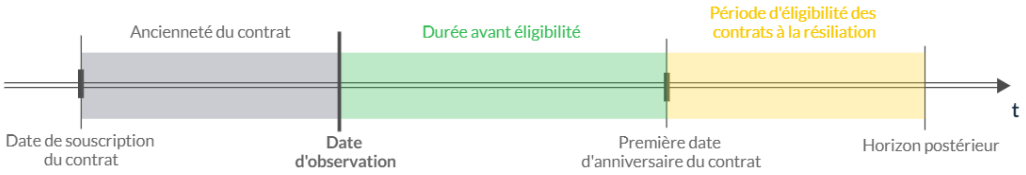
$$\textbf{Indexation antérieure} = \frac{Cotisation\ annuelle(N) - Cotisation\ annuelle(N-1)}{Cotisation\ annuelle(N-1)}$$

$$\textbf{Indexation postérieure} = \frac{Cotisation\ annuelle(N+1) - Cotisation\ annuelle(N)}{Cotisation\ annuelle(N)}$$

La variable **Indexation postérieure** nous permet un certain dynamisme des modèles. En effet, les prédictions peuvent être orientées en fonction de l'indexation postérieure, par le biais de simulations, afin d'analyser l'impact de la valeur d'indexation sur le flux de contrats sortant probable.

Dans le but de fournir aux modèles l'information sur l'éligibilité des adhérents à résilier, nous créons une variable nommée "Durée avant éligibilité" qui donne la durée avant laquelle l'adhérent peut résilier son contrat selon la date d'observation.
Cette variable est calculée en année et dépend de la date d'observation. Par exemple, le calcul de cette variable après l'entrée en vigueur de la résiliation infra-annuelle se présente comme suit :

Pour une date d'observation donnée, cette variable vaut 0 pour les contrats d'ancienneté supérieure à 1 an (donc éligibles à résilier à tout moment).
Dans le cas d'une ancienneté inférieure à 1 an à la date d'observation, le calcul de cette variable se présente comme suit :



$$\begin{cases} DAE = 0 & si\ AC \geqslant 1 \\ DAE = 1 - (AC - \lfloor AC \rfloor) & si\ AC < 1 \end{cases}$$

Avec,

DAE : Durée avant éligibilité
AC : Ancienneté contrat

$\lfloor AC \rfloor$ désigne la partie entière inférieure de l'ancienneté du contrat.

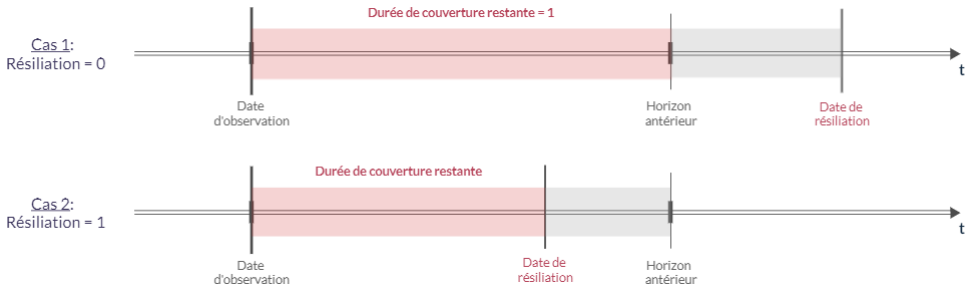
Du coté des variables à expliquer, les variables à prédire pour l'étape I du modèle globale sont la résiliation et la durée de couverture restante. Ces variables sont attribuées aux bases par contrat et définies comme suit :

$$\textbf{Résiliation} = 1_{\{Date\ resiliation\ <\ Date\ horizon\}}$$

$$\textbf{Durée de couverture restante} = \begin{cases} 1 & si \quad Resiliation = 0 \\ DR\ -\ DO & si \quad Resiliation = 1 \end{cases}$$

Avec,

DR : Date de résiliation
DO : Date d'observation



La prédiction de la consommation postérieure dans la partie II se fait par bénéficiaire, les variables à prédire sont pour un poste d'actes i donnée :

— **Coût moyen postérieur (i)** = Remboursements postérieurs (i) / Nombre d'actes postérieurs (i)
— **Fréquence postérieure (i)** = Nombre d'actes postérieurs (i) / Durée de couverture restante

Ces variables sont calculées sur la période de couverture postérieure à l'observation et avec un horizon d'un an.

Enfin, dans le but d'exploiter la totalité des informations à disposition, des dates d'observation ultérieures au 1 mai 2021 ont été prises en compte pour les cas de résiliation avant le 6 mai 2022. En effet, dans le cas d'une résiliation l'horizon se limite à la date de résiliation.

En comparant les distributions des variables explicatives avec la résiliation postérieures des contrats, nous trouvons une différence significative pour les variables suivantes :

Âges des adhérents

La distribution de l'âge des adhérents en fonction de la résiliation du contrat à l'horizon d'un an est représentée de la manière suivante :

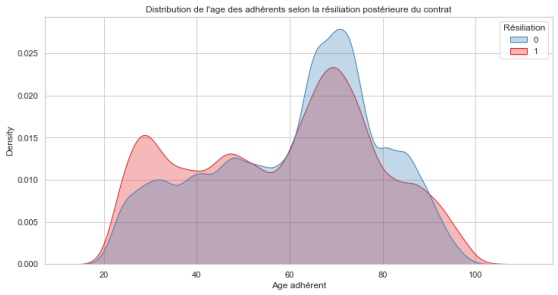


FIGURE 31 – Distribution de l'âge des adhérents selon la résiliation du contrat

Nous constatons que les résiliations sont plus représentées pour les contrats dont les adhérents sont âgées de moins de 50 ans. Cet écart est surtout présent pour les adhérents de moins de 35 ans. D'autre part, la résiliation est aussi mieux représentée pour les contrats des adhérents les plus âgés, ces résiliations sont probablement dues aux décès plus fréquents sur cette tranche d'âge.

Durée avant éligibilité

La distribution de cette variable, sur toute la durée de couverture du portefeuille, en fonction de la résiliation des contrats est représentée comme suit :

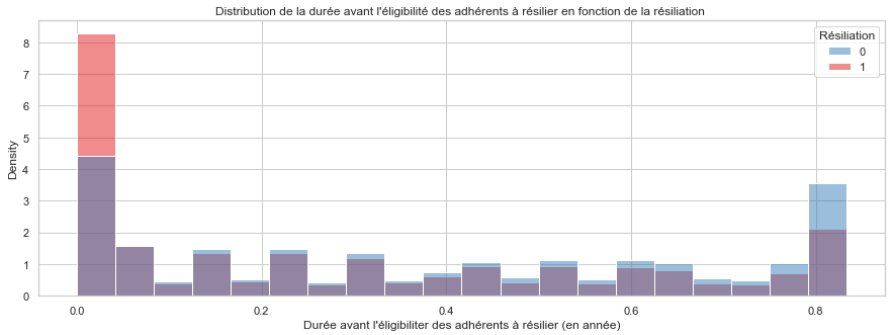
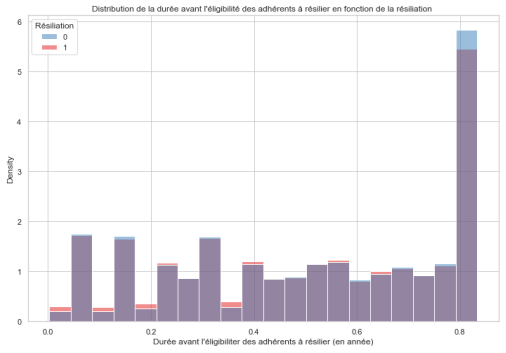
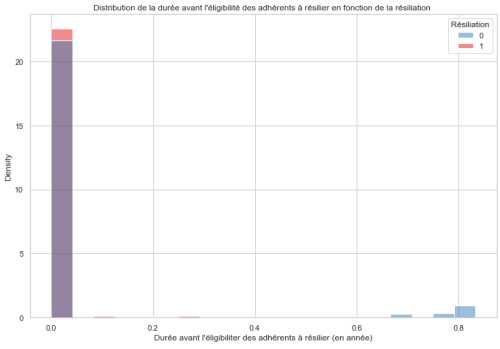


FIGURE 32 – Distribution de la durée avant éligibilité en fonction de la résiliation du contrat

Cette durée a connu un avant et un après entrée en vigueur de la résiliation infra-annuelle en fin 2020. Les distributions de cette variable relatives aux observations de 2019 et 2021 se présentent comme suit :



(a) Distribution de la durée avant éligibilité en 2019

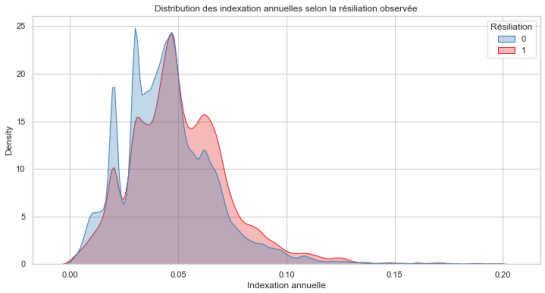


(b) Distribution de la durée avant éligibilité en 2021

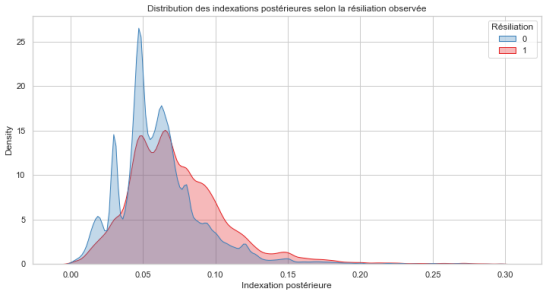
Contrairement à l'année 2019, la majorité des contrats sont éligibles à résilier en 2021 grâce à la nouvelle réglementation.

Indexation antérieure et postérieure

Les distributions de l'indexation antérieure et postérieure en fonction de la résiliation se présentent comme suit :



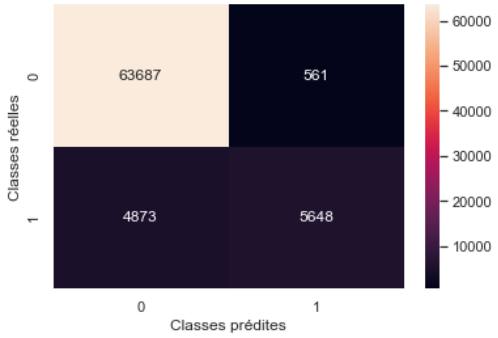
(a) Distribution de l'indexation antérieure



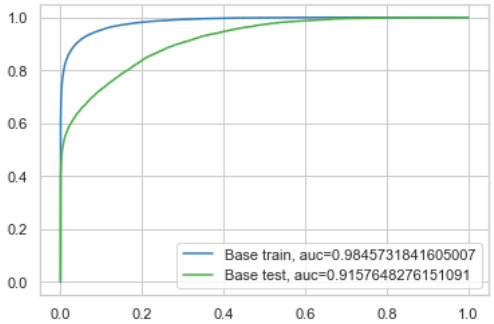
(b) Distribution de l'indexation postérieure

Les indexations de niveau élevé, qu'elles soient antérieures ou postérieures, favorisent le risque de résiliation des contrats d'après ces deux graphes.

Pour la prédiction des résiliations, nous avons opté pour un modèle XGBoost. Afin d'améliorer la performance du modèle, les contrats ayant servis à l'apprentissage des modèles sont relatifs à des résiliations après l'entrée en vigueur de la résiliation infra-annuelle, les résultats de la classification obtenue se présentent comme suit :



(a) Matrice de confusion pour la base test



(b) Courbe ROC

	precision	recall	F1-score	Support
Non-résiliation	0,93	0,99	0,96	64 248
Résiliation	0,91	0,54	0,68	10 521

La sensibilité du modèle a atteint 54% avec une précision de 91%. La prise en compte des contrats présents à l'entrée en vigueur de la résiliation infra-annuelle nous a permis de détecter plus de résiliations avec une meilleure précision.

L'interprétation par la méthode SHAP de la contribution des variables aux prédictions du modèle se présente comme suit :

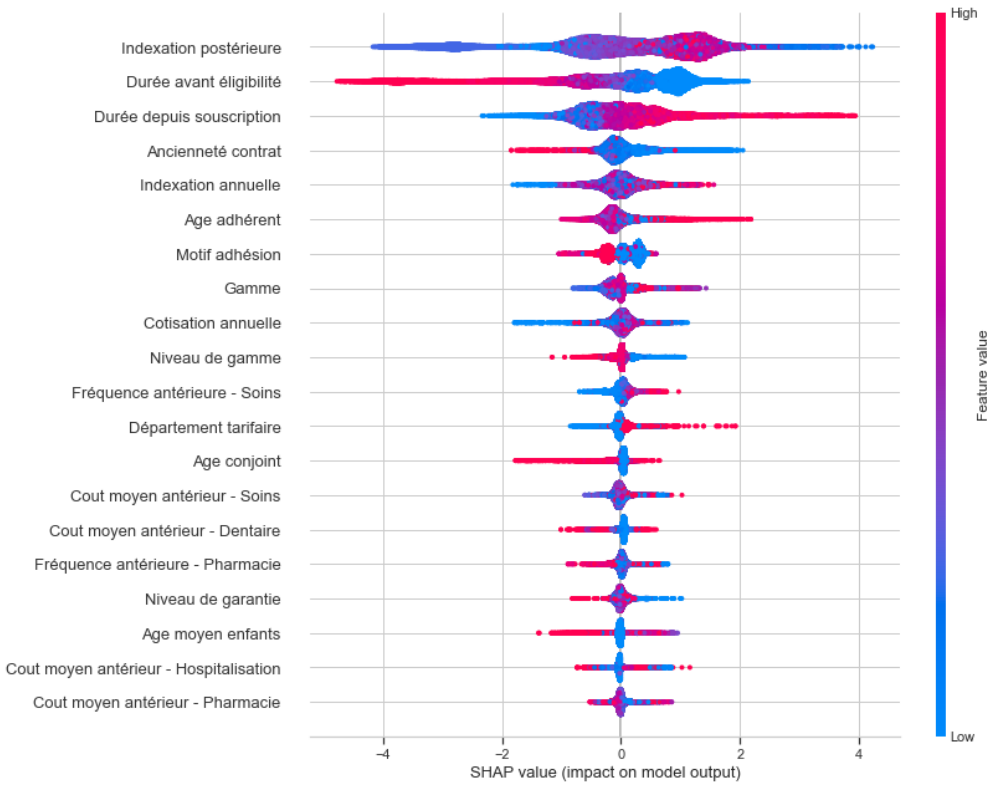


FIGURE 36

D'après les valeurs de Shapley, la variable la plus influente sur la prédiction est l'**indexation postérieure**. La représentation montre que la majorité des indexations postérieures élevées ont fait augmenter la probabilité de résiliation des contrats concernés.

La **durée avant éligibilité** représente la deuxième variable la plus influente sur les prédictions du modèle. On voit qu'une durée avant éligibilité haute diminue la probabilité de résiliation d'un contrat. En effet, plus la valeur que prend cette variable est grande plus la période pendant laquelle un contrat peut être résilié dans l'horizon postérieur est faible. De la même façon, les valeurs faibles de cette variable représentants des contrats qui sont éligibles à la résiliation tout au long de l'année sont représentés avec des valeurs de Shapley positives qui augmentent la probabilité de résiliation.

La troisième variable est la **durée depuis souscription**. Nous pouvons voir que cette variable est croissante en fonction de la valeur de Shapley. En effet, nous avons vu dans le chapitre précédant que la majorité des contrats changent de gamme ou de niveau de garantie au bout de 5 ans, cette variable réfère à la date de souscription à la

gamme et au niveau de garantie actuels. Ainsi une valeur élevée de cette durée augmente la probabilité de résiliation.

De l'autre coté, l'**ancienneté des contrats** est décroissante en fonction de la valeur de Shapley. Cela reflète la fidélité des anciens adhérents.

Afin d'interpréter l'influence de l'**âge de l'adhérent** sur la résiliation des contrats, nous traçons la valeur de Shapley en fonction de l'âge de l'adhérent (nous prendrons en compte la présence d'un conjoint ou non dans le contrat pour une meilleure interprétation) :

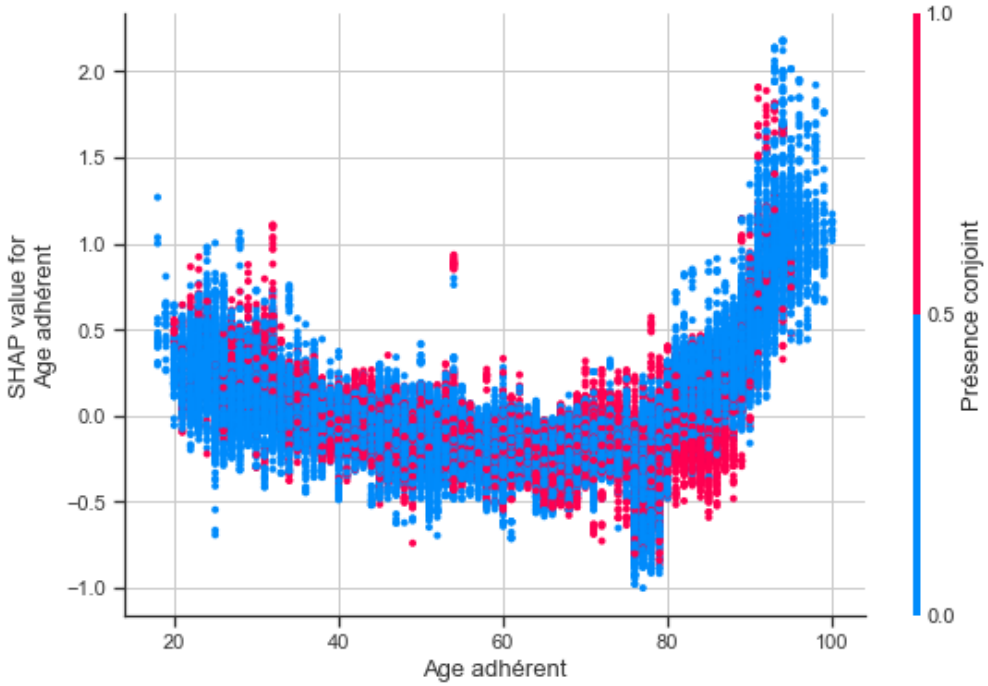


FIGURE 37 – Valeur de Shapley en fonction de l'âge des adhérents

Comme nous l'avons remarqué en comparant les distribution de l'âge des adhérents en fonction de la variable résiliation, les résiliations concernent surtout les jeunes de moins de 40 an et les plus âgés (probablement à cause des décès). Les valeurs de Shapley nous le confirme par le biais de cette représentation, les valeurs de Shapley décroient pour les plus jeunes et croient pour les plus âges. En particuliers, nous retrouvons que la présence d'un conjoint pour les adhérents âges de plus de 80 ans baisse la probabilité de résiliation.

Au niveau de la durée de couverture restante. Nous allons comparer sa distribution en fonction pour deux mois d'observation en fonction de l'année d'observation.

La distribution de la durée de couverture restante pour une observation du mois de mars se présente comme suit :

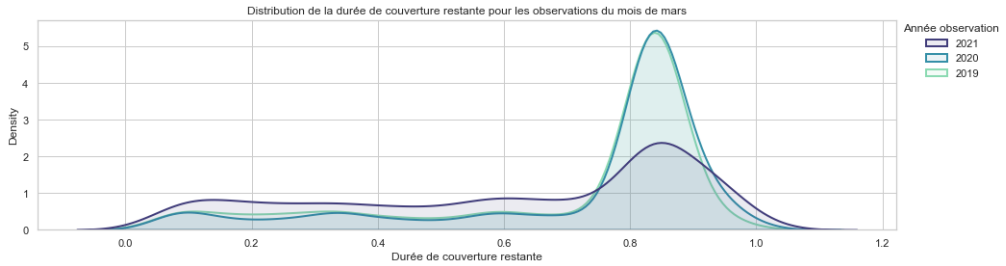


FIGURE 38 – Distribution de la durée de couverture restante pour les observations du mois de **mars**

Les pics observées sur la courbe sont relatifs au mois de janvier. La distribution de la durée de couverture restante est quasi identique pour les observations de 2019 et 2020. En effet, la résiliation infra-annuelle n'entre en vigueur qu'à partir du mois de décembre 2020 donc il n'y a pas une grande différence en terme d'éligibilité à résilier pour une observation en mars.

En revanche, la distribution de 2021 est mieux répartie en 2021 grâce à la résiliation infra-annuelle. Les résiliations observées entre mars et le janvier qui suit (durée entre 0 et 8,33) ont presque doublées par rapport aux deux années précédentes.

La distribution de la durée de couverture restante pour une observation du mois d'octobre se présente comme suit :

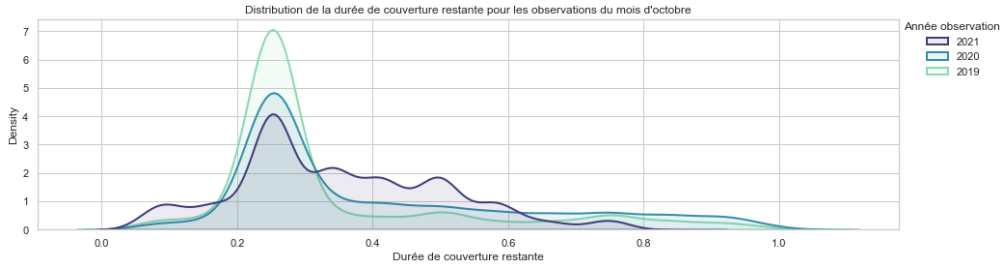
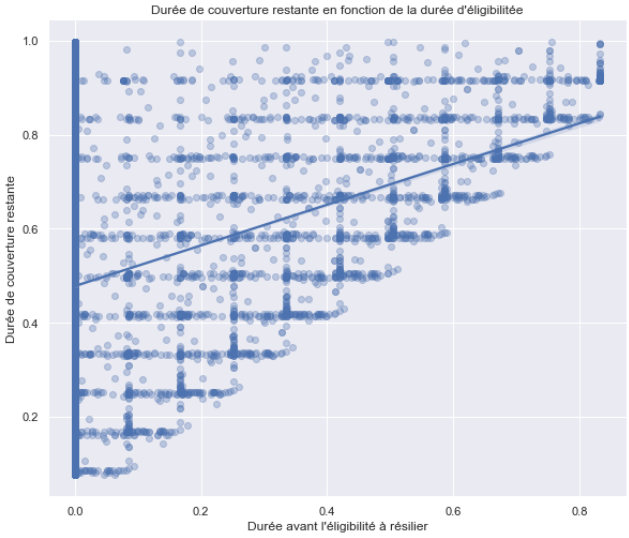


FIGURE 39 – Distribution de la durée de couverture restante pour les observations du mois d' **octobre**

Pour les observations de 2020, l'impact de la résiliation infra-annuelle est visible cette fois. En effet, la réglementation est entrée en vigueur en décembre seulement deux mois après. Par la suite, les résiliations sont moins concentrées en janvier et plus réparties sur la période postérieure.

En 2021, la distribution prend en compte seulement les résiliations datées d'avant le 6 mai 2022 (notre horizon de données postérieur. Cette distribution est plus répartie sur l'ensemble de l'année postérieur à l'observation.

La durée de couverture restante est encadrée par la durée avant éligibilité. En effet, les contrats ne peuvent résilier qu'après que la durée avant laquelle ils seront éligibles à résilier s'écoule. La durée de couverture restante en fonction de la durée avant éligibilité est représentée ci-dessous :



La durée de couverture restante est bien toujours supérieure à la durée avant éligibilité. Cette variable sera donc d'une grande importance pour la prédiction.

Le modèle de régression XGBoost utilisé se base uniquement sur les cas de résiliation, le nombre de contrats résiliés après 2021 est de 6 523 contrats, Suite à l'entraînement du modèle, les résultats du test obtenues se présentent comme suit :

Score d'entraînement	MAE	RMSE
0,78	0,16	0,21

Pour la deuxième partie du modèle, la modélisation de l'évolution de la consommation postérieure du portefeuille se fait par la prédiction du coût moyen et fréquence pour chaque bénéficiaire durant sa période de couverture restante et pour chaque grand poste d'actes.

La distribution des coûts moyens et fréquences postérieurs et antérieurs aux dates d'observation par poste d'actes se présente comme suit :

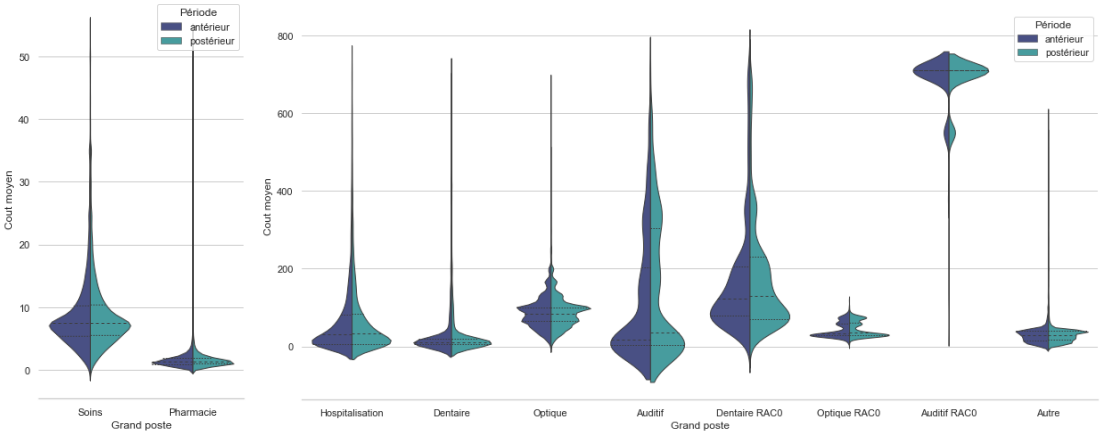


FIGURE 40 – Coûts moyens postérieurs et antérieurs par poste

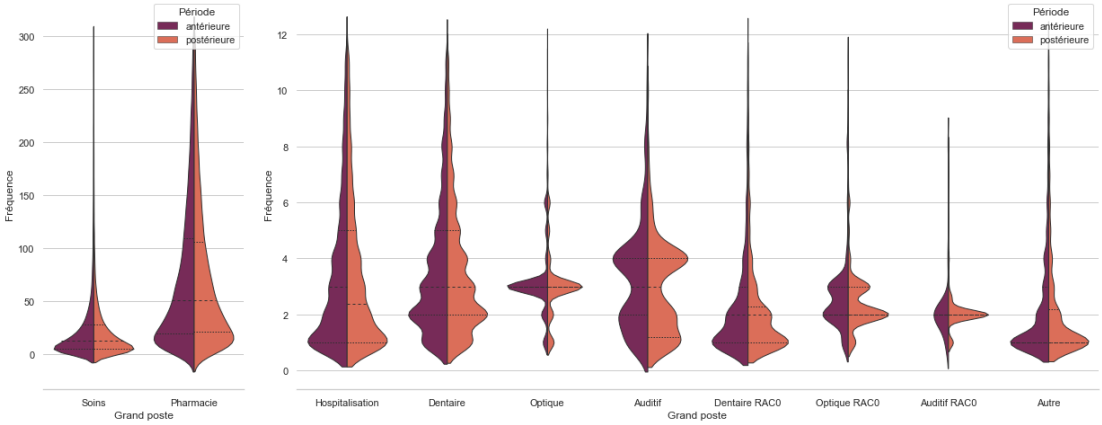


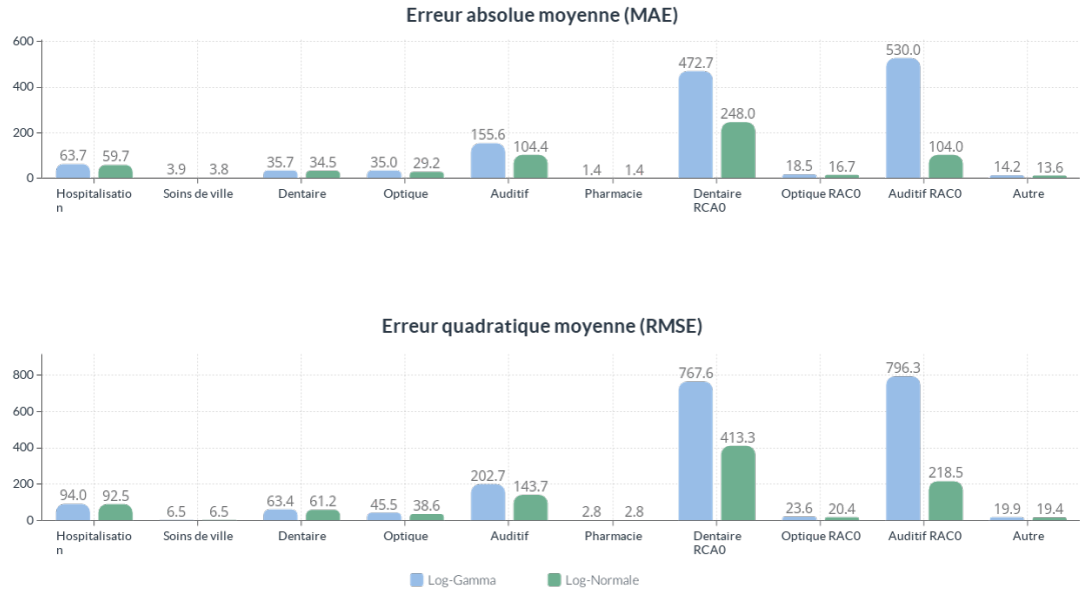
FIGURE 41 – Fréquences postérieures et antérieures par poste

Pour ne pas double compter les coûts moyens et fréquences antérieurs et postérieurs à cause des observations successives et pour prendre compte du 100% santé sur tous les postes concernés, ce graphe est issu d'une unique observation d'avril 2021.

Les distributions des coûts moyens et fréquences postérieures à cette date d'observation et par poste d'actes restent très proches de celle observés sur l'année antérieure à la date d'observation.

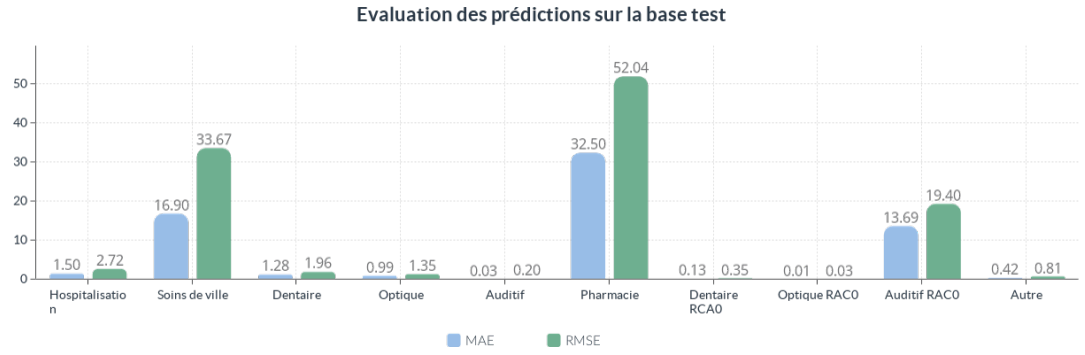
Pour la prédiction du coût moyen postérieur pour chaque poste d'acte, nous allons choisir entre un modèle GLM-LogGamma et un modèle GLM-LogNormale en fonction de leurs performances pour chaque poste d'acte.

Par poste, les résultats des modèles obtenus se présentent comme suit :



Sur tous les modèles obtenus, les modèles Log-Normales présentent de meilleurs résultats. En effet, les valeurs de RMSE et de MAE pour ce modèle sont toujours inférieures à celle des modèles Log-Gamma. Le modèle Log-Normale est donc mieux adapté à nos distributions. Ainsi, nous retenons ce modèle pour la prédiction des coûts moyens postérieurs des actes.

Pour la prédiction de la fréquence, nous utilisons un GLM Poisson. Par poste, les résultats des modèles obtenus se présentent comme suit :



La précision des modèles obtenus par rapport à chaque poste d'actes est acceptable. Les modèles obtenus semblent performants.

Enfin nous procédons à la mise en oeuvre du modèle de résiliation des contrats au cours d'un cas d'application sur une portion de contrats retirée du portefeuille au début de la modélisation. Cette portion concerne 5 965 contrats tous présents au 1 mai 2021.

Le cas d'application consiste à prédire l'évolution du nombre de contrats présents à l'horizon d'un an à partir d'une observation effectuée le 1 mai 2021 pour les contrats en question. L'évolution de la pyramide des âges via l'implémentation du modèle se présente comme suit :

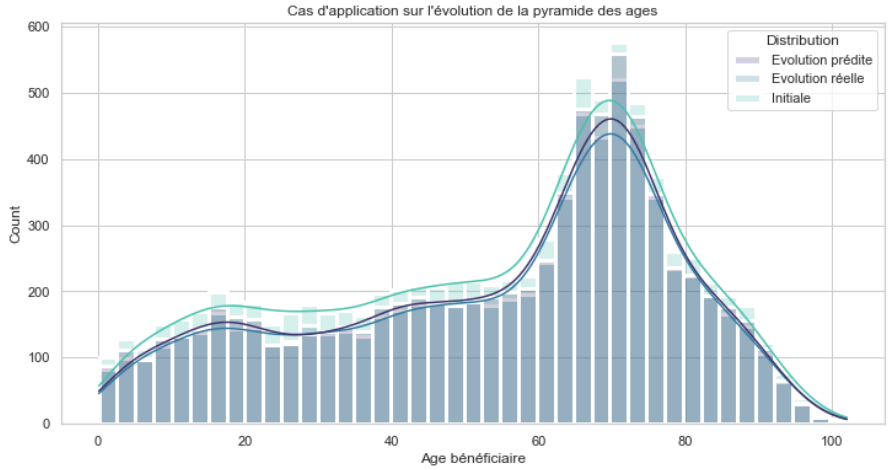


FIGURE 42 – Prédiction de l'évolution de la pyramide des âges

La prédiction du modèle de l'évolution de la pyramide des âges est très proche de la réalité. Cependant, cette dernière reste légèrement supérieure à l'observation réelle de la pyramide au bout d'un an, le modèle a donc prédit moins de résiliations pour certains âges.

Pour avoir une idée sur l'impact de l'évolution de la composition des effectifs sur la consommation du portefeuille, nous traçons les distributions des âges obtenues.

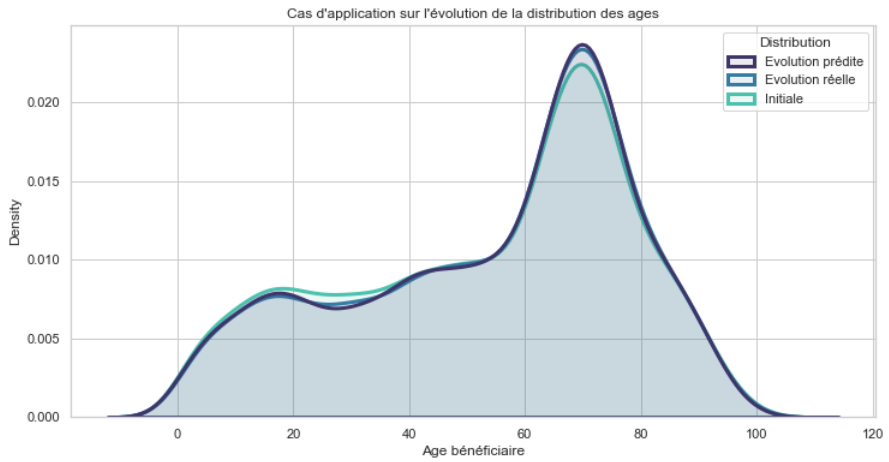


FIGURE 43 – Prédiction de l'évolution de la distribution des âges

La distribution des âges obtenue est très proche de la réalité. Le XGBoost a particulièrement su détecter les résiliations concernant les jeunes âgés de 20 à 40 ans ce qui a permis de prédire le creux observé sur cette tranche d'âge.

Cette évolution de distribution permet de mieux estimer la consommation du portefeuille vu que l'âge des bénéficiaires est le principal indicateur sur la consommation de ces derniers.

L'application des modèles de prédiction des coûts moyens et fréquences par poste nous permet d'obtenir les primes pures pour l'année postérieure à l'observation que nous traçons en fonction de l'âge des assurés comme suit :

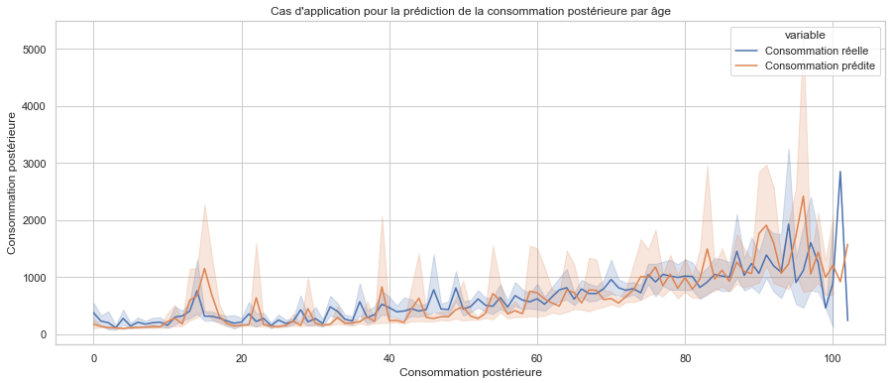


FIGURE 44 – Prédiction de l'évolution des primes pures par âge

Malgré que les modèles GLMs n'étaient pas excellents en terme de précision, vu l'irrégularité des distributions due aux différents forfaits et systèmes de remboursement définis dans les grilles de garanties de la mutuelle, nous arrivons à approcher les primes pures réelles par âge avec une prudence assez bonne. Le fait de modéliser la consommation par grand poste nous a permis de détecter le pic de consommation pour les jeunes de 17 à 19 ans dû aux actes en orthodontie comme vu précédemment.

Remerciements

En premier lieu, je tiens à remercier tous mes proches, pour leur soutien et leur encouragement tout au long de ma scolarité.

Je remercie particulièrement mon tuteur en entreprise M. Pierre COCHAIN de m'avoir accompagné tout au long de mes travaux. Sa disponibilité, son écoute et ses recommandations m'ont été précieux pour mener à bien ce mémoire.

Je tiens également à remercier chaleureusement Mme Annaelle LE BERRE, tutrice EURIA pour sa disponibilité, son suivi et ses conseils pertinents.

Je souhaite remercier toute l'équipe enseignante de l'EURIA pour la qualité de l'enseignement.

Enfin je remercie toutes les personnes m'ayant soutenu pour la rédaction de ce mémoire.

Table des matières

Chapitre 1

Introduction

L'assurance santé en France subie de plus en plus de changements depuis l'entrée en vigueur de plusieurs réglementations telles que le 100% santé et la résiliation infra-annuelle. Face à ces nouvelles lois, les complémentaires santé essayent de retrouver un équilibre face à ces changements impactant leur activité.

Disposant d'un portefeuille en assurance santé individuelle, ce mémoire vise à modéliser l'évolution du portefeuille en terme d'effectifs et de consommation dans un horizon postérieur d'un an.
La modélisation de l'évolution des effectifs repose sur la prédiction des résiliations à l'horizon d'un an des contrats pour une date d'observation donnée puis la prédiction de la durée de couverture restante pour les contrats à risque de résilier.
D'un autre coté, la modélisation de la consommation postérieure à la date d'observation se fait par la prédiction des coûts moyens et fréquences pour chaque poste d'actes.

Pour se faire nous allons introduire les systèmes obligatoire et complémentaire qui composent l'assurance santé en France puis regarder de prêt l'activité des complémentaires santé.
Ensuite nous allons introduire un système permettant la modélisation de cette évolution depuis la base de données à disposition.
Avant de commencer la modélisation, nous allons faire procéder à des analyses univariées et bivariées des variables explicatives en fonction des variables qu'on veut expliquer.
Une fois les modèles entraînés, nous passerons aux tests et à l'interprétation des modèles obtenus.

Enfin, nous passerons à un cas d'application de notre modèle pour la prédiction de l'évolution d'une portion du portefeuille à l'horizon d'un an depuis une observation du 1 janvier 2021.

Chapitre 2

Contexte

Dans ce chapitre nous allons présenter le système au sein duquel sont impliquées les garanties d'assurance santé individuelle. Ainsi, nous allons commencer par définir les régimes, obligatoire et complémentaire, qui forment le mécanisme de remboursement des frais de soins des assurés et les différentes mesures législatives récentes qui ont pu impacter ce système, comme l'entrée en vigueur du 100% santé ou bien la mise en place de la résiliation infra-annuelle. De plus, dans une dernière partie, nous présentons les garanties en santé individuelle.

2.1 Régime général obligatoire

La sécurité sociale protège l'ensemble des résidents en France et distingue 3 types de régimes selon les catégories socioprofessionnelles de la population (Régime général, Régime agricole et régimes spéciaux).
Le régime général représente l'un des principaux régimes de la sécurité sociale. Il se compose de cinq branches en fonction des types de risque pris en charge comme l'illustre le graphe ci-dessous :

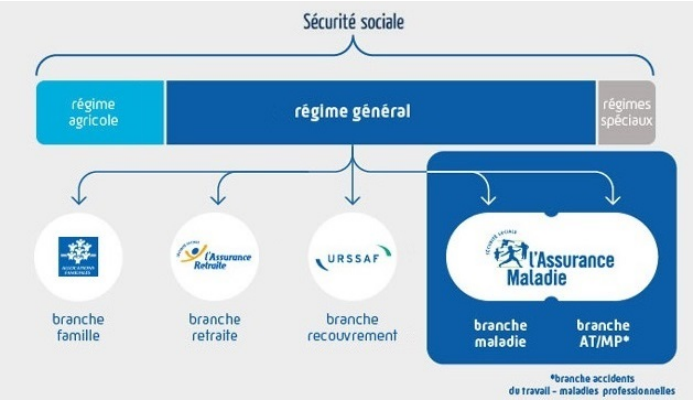


FIGURE 2.1 – Régimes et branches de la sécurité sociale (source)

Parmi les affiliés du régime général, on retrouve les travailleurs salariés (hors agricoles et régimes spéciaux), les travailleurs indépendants (depuis 2018) et les personnes résidant en France.
La branche maladie couvre les risques de maladie, maternité, invalidité et décès. Elle est gérée par la Caisse nationale de l'assurance maladie (CNAM).

Ressources de la branche maladie

Les recettes de la branche maladie du régime général proviennent de plusieurs catégories de sources. La répartition de ces ressources se présente comme suit :

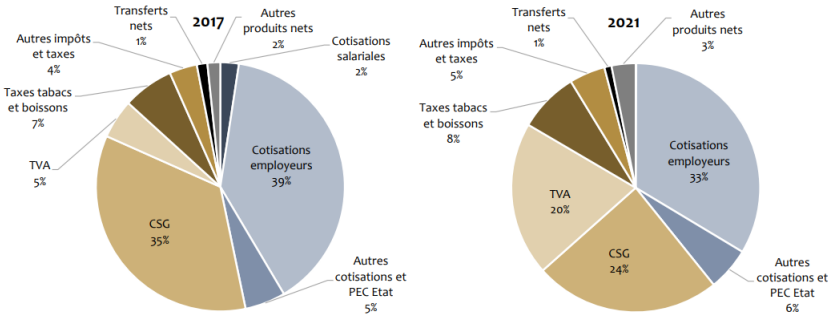


FIGURE 2.2 – Structures des recettes de la branche maladie du régime général en 2017 et 2021 (source)

La **contribution sociale généralisée (CSG)** est un impôt, souvent prélevée à la source, calculée sur l'ensemble des revenus des résidents en France.
La répartition des recettes a bien changé entre 2017 et 2021 avec les transformations législatives telles que la suppression de la part salariale en 2018.

Les recettes de la branche maladie représentent 39% des recettes des régimes de base de la sécurité sociale, soit 207,91 Md€en 2021. Cependant, cette branche reste largement déficitaire à la suite des dépenses engendrées par la crise sanitaire. Les résultats durant les dernières années se présentent comme suit :

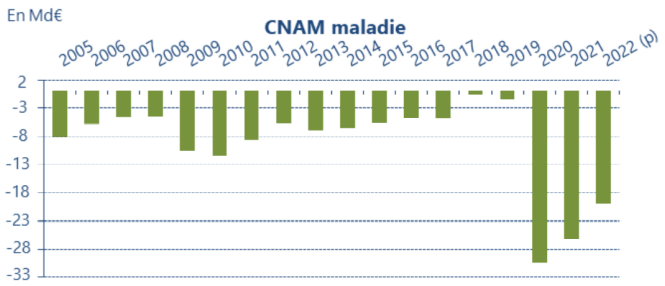


FIGURE 2.3 – Évolution du solde de la branche maladie du régime général (2005-2022) (source)

Dépenses de la branche maladie

Le montant des dépenses de la branche maladie d'élève à 217,6 Md€ et représente 47% des dépenses du régime général.

La part des dépenses consacrée au risque de maladie correspond à la prise en charge d'une partie des frais de soins de ses affiliés. Cette partie est calculée à partir d'un **taux de remboursement** appliqué à un **tarif de référence**, qui dépend de l'acte médical.

Le tarif de référence est appelé **base de remboursement de la Sécurité sociale (BRSS)**, il est composé de la partie remboursée par la sécurité sociale, d'une participation forfaitaire (ou une franchise médicale dans le cas d'actes pharmaceutiques, paramédicaux ou frais de transport) et du **ticket modérateur** qui représente le reste à charge de l'assuré.

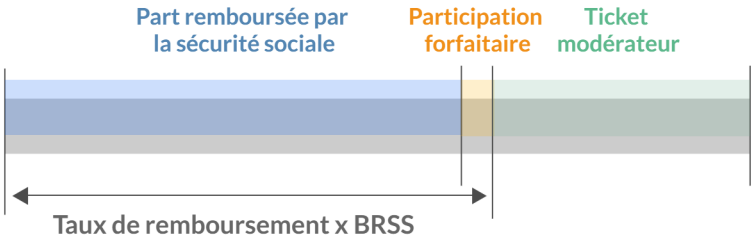


FIGURE 2.4 – Composition de la BRSS

La **participation forfaitaire** est de 1€ et s'applique pour les actes en médecine générale, radiologie et analyses de biologie médicales. Elle a été introduite au début de l'année 2008 et ne concerne que les affiliés âgés de plus de 18 ans et ne peut dépasser les 4€ /jour.
Quant à la **franchise médicale**, elle est de 0,5€ par boite de médicament ou acte paramédical et de 2€ pour un acte de transport. Elle ne concerne que les affiliés âgés de plus de 18 ans et est plafonnée à 50€ /an.

Le **reste à charge (RAC)** d'un acte pour un affilié se compose comme suit :
— La participation forfaitaire ou franchise médicale ;
— Le ticket modérateur ;
— Une éventuelle majoration du ticket modérateur ;
— Un éventuel dépassement d'honoraires ;

Certains actes, dits "hors nomenclature (H.N)", ne sont pas pris en charge par le régime obligatoire. Un régime complémentaire sert à couvrir une partie des frais de santé restant à la charge des affiliés.

2.2 Régimes complémentaires

Les régimes complémentaires agissent en complément du régime général afin de couvrir une part supplémentaire des frais de santé des affiliés.

Afin de couvrir la totalité de ses frais de santé, une personne peut souscrire à un **contrat d'assurance santé complémentaire** auprès d'une société d'assurance, d'une mutuelle ou d'une institution de prévoyance.

2.2.1 Contrat d'assurance santé complémentaire

Un contrat d'assurance santé complémentaire sert à couvrir les frais de santé restant à la charge des assurés lors de la survenance des sinistres contre le versement d'une cotisation. Cette couverture est obligatoire pour les salariés du secteur privé avec une prise en charge des cotisations à hauteur de 50% minimum. Un projet autour de la mise en place d'un tel système pour les salariés du secteur public est en cours de discussion.

Il existe trois types de contrat d'assurance santé complémentaire :

— **Les contrats collectifs** : Ces contrats sont souscrits par les employeurs pour couvrir les salariés de l'entreprise.

— **Les contrats individuels** : Ces contrats peuvent êtres souscrits par les auto-entrepreneurs et les non-salariés.

— **Les contrats collectifs à adhésion facultatifs** : L'adhésion à ces contrats n'est pas obligatoire. Ces contrats sont fiscalement moins avantageux que les contrats collectifs et sont souvent souscrits par des associations.

2.2.2 Les organismes d'assurance complémentaire

Les organismes d'assurances complémentaire se composent de sociétés d'assurance, de mutuelles et d'institutions de prévoyance. Ces organismes assurent la prise en charge de la part des risques non couverte par l'assurance maladie.

Le nombre d'organismes d'assurance complémentaire agréés par l'Autorité de contrôle prudentiel et de résolution (ACPR) ne cesse de diminuer au fil des années. En 2020, on compte 683 organismes dont 369 des mutuelles contre 1631 organismes en 2006 avec 1158 mutuelles, comme le montre le graphe ci-dessous :

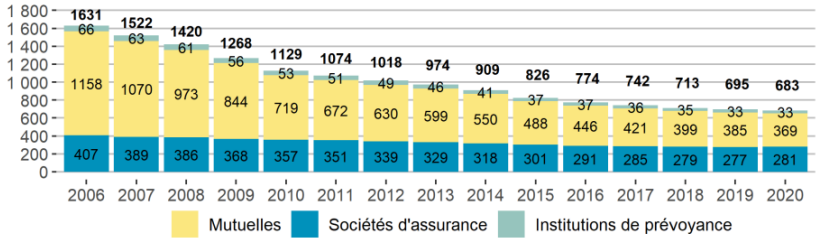


FIGURE 2.5 – Nombre d'organismes d'assurance agréés par l'ACPR (source)

Cette baisse d'organismes peut être induite par le renforcement progressif des contraintes au fil des années sur les fonds propres minimum, requis pour exercer leurs activités ce qui a engendré plusieurs fusions d'organismes.

Les ressources des organismes d'assurance complémentaire proviennent principalement des cotisations perçues par ses adhérents. Leurs charges se composent de prestations et de charges de gestion.

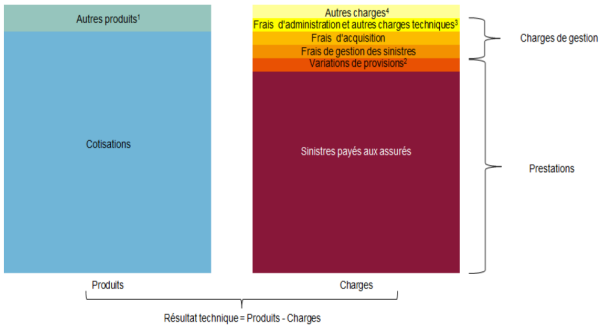


FIGURE 2.6 – Produits et charges des organismes d'assurance complémentaire (source)

Les autres produits peuvent être des produits financiers.

Les prestations sont réparties en sinistres payés et variations de provisions. En assurance santé, les provisions sont essentiellement des provisions pour sinistres à payer (PSAP, des sinistres survenus et en cours de gestion ou non encore déclarés que la mutuelle doit prendre en charge).

Les charges de gestion peuvent être réparties en une partie fixe et une variable. Cette répartition dépend du système de gestion adopté par l'organisme.

2.2.3 L'assurance santé et les organismes d'assurance complémentaire

En santé, une grande partie des dépenses en soins et biens médicaux nécessite l'intervention des organismes d'assurance complémentaire. Ci-dessous la part des dépenses prises en charge par les organismes complémentaires en 2020 pour les postes d'audioprothèses, d'optique et des prothèses dentaires.

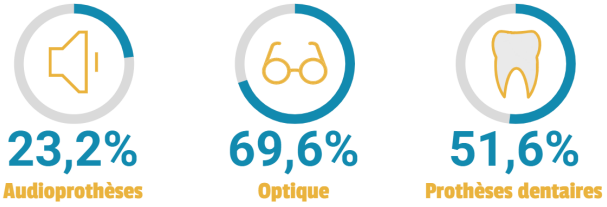


FIGURE 2.7 – Parts des dépenses prises en charge par les complémentaires santé en 2020 (source)

Les organismes d'assurance complémentaire couvrent la majeur partie des dépenses en optique et prothèses dentaires. Ces postes sont peu couvert par l'assurance maladie et nécessitent une couverture complémentaire.

La part de l'assurance santé, en termes de chiffre d'affaires, dans les activités des organismes d'assurance complémentaire diffère d'une catégorie d'organisme à une autre. En 2019, la répartition de l'activité par type d'organisme en 2019 (en % des cotisations collectées) se présente comme suit :

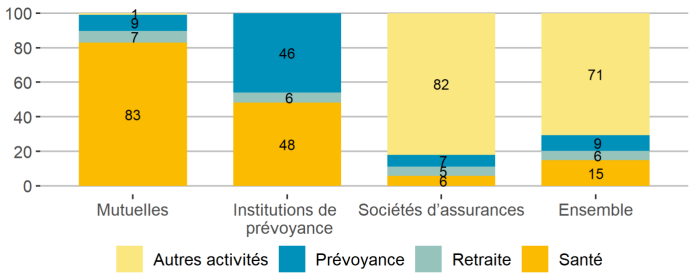


FIGURE 2.8 – Activités des organismes d'assurance complémentaire (source)

L'assurance santé représente 83% de l'activité des mutuelles santé en 2019 en termes de cotisations encaissées. A partir de la répartition de l'ensemble, on déduit que les mutuelles détiennent 41% en France, en termes de chiffre d'affaires, de l'activité santé devant les sociétés d'assurances qui en détiennent 37%.

2.3 La santé et la mutualité

2.3.1 Les mutuelles santé

Les mutuelles santé sont des sociétés de personnes à but non lucratif. Ces dernières exercent de l'assurance de personnes et sont régies par le Code de la mutualité.

Pour le remboursement des frais de santé, les mutuelles proposent des grilles de garanties sur plusieurs niveaux afin de satisfaire les besoins de l'ensemble de ses adhérents.

2.3.2 Grilles de garantie et modalités de remboursement

Une grille de garanties est un tableau contenant les éléments nécessaires expliquant le montant des remboursements de la mutuelle en cas de sinistres.

Les colonnes représentent généralement les niveaux des garanties (plus le niveau de la garantie est élevée et plus les taux de remboursement sont hauts) allant de la garantie de base à la garantie de luxe.
Les lignes représentent des groupements de type d'acte et sont souvent organisées par postes de soins. Les postes sont l'hospitalisation, les soins courants, l'optique, le dentaire, l'auditif, les actes H.N. et d'autres.

Pour un acte donnée, la part remboursée par la mutuelle est calculée à partir d'un paramètre dépendant de la nature de l'acte.

Le montant de dépenses pris en charge par la mutuelle peut être basée sur un **taux** de la **BRSS** (ou **ticket modérateur (TM)**), sous la forme d'un **forfait**, la combinaison des deux ou un taux des frais réelles.
Les forfaits peuvent êtres par acte, par jour, par an, par bénéficiaire ou une combinaison de ces derniers.

Prenons le cas d'une visite chez un ophtalmologue conventionné secteur 1 comme exemple illustratif. La BRSS est de 30€ et le taux de remboursement de la sécurité sociale est de 70%. Le prix de la consultation est supposé de 50€.

Cas d'une couverture à 150% BRSS

- Frais réels (FR) = 50€
- Dépassement honoraires = FR - BRSS = 50 - 30 = 20€
- TM = BRSS x (1 - Taux remboursement SS) = 30 x (1 - 0.7) = 9€
- Participation forfaitaire (PF) = 1€

  La participation forfaitaire est considérée incluse dans la partie prise en charge par la sécurité sociales (SS) lors du calcul du remboursement mutuelle.

- **Remboursement SS** = BRSS - TM - PF = 30 - 9 - 1 = 20
- **Remboursement mutuelle** = 150% x BRSS - Remboursement SS - PF = 45 - 20 - 1 = 24€
- **RAC** = FR - Remboursement SS - Remboursement mutuelle = 50 - 20 - 24 = 6€

On peut schématiser le calcul de la façon suivante :

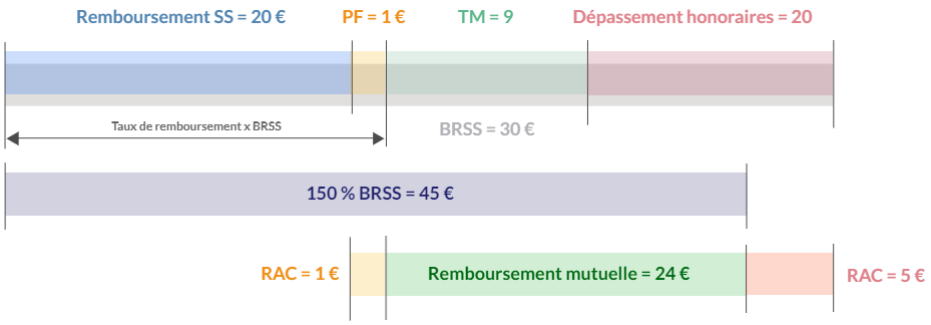


FIGURE 2.9 – Calcul du remboursement d'un acte en ophtalmologie

Le reste à charge pour le bénéficiaire est de 6€ (1€ de participation et 5€ dépassant le seuil de 150% de la BRSS).

2.4 Évolutions législatives

Durant les dernières années, l'assurance santé en France a connue plusieurs évolutions en matière de réglementation. L'arrivée du 100% santé et l'entrée en vigueur de la résiliation infra-annuelle ont bouleversé le système de santé français.

2.4.1 Contrats responsables

Les contrats responsables sont des contrats d'assurance santé qui respectent les conditions définies par le décret n° 2014-1374 du 18 novembre 2014.
Ils ont pour objectif l'amélioration du niveau minimal des garanties santé par l'encadrement des tarifs des professionnels de santé.

Ces contrats ont connu des évolutions en avril 2015 pour les contrats individuels et janvier 2016 pour les contrats collectifs. A compter de ces dates des obligations, que les organismes complémentaires doivent respecter, ont été fixées telles que :

- Mise en place de plafonds et planchers de remboursement ;
- Couverture de l'intégralité des tarifs de base de l'assurance maladie obligatoire pour les actes en soins courants, optique, hospitalisation et dentaire ;
- Couverture de l'intégralité des forfaits journaliers hospitaliers sans limitation de durée ;
- Couverture des prestations liées à la prévention ;
- Absence de remboursement de franchises et participations forfaitaires ;
- Pénalisation des actes hors parcours de soins.

Les contrats responsables présentent plusieurs avantages fiscaux pour les entreprises et leurs salariés tels qu'une taxe de solidarité à 13,27% à la place de 20,27%, ainsi qu'une exonération de charges sociales sur les parts de cotisations payées par l'employeur et une déduction fiscale des impôts sur le revenu concernant la part de cotisations à la charge des salariés.

Dans le même objectif de faciliter l'accès aux soins essentiels dont les restes à charges restent élevés, des paniers de soins en optique, auditif et prothèses dentaires dits " Paniers 100% santé" viennent s'ajouter aux contrats responsables avec la **réforme 100% santé**.

2.4.2 Réforme du 100% santé

Instaurée par le décret n° 2019-21 en 2019, la réforme 100% santé vise à proposer des paniers de prestations de base totalement pris en charge par les complémentaires santé (après le remboursement de l'assurance maladie obligatoire) pour les postes ayant le plus de reste à charge.
Trois postes sont concernés par cette réforme, soit le poste audiologie pour les prothèses auditives, le poste optique pour les lunettes et le poste dentaire pour les prothèses dentaires.

Cette mesure a été mise en place progressivement sur la période du 1er janvier 2019 au 1er janvier 2021 de la façon suivante :

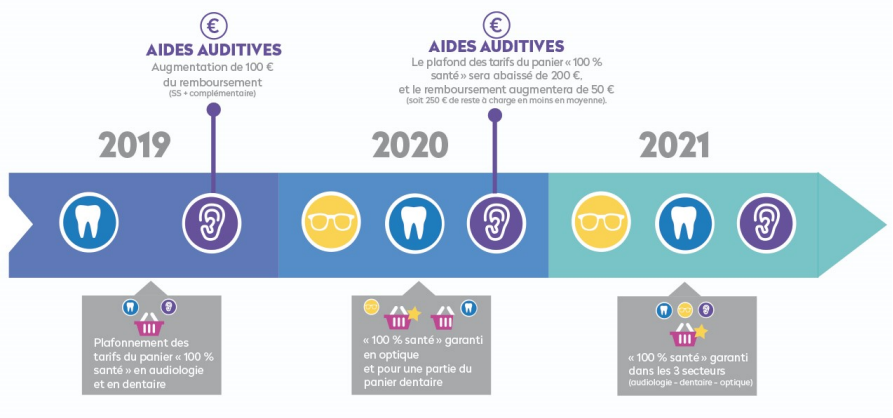


FIGURE 2.10 – Calendrier de mise en place du 100% santé (source)

Depuis le 1er janvier 2021, les paniers à zéro reste à charge sont garanties sur les trois postes pour la totalité des contrats responsables soit 95% des contrats complémentaires santé vendu en France.(source).

Une grande partie de la population ignorent encore l'existence de ce système selon une enquête menée par la DREES. Le graphe ci-dessous nous présente une estimation des % de la population ayants entendu parler du panier 100% santé par niveau de vie sur les année 2020 et 2021 :

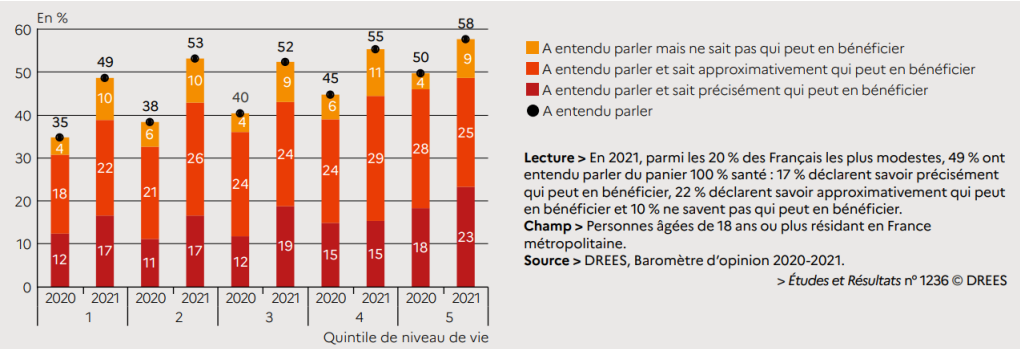


FIGURE 2.11 – Connaissance du panier 100 % santé par quintile de niveau de vie (source)

La réforme 100% commence à rentrer progressivement dans la vie des Français. Malgré qu'entre les années 2020 et 2021, seuls 10% de plus des Français ont entendus parler de la réforme, prêt de la moitié de la population majeure reste encore dans l'ignorance.
On remarque que la part de l'ignorance de la réforme est plus importante chez les Français les plus modestes, cela peut s'expliquer par la facilité d'accès à l'information par quintile de niveau de vie.

Depuis le 1er janvier 2022 les complémentaires santé sont dans l'obligation de faire bénéficier à leurs assurés d'un système de tiers payant pour les actes en 100% santé dans le but de lutter contre le renoncement aux soins.

2.4.3 Résiliation infra-annuelle

Instaurée par le décret n° 2020-1438 du 24 novembre 2020 relatif au droit de résiliation sans frais de contrats de complémentaire santé et mise en œuvre le 1er décembre 2020, la résiliation infra-annuelle permet aux assurés la possibilité à résilier à tout moment de l'année après un an d'ancienneté. Cette réforme concerne les contrats couvrant les risques liés à la santé.

Avant la résiliation infra-annuelle, les assurés n'avaient qu'un délai de 3 mois à partir de la date d'anniversaire de leurs contrats pour faire une demande de résiliation avant tacite reconduction. Certains assurés devaient attendre cette date malgré des besoins, induits par un changement de leurs situations, qui nécessite le changement de leurs garanties comme l'arrivée d'un nouveau-né ou la contraction d'une maladie. Cette réforme répond en partie à ces besoins.

Les modalités de la résiliation infra-annuelle ont été revues par le décret n°2022-388 du 17 mars 2022 relatif au fonctionnement des mutuelles et unions et aux institutions de prévoyance pour permettre la résiliation infra-annuelle aux contrats santé comportant des garanties liées à la perte d'autonomie.

2.5 Cadre de l'étude

Les différentes évolutions qu'a connu le marché de la santé en France ont eu un impact non négligeable sur le comportement des assurés en France.
Les mutuelles santé, dont les méthodes de tarification ne prennent pas toujours en compte l'effet du 100% santé et la résiliation infra-annuelle, peuvent rencontrer des difficultés dans l'estimation de la consommation de leurs portefeuilles.

D'un côté, la consommation d'un assuré peut changer avec la mise en place de garanties à zéro reste à charge ainsi que des systèmes de tiers payant permettant de ne pas à avoir à avancer les frais. En effet, en absence de contraintes pour l'accès aux soins, les assurés peuvent avoir recours aux soins plus fréquemment.
De l'autre côté, la résiliation infra-annuelle expose les assurés à un marché plus concurrentiel et présente un risque d'anti-sélection pour certaines mutuelles. En conséquence, la composition du portefeuille peut changer avec une hausse de la part des mauvais risques et des départs plus fréquents.

Dans ce contexte, nous étudions le portefeuille santé individuel d'une mutuelle santé afin de prédire au mieux l'évolution de sa composition et de sa consommation à horizon 1 an. La consommation du portefeuille dépend essentiellement de l'évolution de sa composition et du comportement des assurés. La modélisation de l'évolution du portefeuille, en terme d'effectifs et profils de risques, est donc essentielle à la prédiction.

Une modélisation de l'évolution des effectifs d'un portefeuille doit prendre en compte les flux d'entrées et de sorties des contrats. La modélisation des deux flux représentant deux sujets d'études différents, nous avons décidé de dédier ce mémoire aux flux sortants des contrats, soit la modélisation de l'évolution d'un portefeuille fermé en assurance santé.

Disposant des données d'un portefeuille fermé, sur période d'un an passé, à une date t, la modélisation de l'évolution à l'horizon d'un an est réalisée via un processus de prédictions formé de deux parties principales.

La première partie concerne l'évolution des effectifs du portefeuille dans le temps. Pour cela, nous allons commencer par estimer les probabilités de résiliation des contrats à l'horizon d'un an. Ayant la liste des contrats qui risquent de résilier dans l'année, nous prédisons pour chaque contrat la durée restante avant résiliation.

La seconde partie consiste à modéliser la consommation des contrats pendant leurs périodes de couverture restantes probables. Cette prédiction est réalisée par la prédiction des coûts moyens et fréquences des bénéficiaires de chaque contrat prenant en compte la probabilité de résiliation et la durée de couverture restante probable des contrats.

Chapitre 3

Construction de la base de données

Ce chapitre présente les démarches entrainant la conception des bases de données et la structure d'un modèle permettant la prédiction de l'évolution d'un portefeuille fermé en assurance santé individuelle. Cette évolution du portefeuille se traduit par l'évolution de son effectif et de sa consommation au cours du temps.

3.1 Conception du modèle

La modélisation de l'évolution du portefeuille a nécessité la mise en place d'un processus à deux étapes principales qui se présente comme suit :

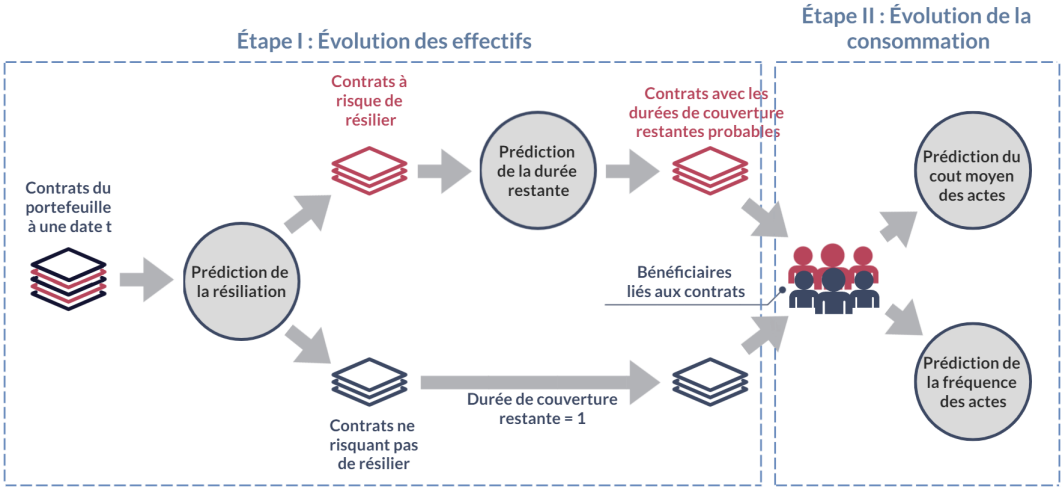


FIGURE 3.1 – Schéma du processus de modélisation

La première étape concerne l'évolution de l'effectif du portefeuille. En observant le portefeuille à une date t, on souhaite connaître si un assuré risque de résilier dans l'année qui suit et sa durée de couverture restante dans le cas d'une résiliation probable. Dans cette partie, les prédictions sont faites par contrat puisque la résiliation d'un assuré

concerne, dans la majorité des cas, l'ensemble des bénéficiaires couverts par le même contrat.

Une fois que la durée de couverture restante, à un horizon d'un an, est connue pour chaque contrat, on passe aux données relatives à chaque bénéficiaire pour prédire sa consommation durant sa période de couverture restante par la prédiction du coût moyen de ses actes et de sa fréquence de consommation. Ce modèle permet la détection d'un changement comportemental relatif à une résiliation probable.

La prédiction des variables à expliquer, à une date donnée d'observation du portefeuille, nécessite la liaison des évènements antérieurs à cette date aux valeurs prises par ces variables sur une durée postérieure.

La construction des bases de données est donc réalisée par la concaténation de données provenant d'une succession d'observations mensuelles du portefeuille sur la période d'activité étudiée. Les observations sont faites le premier de chaque mois. Les données explicatives, respectivement à expliquer, concernent une période d'un an avant, respectivement un an après, la date d'observation, comme le montre le graphe si dessous :

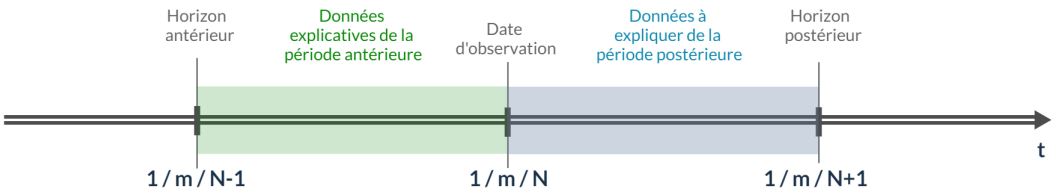


FIGURE 3.2 – Périodes d'observations du portefeuille

L'étape suivante de l'étude est donc la prise en compte des données communiquées par la mutuelle pour la construction des variables explicatives pertinentes pour les prédictions.

3.2 Bases de données à disposition et RGPD

Les données relatives au portefeuille sont des données personnelles et de santé, la mutuelle est donc dans l'obligation de respecter le **Règlement général sur la protection des données (RGPD)** lors de la transmission des données.

3.2.1 Règlement général sur la protection des données (RGPD)

Entrée en vigueur le 25 mai 2018, cette loi européenne vient s'ajouter à la Loi Française Informatique et Libertés de 1978 et concerne tout organisme, établis sur le territoire de l'Union Européen ou dont l'activité concerne des résidents européens, exerçant de la collecte et/ou traitement de données indépendamment de sa taille et son secteur d'activité.

Cette dernière vise à protéger les données personnelles en renforçant les responsabilités des organismes les manipulant.

Les **données personnelles** sont les données depuis lesquelles on peut identifier directement ou indirectement une personne physique. Dans notre cas, l'identification directe correspondrait, par exemple, aux variables noms et prénoms des assurés alors que l'identification indirecte correspondrait à la présence de leurs numéros de sécurité sociale ou adresses postales...

Afin de respecter le RGPD, il est recommandé pour la mutuelle de pseudo-anonymiser les bases de données relatives au portefeuille.

La **pseudonymisation** des données consiste à remplacer les données d'identification directe par des références indirectement identifiantes. Cette méthode est réversible contrairement aux méthodes d'anonymisation.

Les bases de données transmises ne devront donc pas contenir des données permettant de remonter à l'identité d'un assuré du portefeuille.

3.2.2 Bases de données à disposition

Trois bases de données nous ont été transmises par la mutuelle : une base "effectif", une base "cotisation" et une base "prestations". Les variables retenues des bases de données transmises se présentent comme suit :

Effectifs	**Cotisations**	**Prestations**
Produit	Produit	Produit
Garantie technique	Garantie technique	Garantie technique
Numéro contrat individuel	Numéro contrat individuel	Numéro contrat individuel
Numéro bénéficiaire	Numéro bénéficiaire	Numéro bénéficiaire
Offre	Offre	Offre
Date ancienneté contrat	Exercice	Date de paiement
Date ancienneté bénéficiaire	Cotisation HT	Date début soins
Date naissance	Taxe	Date fin soins
Sexe		Code acte
Type bénéficiaire		Famille acte
Situation familiale		Taux remboursement SS
Code postal		Taux remboursement mutuelle
Département tarifaire		Frais réels
Date effet adhésion		Remboursement SS
Date effet radiation		Ticket modérateur
		Remboursement mutuelle

Les numéros bénéficiaire et numéros contrat individuel sont des références attribuées par la mutuelle aux assurés et contrats les liant et ne nous permettent en aucun cas l'identification d'un assuré du portefeuille.

La consommation, les cotisations perçues et les caractéristiques de chaque assuré du portefeuille est identifiable par les cinq variables suivantes :

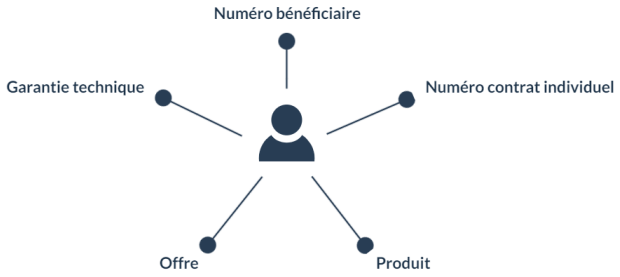
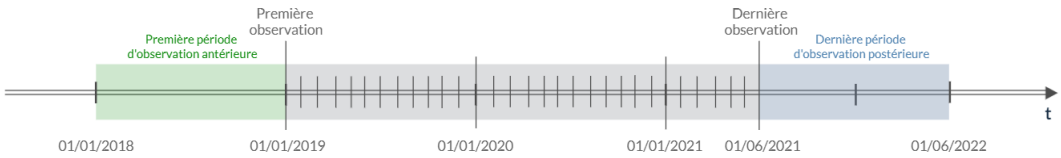


FIGURE 3.3 – Clé primaire des bases de données

A partir des bases à disposition, on va pouvoir fournir aux modèles les variables décrivant l'activité, antérieure aux dates d'observation, de chaque contrat et assuré.

3.3 Construction des bases de données

Les données transmises par la mutuelle correspondent à la période du **1 janvier 2018** au **6 mai 2022**. Vu la contrainte sur les horizons antérieurs et postérieurs, les observations mensuelles du portefeuille seront du **1 janvier 2019** au **1 mai 2021**.



3.3.1 Variables explicatives

Les variables décrivant l'activité antérieure d'un assuré par rapport à une date d'observation donnée se décomposent de la manière suivante :

— Variables relatives aux **profils des assurés** : Âge, sexe, type, code postal, ...
— Variables relatives aux **contrats** : Gamme, garantie, cotisations, taxes, ...
— Variables relatives à la **consommation** : Remboursement mutuelle par poste d'actes, nombre d'actes, ... (sommées sur une année antérieure à la date d'observation.)

Les variables qui varient en fonction du temps telles que l'âge du bénéficiaire, l'ancienneté du contrat, l'ancienneté du bénéficiaire, ... sont calculées en fonction de la date d'observation. Pour une meilleure interprétation des modèles, il a été nécessaire d'ajouter quelques variables caractérisant la situation de chaque contrat et assuré.

Niveaux des gammes et garanties

Chaque combinaison des variables "Garantie technique", "Produit" et "Offre" réfère à une **garantie** d'une **gamme** commercialisée par la mutuelle. Sur la période étudiée, les contrats souscrits dans le portefeuille réfèrent à 91 garanties de 17 gammes différentes. Pour une question de confidentialité, les gammes sont nommées de G1 à G17.

Une classification par Niveau de gamme et Niveau de garantie permettait une meilleure interprétabilité et simplification de ces variables.

La classification des garanties et gammes a été basée sur la tarification de ces dernières. En effet, Le nombre de contrats souscrits pour certaines garanties n'était pas suffisant pour une classification basée sur la consommation et ne donnait pas des résultats cohérents avec leurs niveaux réels de remboursement.

La tarification des garanties pour chaque gamme est faite par assuré et dépend des variables suivantes :

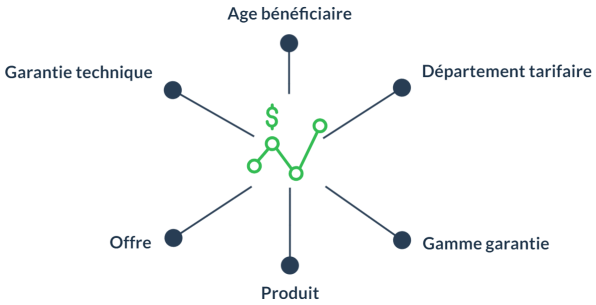


FIGURE 3.4 – Variables prises en compte dans la tarification de la mutuelle

La première étape de la classification était de faire une moyenne par âge des cotisations annuelles perçues par un assuré afin de supprimer les variations dues aux autres variables. Le nombre d'assurés par gamme par exercice se présente comme suit :

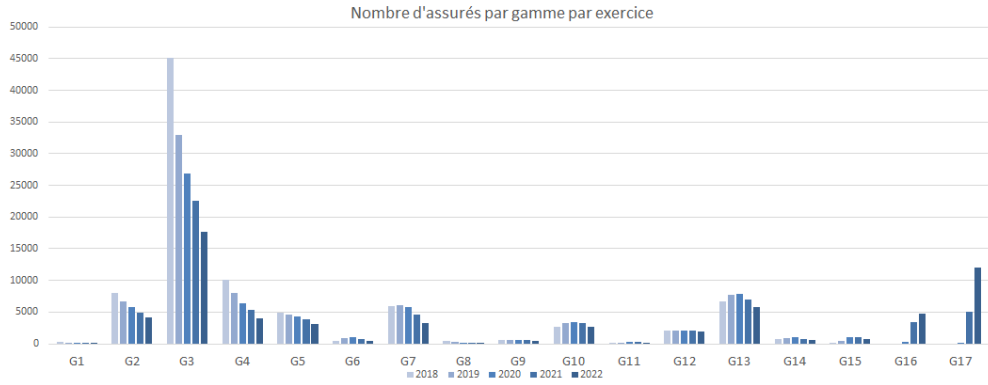


FIGURE 3.5 – Nombre d'assurés par gamme par exercice

Les cotisations annuelles moyennes par gamme en fonction de l'âge des assurés se présentent comme suit :

FIGURE 3.6 – Cotisations annuelles moyennes par âge par gamme

La méthode de classification adoptée consiste à localiser les allures des courbes de cotisations moyennes à l'âge 60 et attribuer les classes de la manière à ce qu'une classe c corresponde à une gamme dont la cotisation annuelle moyenne pour un assuré âgé de 60 ans appartient à l'intervalle $](c-1) \times 100, c \times 100]$.
Par exemple, une gamme de classe 7 correspond à une gamme dont la cotisation moyenne annuelle, pour un assuré âgé de 60, est entre 600 et 700 euros.

On obtient la classification des gammes suivante :

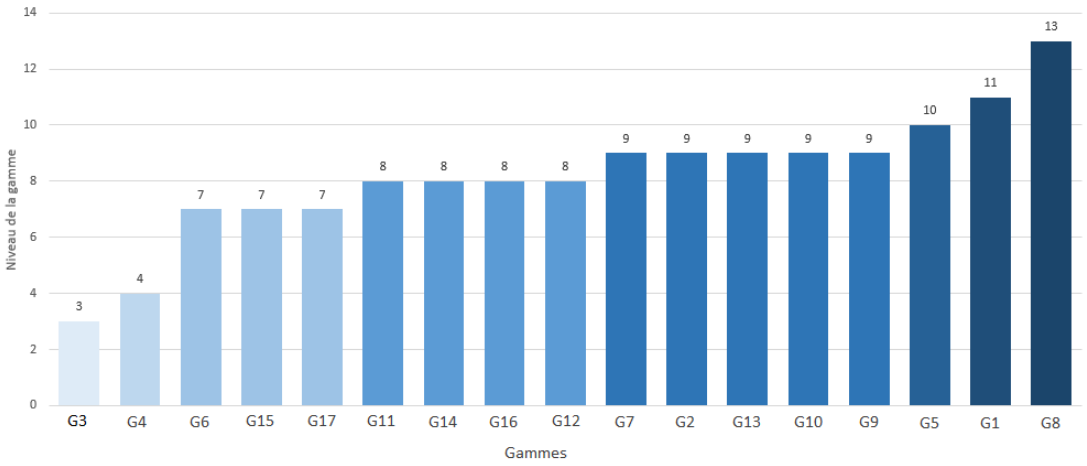


FIGURE 3.7 – Niveaux des gammes

Le choix de l'âge de 60 ans vient du fait que c'est l'âge le plus représenté dans le portefeuille et la majorité des gammes. De plus, la totalité des courbes sont superposées et ne croisent pas aux alentours de cet âge.

Cette méthode de classification nous permet de prendre en compte le réel écart entre les niveaux. Par exemple, cela permet de classer la gamme G6 à trois niveaux au-dessus de la gamme G4 alors qu'il n'existe pas de gammes intermédiaires.

Avec la même méthode, on classe les garanties de chaque gamme. L'application sur un exemple de quatre gammes se fait comme suit :

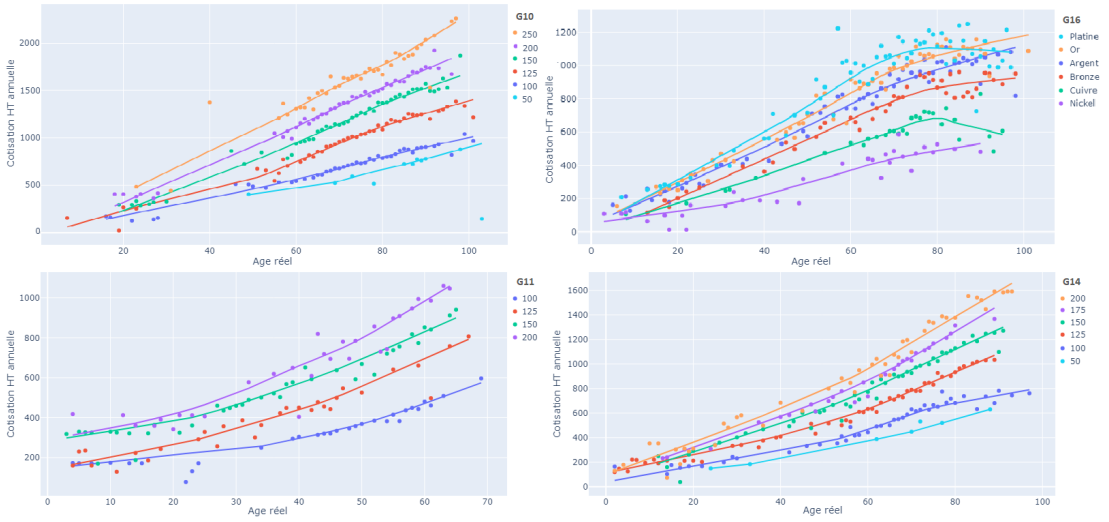


FIGURE 3.8 – Cotisations annuelles moyennes par âge par garantie pour les gammes G10, G11, G14 et G16

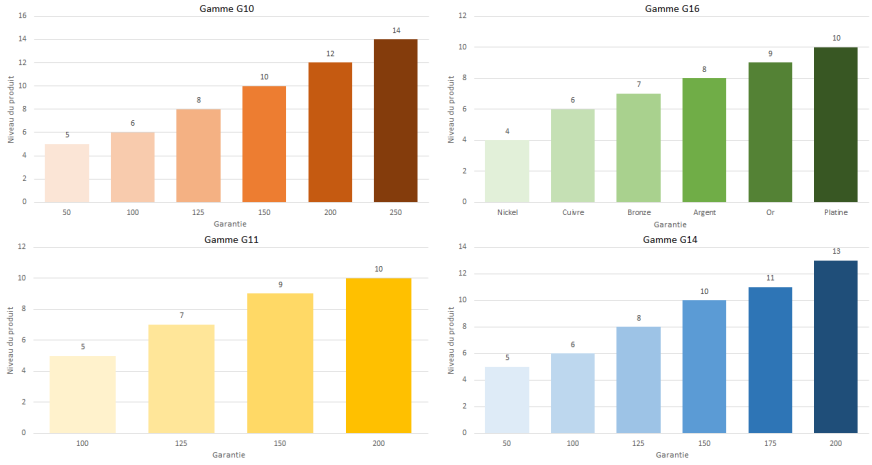


FIGURE 3.9 – Niveaux de garanties des gammes G10, G11, G14 et G16

L'attribution de niveaux aux gammes et garanties nous permet de remplacer les variables "Garantie technique", "Offre" et "Produit".

Certaines gammes ne sont plus commercialisées et l'appartenance d'un contrat à ces gammes peut être un facteur significatif pour la probabilité de résiliation en raison de la commercialisation de nouvelles gammes concurrentes au sein de la mutuelle, il a été donc préférable de garder la variable "Gamme" en tant que variable qualitative.

Cotisations et indexations annuelles

Le montant que cotise un assuré représente un facteur direct déterminant pour la probabilité de résiliation et la consommation postérieure. En effet, le paiement d'une prime élevée répond naturellement à un besoin de l'assuré et dans le cas d'une faible consommation une résiliation probable. Les cotisations payées par un assuré entre le début de l'exercice et la date d'observation sont donc rapportées à une **cotisation annuelle** et intégrées dans la base.

De même, les niveaux d'**indexations annuelles** antérieures et postérieures aux observations sont aussi intégrés. Ces indexations sont supposées appliquées au premier jour de chaque année.

Pour une date d'observation le 1/m/N, les variables d'indexation se présentent comme suit :

$$\textbf{Indexation antérieure} = \frac{Cotisation\ annuelle(N) - Cotisation\ annuelle(N-1)}{Cotisation\ annuelle(N-1)}$$

$$\textbf{Indexation postérieure} = \frac{Cotisation\ annuelle(N+1) - Cotisation\ annuelle(N)}{Cotisation\ annuelle(N)}$$

La variable **Indexation postérieure** nous permet un certain dynamisme des modèles. En effet, les prédictions peuvent être orientées en fonction de l'indexation postérieure, par le biais de simulations, afin d'analyser l'impact de la valeur d'indexation sur le flux de contrats sortant probable.

Une variable concernant le temps restant entre la date d'observation et la prochaine indexation vient compléter les dernières variables. Pour une observation le 1/m/N, cette valeur prend la durée (en année) entre le 1/m/N et le 1/1/N+1.

Résiliation infra-annuelle et éligibilité à résilier

La résiliation infra-annuelle ne permet la résiliation des contrats qu'à partir d'un an d'ancienneté. A une date d'observation donnée, certains des assurés n'auront pas atteint un an d'ancienneté et ne pourrons résilier leurs contrats qu'à partir d'une durée donnée de la date d'observation.

Cependant, notre première observation du portefeuille date du 1 janvier 2019 alors que la résiliation infra-annuelle n'entre en vigueur qu'à partir du 1 décembre 2020. Avant

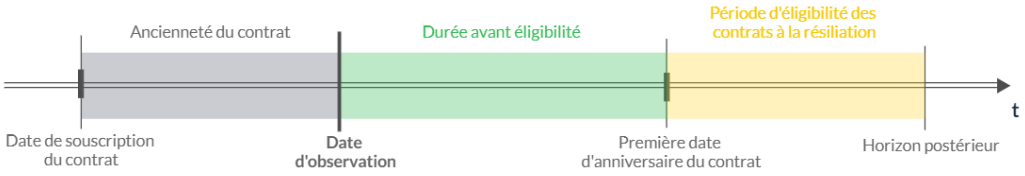
cette date, les assurés n'avaient que 2 mois à partir de la date d'anniversaire pour résilier leurs contrats.

Dans le but de fournir aux modèles l'information sur l'éligibilité des adhérents à résilier, nous créons une variable nommée "Durée avant éligibilité" qui donne la durée avant laquelle l'adhérent peut résilier son contrat selon la date d'observation.
Cette variable est calculée en année et dépend de la date d'observation.

Après le 1 décembre 2020 :

Après l'entrée en vigueur de la résiliation infra-annuelle, le calcul de la durée avant éligibilité dépend uniquement de l'ancienneté des contrats.

Pour une date d'observation donnée, cette variable vaut 0 pour les contrats d'ancienneté supérieure à 1 an (donc éligibles à résilier à tout moment).
Dans le cas d'une ancienneté inférieure à 1 an à la date d'observation, le calcul de cette variable se présente comme suit :



$$\begin{cases} DAE = 0 & si\ AC \geqslant 1 \\ DAE = 1 - (AC - \lfloor AC \rfloor) & si\ AC < 1 \end{cases}$$

Avec,

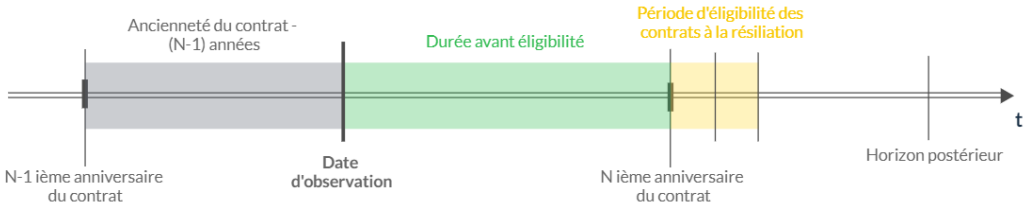
DAE : Durée avant éligibilité
AC : Ancienneté contrat

$\lfloor AC \rfloor$ désigne la partie entière inférieure de l'ancienneté du contrat.

Avant le 1 décembre 2019 :

Un an avant l'entrée en vigueur de la résiliation infra-annuelle (tant que notre horizon d'observation ne dépasse pas le 1 décembre 2020), le calcul de la durée avant éligibilité se base uniquement sur la date d'anniversaire des contrats.

À la date d'anniversaire du contrat, une période d'éligibilité à la résiliation est lancée et dure 2 mois, la durée avant éligibilité pour les observations incluses dans cette période est donc de 0. Pour les observations avant la date d'anniversaire du contrat, la durée avant éligibilité se présente comme suit :



$$\begin{cases} DAE = 0 & si \quad 0 \leqslant AC - \lfloor AC \rfloor \leqslant \frac{2}{12} \\ DAE = 1 - (AC - \lfloor AC \rfloor) & si \quad \frac{2}{12} < AC - \lfloor AC \rfloor < 1 \end{cases}$$

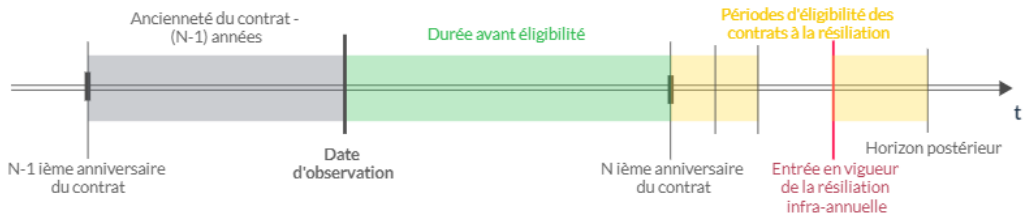
Entre le 1 décembre 2019 et le 1 décembre 2020

Sur cette partie, la résiliation infra-annuelle entre en vigueur au milieu de la période postérieure observée. Les contrats ayant donc une ancienneté de plus d'un an à la date d'entrée en vigueur de la réglementation pourront désormais résilier. Le calcul de la durée avant éligibilité dépend de trois cas possibles.

Les deux premiers cas concernent les contrats ayant une ancienneté supérieure à un an à la date d'entrée en vigueur de la résiliation infra-annuelle.

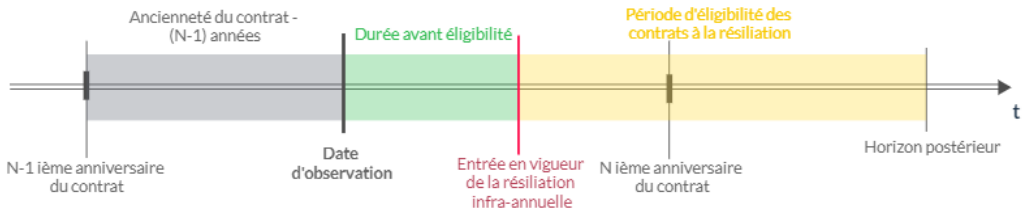
Cas n° 1 :

Le premier cas correspond aux contrats dont la date d'anniversaire est avant le 1 décembre 2020. La durée avant éligibilité serait donc la durée entre la date d'observation et la date anniversaire du contrat.



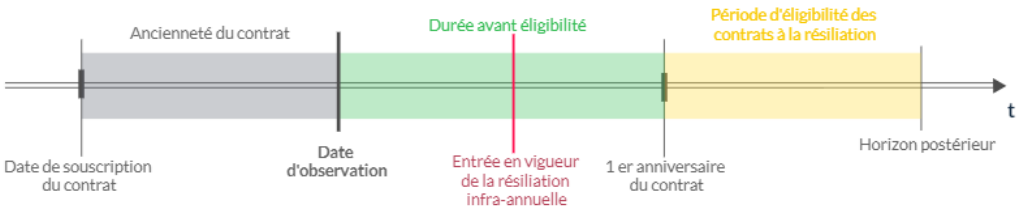
Cas n° 2 :

Le deuxième cas correspond aux contrats dont la date d'anniversaire est après le 1 décembre 2020. La durée avant éligibilité serait donc la durée entre la date d'observation et la date d'entrée en vigueur de la résiliation infra-annuelle.



Cas n° 3 :

Le troisième et dernier cas correspond aux contrats dont l'ancienneté est inférieure à un an au 1 décembre 2020. La durée avant éligibilité serait donc la durée entre la date d'observation et la date du premier anniversaire du contrat.



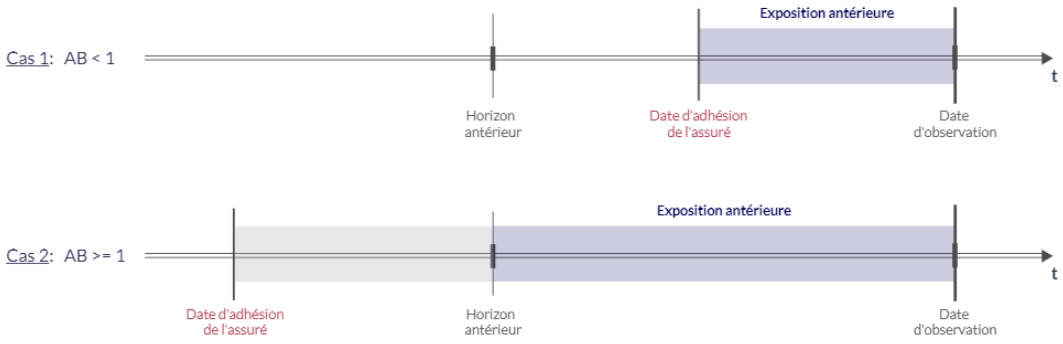
La variable durée avant éligibilité permet d'intégrer le seuil à partir duquel les adhérents pourront résilier. La prédiction de la durée de couverture restante des contrats dépendra fortement de cette variable car cette dernière serait toujours supérieure à la durée avant éligibilité. On s'attend aussi à ce que la probabilité de résiliation décroît en fonction de cette variable.

Avant de passer aux variables liées à la consommation des assurés, on intègre une variable nommée **Exposition antérieure**. Cette variable retourne la durée de couverture de l'assuré sur la période antérieure observée et est égale à :

$$\begin{cases} EA = AB & si \quad AB < 1 \\ EA = 1 & si \quad AB \geqslant 1 \end{cases}$$

Avec,

EA : Exposition antérieure
AB : Ancienneté bénéficiaire

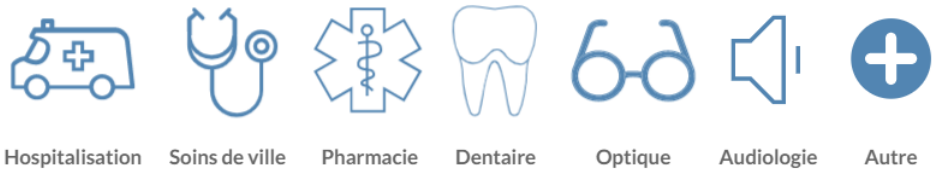


Cette variable est essentielle pour le calcul de fréquence des prestations antérieurs dans la suite de ce mémoire.

Maintenant qu'on dispose des variables décrivant les assurés et leurs contrats, on passe à l'intégration de leurs consommations antérieures aux dates d'observation.

Consommations antérieures des assurés

On distingue 7 grands postes sur lesquelles les prestations sont réparties. Ces grands postes représentent des regroupements d'actes par type de soins et sont les suivants :



Les prestations en 100% santé sont distinguées des autres prestations et classées dans les postes suivants : Dentaire RAC0, Optique RAC0 et Auditif RAC0 afin de les étudier séparément.

Pour chacun de ces postes, nous avons pris en compte les variables suivantes :

— Remboursements antérieurs
— Nombre d'actes
— Reste à charge

Ces variables sont sommées sur l'exposition antérieure observée des assurés. À partir de ces dernières, nous avons ajouté les variables :

— Coût moyen antérieur = Remboursements antérieurs / Nombre d'actes antérieurs
— Fréquence antérieure = Nombre d'actes antérieurs / Exposition antérieure

Regroupement des données par contrat

L'étape suivante consiste à rassembler les données par contrat pour la prédiction de leur résiliation et durée de couverture restante. Pour un contrat, les données prises en compte sont les suivantes :

Lors de chaque observation, seuls les contrats d'au moins un mois d'ancienneté et un mois de durée de couverture restante sont pris en compte.

Une fois les données antérieures rassemblées, on procède à la formation des bases de données qui vont servir aux prédictions par l'ajout des variables explicatives.

3.3.2 Variables à expliquer

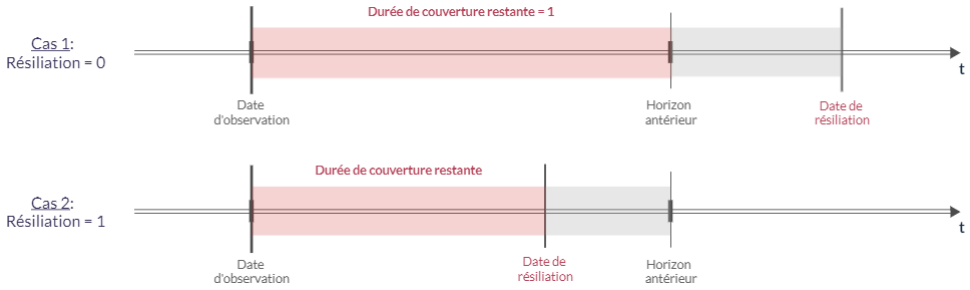
Les variables à prédire pour l'étape I du modèle globale sont la résiliation et la durée de couverture restante. Ces variables sont attribuées aux bases par contrat et définies comme suit :

$$\textbf{Résiliation} = 1_{\{Date\ resiliation\ <\ Date\ horizon\}}$$

$$\textbf{Durée de couverture restante} = \begin{cases} 1 & si \quad Resiliation = 0 \\ DR\ -\ DO & si \quad Resiliation = 1 \end{cases}$$

Avec,

DR : Date de résiliation
DO : Date d'observation



La prédiction de la consommation postérieure dans la partie II se fait par bénéficiaire, les variables à prédire sont pour un poste d'actes i donnée :

- **Coût moyen postérieur (i)** = Remboursements postérieurs (i) / Nombre d'actes postérieurs (i)
- **Fréquence postérieure (i)** = Nombre d'actes postérieurs (i) / Durée de couverture restante

Ces variables sont calculées sur la période de couverture postérieure à l'observation et avec un horizon d'un an.

Enfin, dans le but d'exploiter la totalité des informations à disposition, des dates d'observation ultérieures au 1 mai 2021 ont été prises en compte pour les cas de résiliation avant le 6 mai 2022. En effet, dans le cas d'une résiliation l'horizon se limite à la date de résiliation.

Chapitre 4

Analyse descriptive des données

L'analyse descriptive des données permet une meilleure compréhension des variables dans le but de sélectionner les modèles les plus adéquats aux prédictions.

Ce chapitre permet dans un premier temps de découvrir le profil des assurés et des contrats sur lesquels est basée l'étude ainsi que leur consommation tout au long de leurs périodes de couverture observées.
Dans un second temps, l'attention sera dirigée vers les variables à expliquer afin d'étudier leurs distributions et comportements selon les différentes variables explicatives.

4.1 Présentation du portefeuille

Les contrats et assurés figurant dans les bases de données construites dans le chapitre précédant sont issues d'observations mensuelles du portefeuille sur la période du 1 janvier 2019 au 1 mai 2021 avec des horizons antérieurs et postérieurs d'un an. Le nombre de contrats et bénéficiaires concernés sur les exercices 2019 à 2022 se présente comme suit :

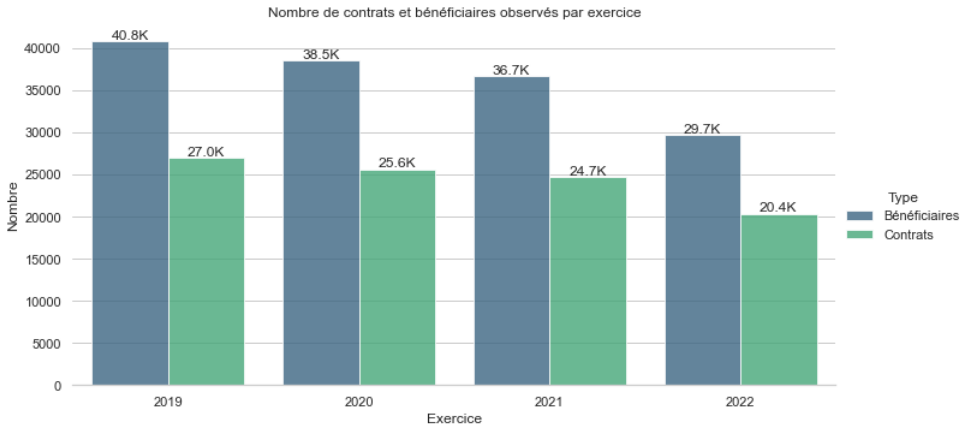


FIGURE 4.1 – Nombre de contrats et bénéficiaires par exercice

La forte baisse observée en 2022 vient du fait que les contrats et assurés ayant moins d'un an d'ancienneté au 6 mai 2022 (Limite de nos données) ne sont pas pris en compte vu qu'ils ne sont pas encore été concernés par la résiliation infra-annuelle à la fin de la période étudiée. Ils n'ont donc pas la possibilité de résilier.

L'évolution de la répartition des assurés par type de bénéficiaire par exercice est la suivante :

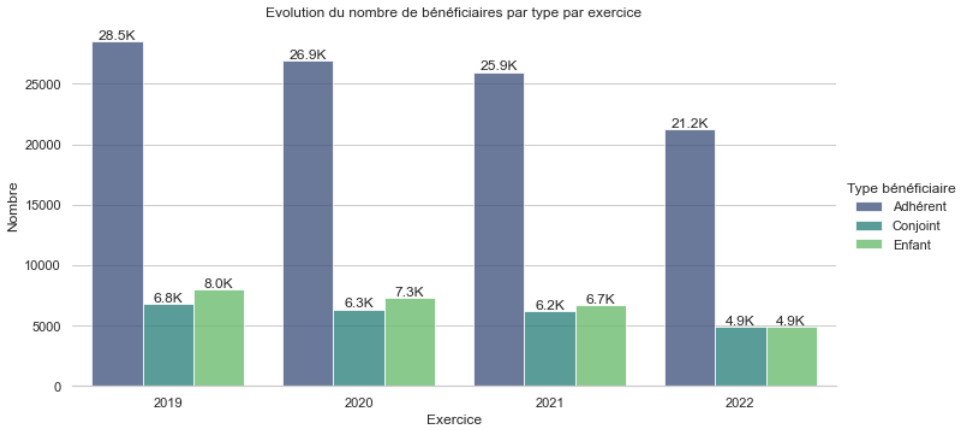


FIGURE 4.2 – Répartition des assurés par type de bénéficiaire par exercice

La proportion des enfants ne cesse de diminuer avec le temps passant de 18,5% en 2019 à 15,8% en 2022. Quant à la proportion des conjoints, elle se maintient entre 15,7% et 15,8%. Cette baisse pourrait donc être induite par la baisse des contrats couvrant des familles.

L'évolution de la pyramide des âges des bénéficiaires en fonction de l'exercice se présente comme suit :

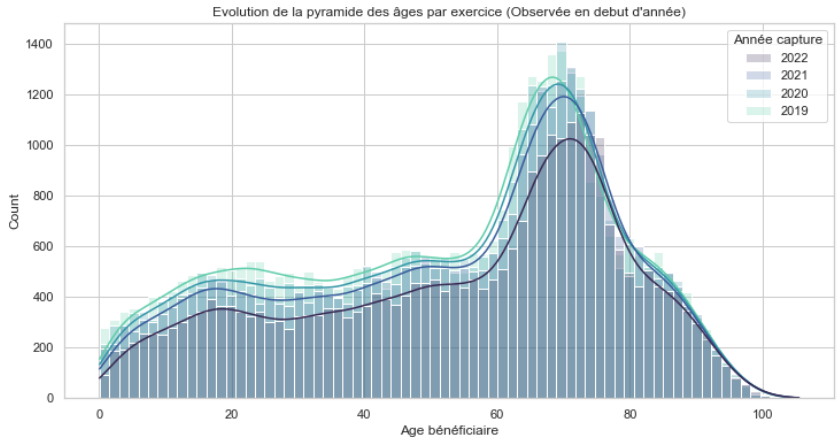


FIGURE 4.3 – Évolution de la pyramide des âges

Une baisse du nombre de personnes protégées concernant l'ensemble des âges est observée. Le traçage de l'évolution de la distribution des âges permet l'identification des âges les plus concernés par cette baisse.

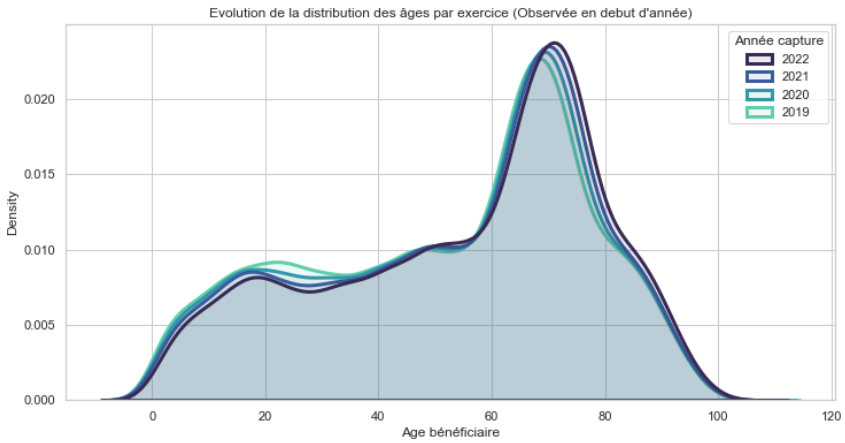
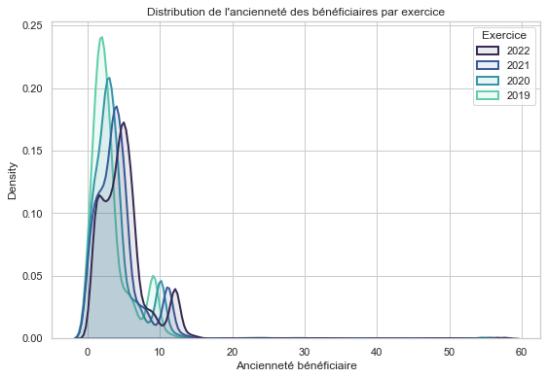


FIGURE 4.4 – Évolution de la distribution des âges

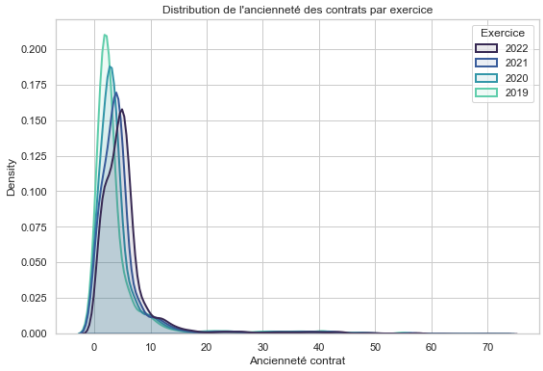
La déformation de la distribution des âges résulte de la sortie des assurés du portefeuille (déformations verticales) et le vieillissement de la population assurée (translation horizontale de la distribution). La tranche d'âges la plus impactée par la baisse est celle des jeunes d'entre 20 et 35 ans (en phase avec la baisse de la proportion des enfants vue précédemment).

La distribution des assurés les plus âgés se maintient au cours des précédents exercices, la fidélité de certains assurés peut être un élément clé de ce maintien.

Les distributions d'anciennetés des bénéficiaires et contrats par exercice sont représentées ci-dessous :



(a) Distribution de l'ancienneté des bénéficiaires par exercice



(b) Distribution de l'ancienneté des contrats par exercice

La distribution de l'ancienneté des bénéficiaires est constituée de pics faisant références à des lancements de gammes. Par exemple, le pic au niveau de l'ancienneté 10 ans pour l'exercice 2019 est formé par les bénéficiaires des gammes G1 et G2. (voir l'ancienneté des bénéficiaires par exercice en annexe).

L'évolution de cette distribution au cours des exercices fait apparaître en 2022 un troisième pique, au niveau de l'ancienneté 1,5 ans. Ce pique est relatif au lancement des gammes G16 et G17 en 2020.

La distribution de l'ancienneté des bénéficiaires diffère de celle des contrats. Cette différence est issue des changements de contrats au sein de la mutuelle. En effet, l'absence d'un pique au niveau de 10 ans d'ancienneté contrat contrairement aux anciennetés bénéficiaires nous permet de conclure que la majorité des bénéficiaires issus des gammes G1 et G2 ont désormais changé de gamme.

L'évolution du nombre de bénéficiaires par niveau de gamme se présente comme suit :

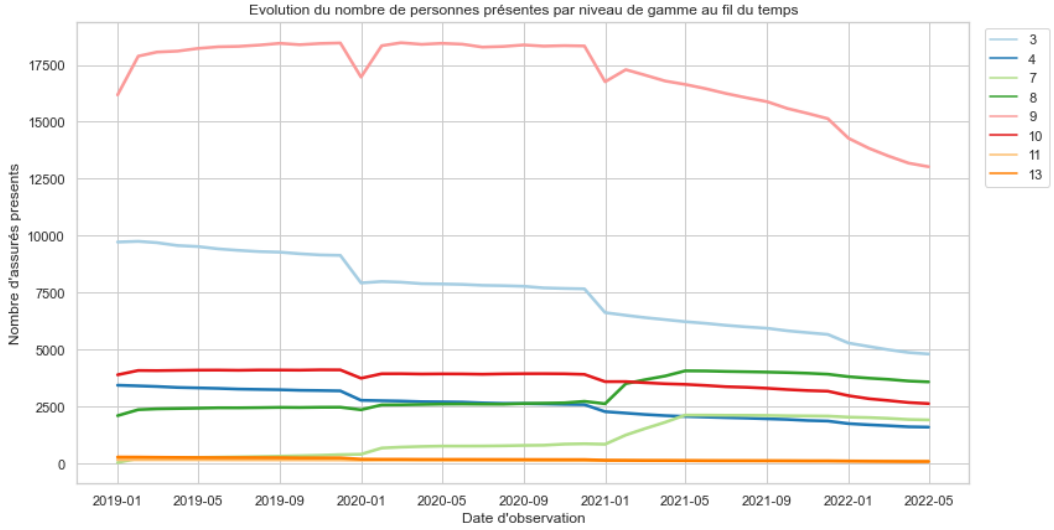


FIGURE 4.6 – Évolution du nombre de bénéficiaires par niveau de gamme

Ce graphique peut être décomposées en deux parties correspondant à un avant et un après l'entrée en vigueur de la résiliation infra-annuelle en décembre 2020.

Avant le 1 décembre 2020, nous observons des paliers annuels venant du fait que la majorité des dates d'anniversaire des contrats étaient en début d'année. Nos observations prenant uniquement les contrats ayant un minimum d'un mois de durée de couverture restante et d'ancienneté, des creux se forment au premier de chaque année reflétant la proportion des bénéficiaires résiliant et souscrivant chaque année.
La majorité des courbes se maintiennent en milieu d'année sauf pour les gammes G3 et G4 qui ne sont plus commercialisées par la mutuelle et qui subissent une baisse de leurs effectifs au cours du temps.

Après le 1 décembre 2020, les allures des courbes changent et subissent des variations continues dans le temps en conséquence de l'entrée en vigueur de la résiliation infra-annuelle.
La hausse d'effectif des niveaux 7 et 8, liée au lancement de nouvelles gammes, s'arrête au 5 mai 2021 en raison de l'ancienneté minimale d'un an, à l'horizon 1 mai 2022, imposée par le modèle. Les nouvelles souscriptions de contrats après le 5 mai 2021 ne sont pas pris en compte car ces contrats ne sont pas éligibles à la résiliation sur la période étudiée. À partir de cette date, seules les résiliations sont prises en compte.

De légers changements de pente sont encore observés à la fin de l'année 2021. Ces rassemblements de résiliations à cette date peut être dû aux anciennes habitudes des assurés à résilier aux dates d'anniversaire ou à l'ignorance de la nouvelle réglementation concernant la résiliation infra-annuelle.

L'évolution de la répartition du nombre de personnes protégées par niveau de gamme est la suivante :

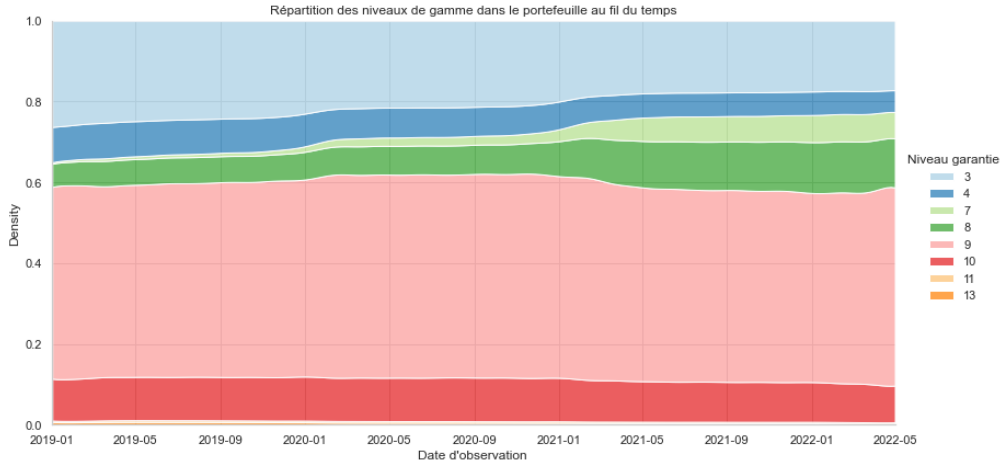


FIGURE 4.7 – Évolution de la répartition des assurés par niveau de gamme

Ce graphe présente une hausse de la part du niveau 7 qui correspond au lancement des nouvelles gammes G16 et G17 en 2020, ainsi qu'une baisse de la proportion des gammes de niveaux 9 qui sont les plus représentées au sein du portefeuille.

Les distributions d'âges observées en 2021 pour chaque niveau de gamme sont représentées comme suit :

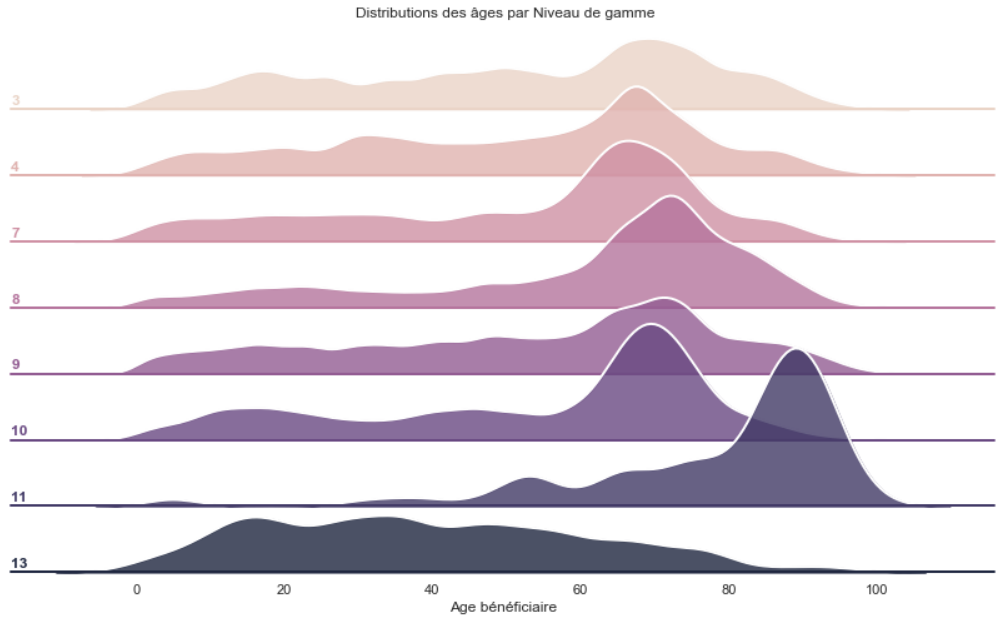


FIGURE 4.8 – Distributions de l'âge des assurés par niveau de gamme

Ci-dessous le nombre de bénéficiaires pour chaque niveau de gamme :

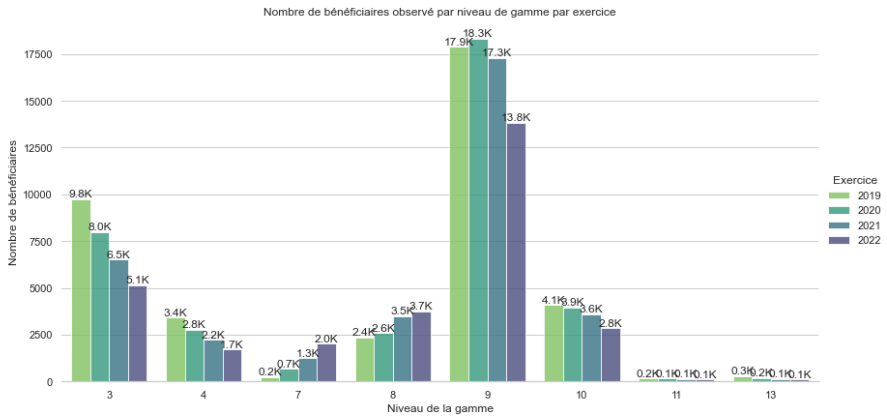


FIGURE 4.9 – Nombre de bénéficiaires par niveau de gamme par exercice

Les distributions d'âges des deux meilleurs niveaux de gammes sont atypiques (niveau 11 et 12). Ces deux garanties sont les moins représentées dans le portefeuille (en 2021 chacune de ces gammes protège une centaine de personnes environ) ce qui peut expliquer le caractère atypique.

Par ailleurs, les bénéficiaires de la gamme 11 sont en majorité des personnes âgées de plus de 80 ans. D'après les cotisations annuelles moyennes par gamme (figure 3.6), les cotisations annuelles moyennes pour ce niveau de gamme pour les âges de plus de 80 ans sont inférieures à ceux de certaines gammes de niveaux inférieurs. Cette différence peut expliquer l'âge moyen des bénéficiaires de ce niveau de gamme.

D'autre part, les cotisations annuelles moyennes des gammes du niveau 11 sont proches de celles des niveaux inférieurs pour les personnes jeunes. L'écart des cotisations annuelles moyennes entre ce niveau et le reste des niveaux de gamme s'élargit en fonction de l'âge et dépasse les 300 €d'écart pour les personnes de plus de 80 ans ce qui explique l'absence de cette tranche d'âge dans ce niveau de gamme.

Nous avons regardé la répartition des bénéficiaires du portefeuille étudié sur les départements français et avons obtenu le graphe suivant :

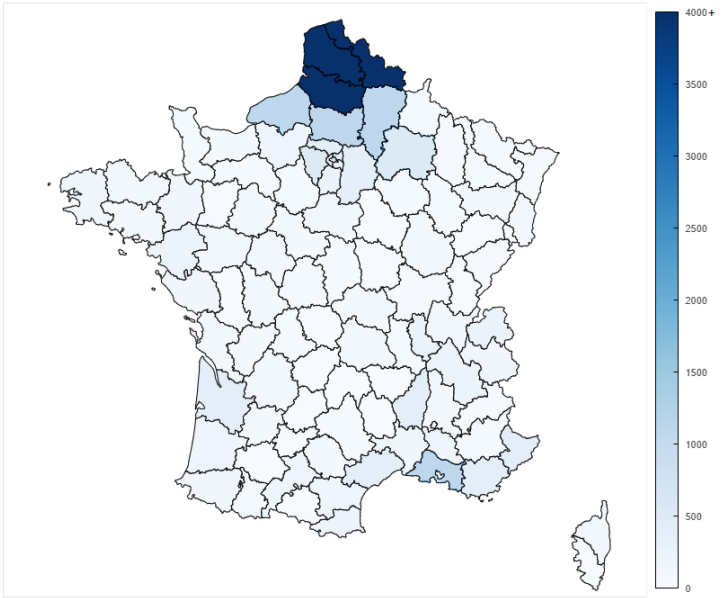


FIGURE 4.10 – Répartition des bénéficiaires par département

Nous constatons que la quasi-totalité des bénéficiaires sont répartis sur 3 départements voisins du nord de la France. Dans ce cadre, l'impact des données départementales disponibles en libre accès telles que le niveau de vie, le taux de chômage, le nombre de praticiens par spécialité... n'auront pas un impact significatif sur les modèles.

4.2 Consommations du portefeuille

Pour la description de la consommation du portefeuille, nous avons commencé par regarder les remboursements mensuels effectués par la mutuelle durant la période de couverture par date de survenance.

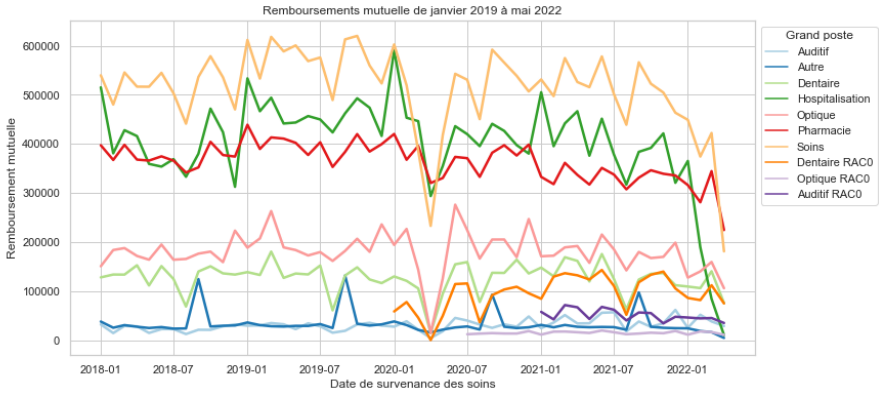


FIGURE 4.11 – Remboursements mutuelle par grand poste

La première zone du graphe attirant notre attention correspond à la chute des remboursements mutuelle en mars 2020 due à la crise covid et le confinement en France. Seul le poste pharmacie n'a pas subi de chute des remboursements sur cette période.

A partir de l'entrée en vigueur du 100% santé au début de l'année 2020, près de la moitié des remboursements en dentaire sont à zéro reste à charge. Une hausse des prestations en auditif est notamment observée avec la prise en compte de certains équipements dans les paniers 100% santé en début 2021.

Pour certains postes d'actes, une périodicité annuelle des remboursements mutuelle peut être repérée. Les remboursements par mois de survenance se présentent comme suit :

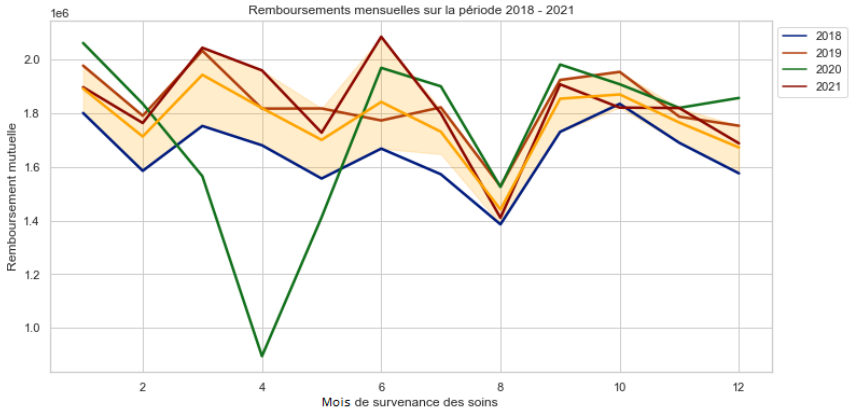


FIGURE 4.12 – Remboursements mutuelle par mois de survenance par exercice

La saisonnalité des remboursements est mieux représentée par ce graphe, la courbe jaune représente la moyenne des remboursements mensuelles pour les exercices 2018, 2019 et 2021.

Avec la résiliation infra-annuelle, la saisonnalité peut avoir un impact direct sur l'estimation de la prime pure des assurés.
Souvent calculée sur toute l'année, l'estimation de la prime pure peut varier pour un calcul sur une période plus courte de l'année.

Afin d'estimer cette variation, nous nous plaçons au début de chaque exercice et calculons les remboursements annuels moyens par bénéficiaire (primes pures) en se basant sur une durée de couverture allant du 1 janvier à une date t qui évolue entre le 1 février et la fin de l'exercice.
Pour un exercice N et t l'écart en nombre de jours entre 1/1/N et le j/m/N, la prime pure annuelle basée sur la période du 1/1/N au j/m/N s'écrit :

$$Prime\,pure(t) = \frac{\sum_{s=0}^{t} Remboursements\,mutuelle(s)}{\sum_{s=0}^{t} Exposition\,totale(s)}$$

Avec,
— Remboursements mutuelle(s) : Somme des remboursements mutuelle référant au jour de survenance s.
— Exposition totale(s) : Somme de l'exposition des bénéficiaires le jour s (en année).
Nous obtenons donc les courbes suivantes :

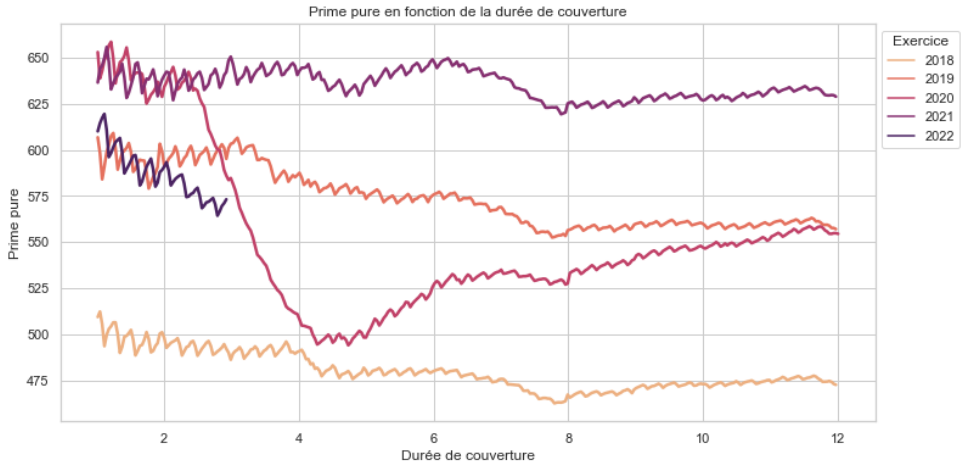


FIGURE 4.13 – Primes pures en fonction de la durée de couverture de l'année

La durée de couverture est ici exprimée en mois pour une meilleure interprétabilité. Les variations liées à la saisonnalité hebdomadaire des courbes sont supprimées à l'aide de moyennes mobiles.

Les courbes obtenues pour les exercices 2018, 2019, et 2021 sont les suivantes :

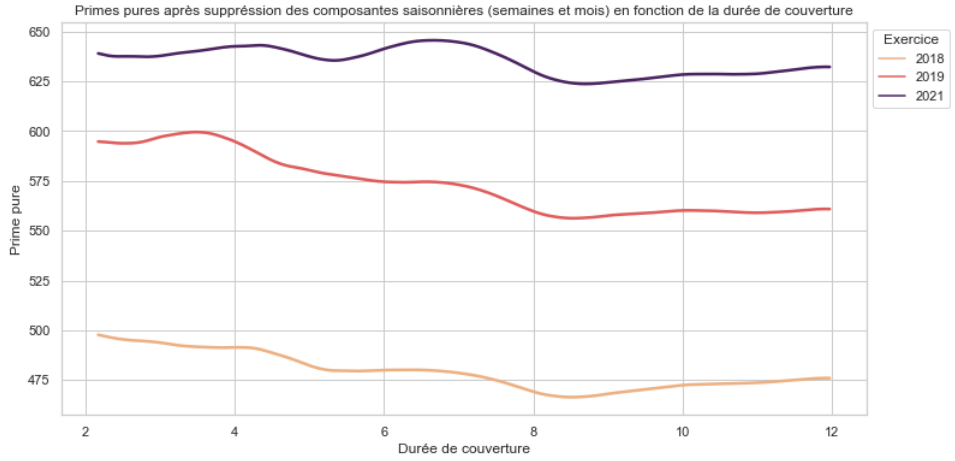


FIGURE 4.14 – Primes pures lissées en fonction de la durée de couverture de l'année

Les primes pures annuelles suivent une saisonnalité similaire pour ces trois exercices. En rapportant la fin des courbes à zéro (par une translation verticale) et en moyennant sur les trois exercices, nous obtenons la variation moyenne de la prime pure annuelle en fonction de la durée de couverture par rapport à une couverture d'une année complète :

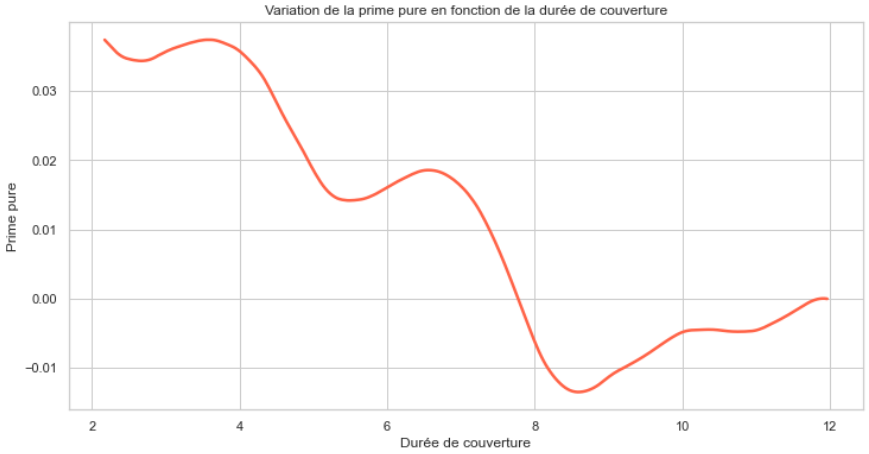


FIGURE 4.15 – Variation de la prime pure annuelle en fonction de la durée de couverture

Supposons qu'on a estimé les primes pures en se basant sur toute l'année de couverture. Pour un assuré qui résilie en avril, sa prime pure serait donc en moyenne sous-estimée de plus de 3%. De même, un assuré qui résilie en septembre, sa prime pure est en moyenne sur-estimée à 1%.

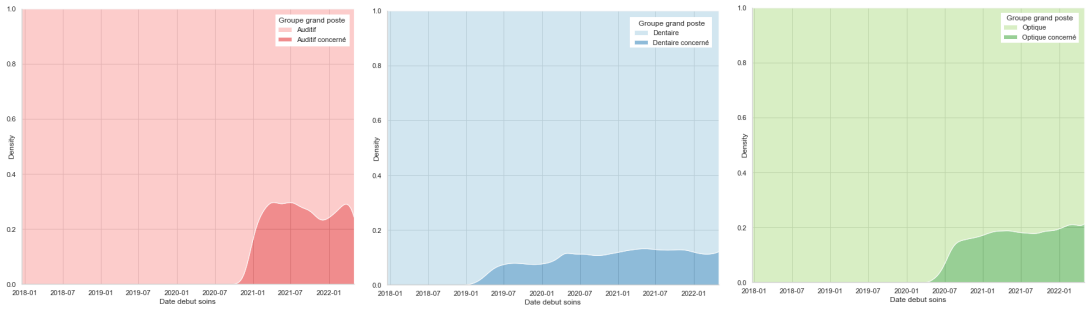
Les cotisations étant perçues proportionnellement aux périodes couvertes par la mutuelle, les résiliations en milieu d'année pourraient détériorer le ratio de sinistralité.

L'effet de la saisonnalité sur les primes pures est pris en compte dans le processus de modélisation mis en place, qui prend en compte la durée de couverture restante des assurés en cas de résiliation probable dans la prédiction des coûts moyens et fréquences.

En plus de la saisonnalité qui impacte les primes pures, la mise en place du 100% santé pourrait impacter la consommation des assurés pour certains actes.

4.2.1 Impact du 100% Santé sur la consommation

Afin de regarder l'impact de l'entrée en vigueur du 100% santé sur la consommation des actes concernés par la réglementation, nous avons commencé par regarder l'évolution de la proportion des codes actes pris en charge par le 100% santé dans le temps pour chaque grand poste concerné.



(a) Proportion des actes concernés par le 100% santé en Audiologie

(b) Proportion des actes concernés par le 100% santé en Dentaire

(c) Proportion des actes concernés par le 100% santé en Optique

La superposition de ces distributions pour les trois postes se présente comme suit :

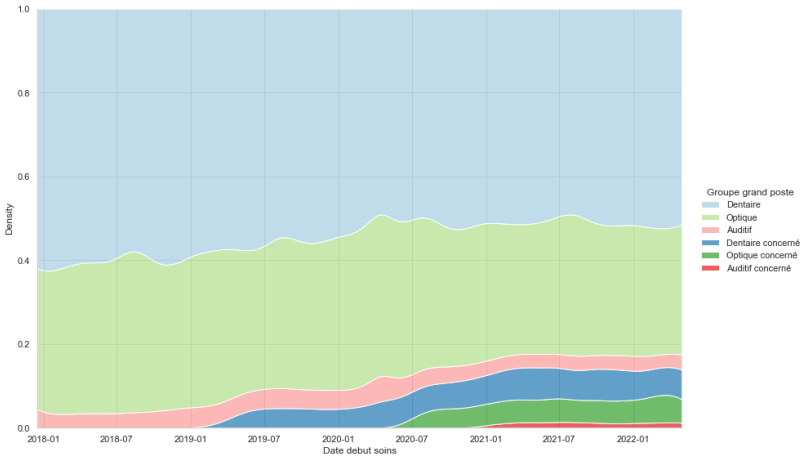


FIGURE 4.17 – Evolution de la consommation des actes pris en charge par le 100% au cours du temps

Les trois distributions dans la partie basse du graphe représentent la proportion (en nombre d'actes) des actes concernés par le reste à charge zéro pour les trois postes concernées par rapport aux restes des actes.

La proportion des actes concernés est restée croissante jusqu'à la mise en place totale des paniers 100% santé. Les assurés ont par la suite plus recours aux types d'actes pris en charges par le 100% santé après l'entrée en vigueur de la réglementation. Ces actes représentent désormais 15% des actes du portefeuille.

L'évolution du coût moyen d'un acte par mois selon s'il est concerné par la réglementation pour les 3 derniers postes se présente comme suit :

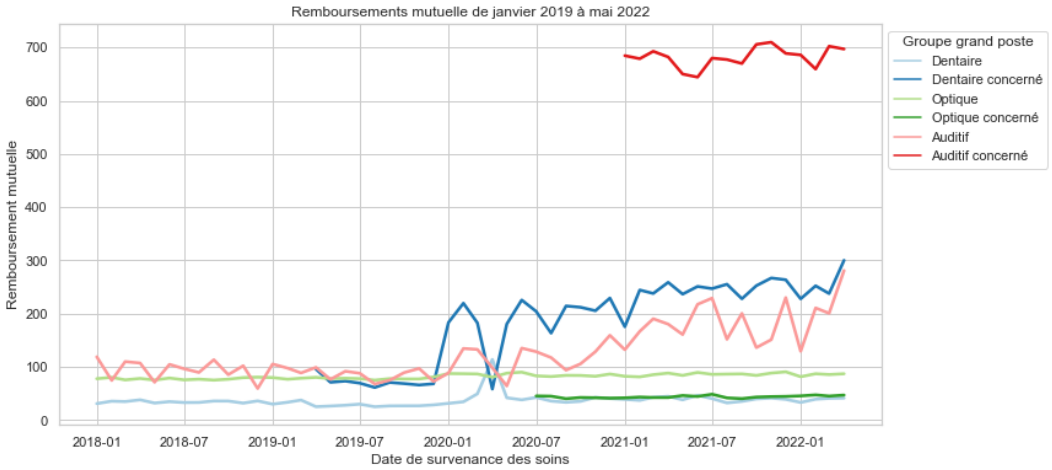


FIGURE 4.18 – Evolution du cout moyen des actes concernés par le 100% santé par mois

Depuis leur prise en charge par le 100% santé en 2021, le coût moyen des actes concernées en audiologie a connu une forte hausse (le coût moyen d'un acte pris en charge par la réglementation d'élève à plus de 600 €). D'autre part, le coût moyen des actes en audiologie non pris en compte a connu à son tour une hausse, cette hausse peut être liée à l'augmentation des garanties.

Par ailleurs, le coût moyen des actes concernés par le 100% santé est toujours supérieur au reste des actes du même groupe homogène.

Enfin l'augmentation de la proportion des actes pris en charge par le 100% santé et leurs coûts moyens plus élevés du reste des actes donneront une hausse de la consommation.

Notre base est composée en partie d'observations avant la mise en place de la réglementation. La prise en compte de cette dernière dans les prédictions se fait par l'ajout de deux variables relatives au nombre de jours, couverts par le 100% santé, antérieurs et postérieurs à la date d'observation.

L'évènement récent qui a le plus impacté le secteur de la santé en France reste la crise covid-19. Cette crise comprend plusieurs vagues épidémiques en France depuis le début de l'année 2020.

4.2.2 Impact de la Covid sur la consommation

Dans le but de découvrir les possibles impacts de cette crise sur la consommation de notre portefeuille, nous traçons le coût mensuel moyen d'un assuré sur un mois glissant de survenance sur la période du 1 janvier 2018 au 5 mai 2022.

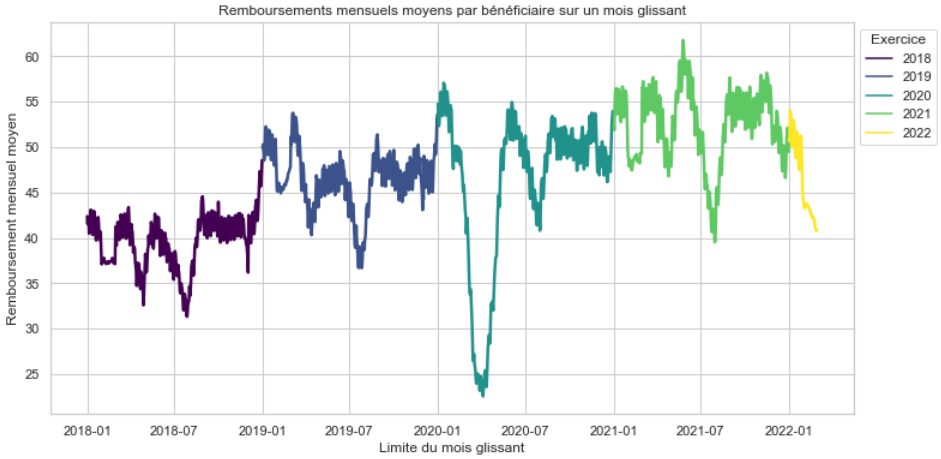


FIGURE 4.19 – Coût mensuel moyen d'un bénéficiaire sur un mois glissant de survenance

Une moyenne mobile permettant la suppression des variations dues à la saisonnalité hebdomadaire, nous donne le résultat suivant :

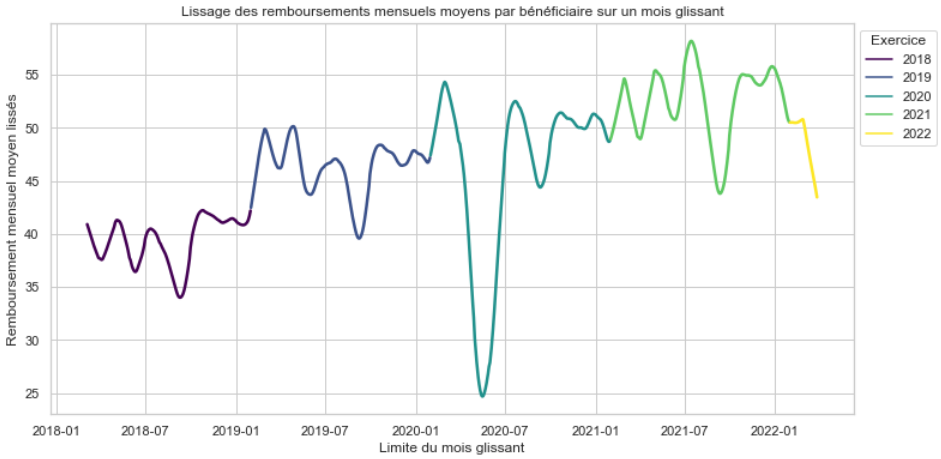


FIGURE 4.20 – Coût mensuel moyen d'un bénéficiaire sur un mois glissant de survenance

La chute des remboursements en 2022 est relative au manque de recul concernant les prestations survenues sur cette période.

La crise COVID a entraîné une chute des prestations sur la période de mars à mai 2020. A partir de ce point, il n'est plus possible de repérer la saisonnalité observée en 2018 et 2019. L'impact de la covid a donc continué sur la période suivant son apparition avec les différentes vagues vécues.

Afin de repérer les différentes évolutions connues durant la crise, nous estimons la variation des prestations dues à la saisonnalité sur les deux premières années. Nous obtenons une variation de prestations moyenne sur un exercice représentée comme suit :

FIGURE 4.21 – Variation moyenne des prestations sur un exercice basée sur les années 2018 et 2019

La soustraction de cette variation aux remboursements mensuels moyens par bénéficiaire de 2020 et 2021 retourne l'évolution des remboursements moyens par bénéficiaire sur la période de crise par rapport aux années 2018 et 2019 (l'impact du 100% santé étant supposé négligeable par rapport à l'effet de la crise sanitaire).

Afin de comparer l'évolution obtenue avec la taille des différentes vagues épidémiques rencontrée, nous sommes allés chercher des indicateurs sur ces dernières depuis la base de données "Synthèse des indicateurs de suivi de l'épidémie COVID-19" en libre accès, publiée quotidiennement par le gouvernement. Parmi ses variables se trouve le nombre de nouveaux cas, le nombre d'hospitalisations, le taux de saturation hospitalière, le nombre de décès liées au covid, ...
Dans notre cas, la variable nombre d'hospitalisations liées au covid est retenue. Cette variable est en lien direct avec la consommation des bénéficiaires de complémentaires santé et représente les risques graves.

La superposition de l'évolution des remboursements mensuels moyens par bénéficiaire après la suppression de la variation moyenne des remboursements moyens par bénéficiaire avant le covid et le nombre d'hospitalisations liées au covid durant la crise se présente comme suit :

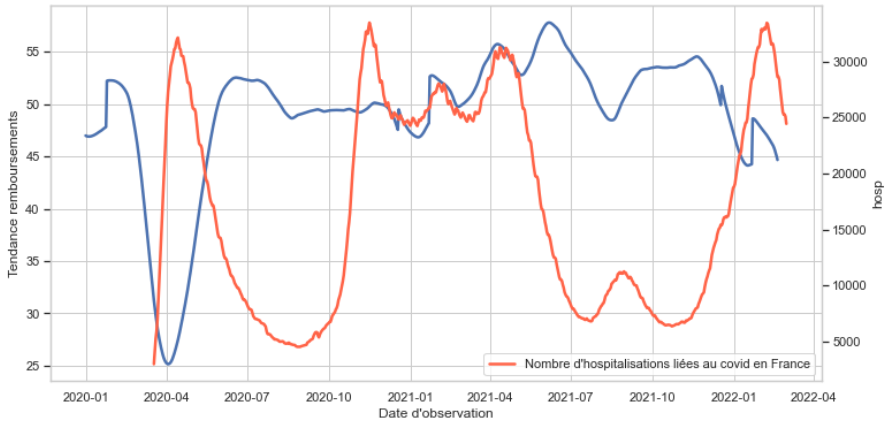


FIGURE 4.22 – Evolution des remboursements mensuels moyen par bénéficiaire et le nombre d'hospitalisations covid en France

Les deux courbes présentent désormais des zones en corrélation. Pour une meilleure interprétation, il était nécessaire de séparer les coûts liés aux hospitalisations des autres prestations. Pour ce faire, la même démarche a été suivie pour les actes hors hospitalisations, puis de déterminer la part liée à l'hospitalisation par la différence des deux courbes. Le résultat obtenu pour les remboursements mensuels moyens en hospitalisation se présente comme suit :

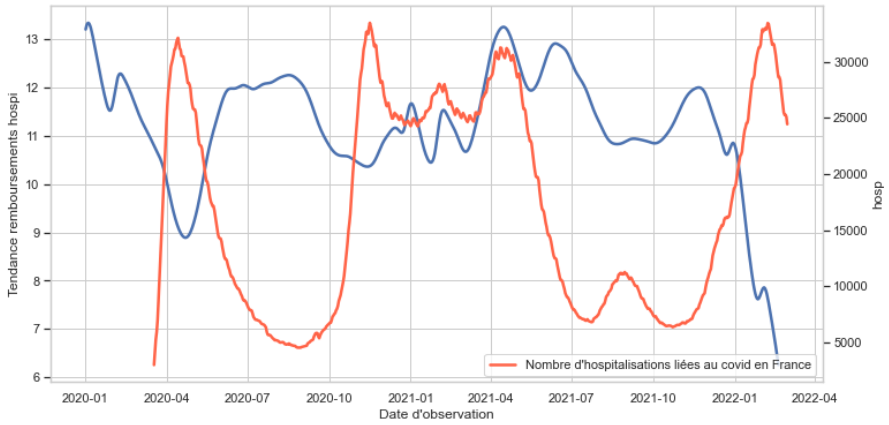


FIGURE 4.23 – Évolution des remboursements mensuels moyen en hospitalisations par bénéficiaire et le nombre d'hospitalisations covid en France

Puis l'évolution du remboursement mensuel moyen hors hospitalisations par bénéficiaire se présente comme suit :

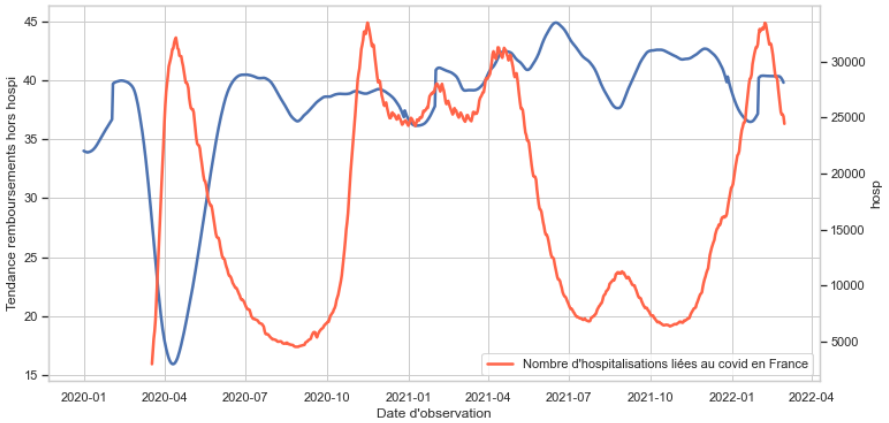


FIGURE 4.24 – Évolution des remboursements mensuels moyen hors hospitalisations par bénéficiaire et le nombre d'hospitalisations covid en France

En septembre 2021, l'apparition d'un pique au niveau des remboursement moyens en hospitalisation vient accompagner le pique de la vague épidémique, alors qu'un creux est observé sur la même période pour le reste des actes.

La prise en compte des vagues épidémiques dans nos prédictions est réalisée par la création des variables suivantes :

— **Hospitalisations covid antérieures** : Nombre d'hospitalisations liées à la covid sur la période d'observation antérieure.
— **Hospitalisations covid postérieures** : Nombre d'hospitalisations liées à la covid sur la période d'observation postérieure.
— **Durée confinement antérieure** : Nombre de jours de confinement sur la période d'observation antérieure.
— **Durée confinement postérieure** : Nombre de jours de confinement sur la période d'observation postérieure.

Les variables postérieures nous permettent d'orienter les prédictions en cas de nouvelle vague épidémique probable.

4.3 Analyse axée sur la résiliation des contrats (Modèle 1)

La première étape de la modélisation des flux sortant du portefeuille est la prédiction de la résiliation des contrats à l'horizon d'un an. Dans cette partie, une analyse de la distribution des différentes variables explicatives en fonction de la résiliation permettra l'identification du profil des contrats les plus soumis à la résiliation.

Nous commençons par l'étude de l'impact sur la résiliation des variables définissant la composition des bénéficiaires des contrats, les caractéristiques des contrats puis les consommations antérieures à la date d'observation par contrat.

Pour une meilleure interprétabilité des graphes, les distributions des contrats résiliés et non-résiliés ne seront pas représentées sous une norme commune. Cela permettra de comparer les différentes distributions indépendamment de la taille des deux populations de contrats (résiliés et non-résiliés).

Âges des adhérents

La distribution de l'âge des adhérents en fonction de la résiliation du contrat à l'horizon d'un an est représentée de la manière suivante :

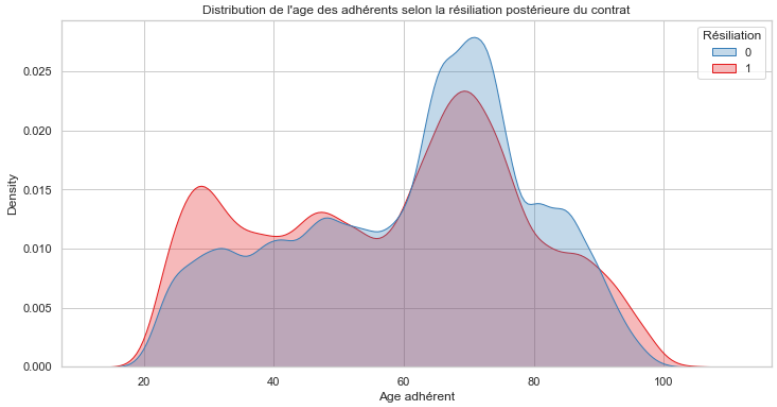
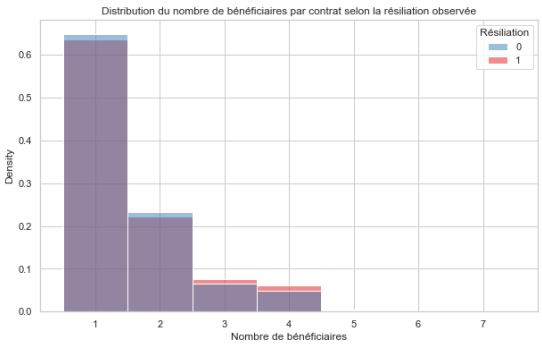


FIGURE 4.25 – Distribution de l'âge des adhérents selon la résiliation du contrat

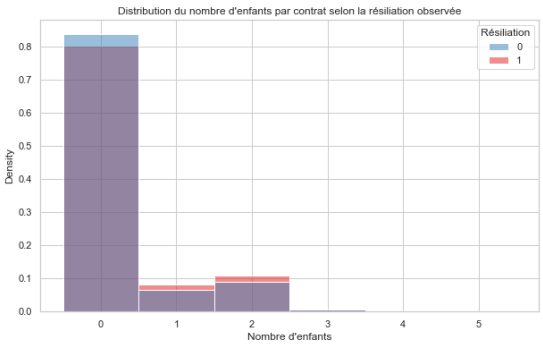
Nous constatons que les résiliations sont plus représentées pour les contrats dont les adhérents sont âgées de moins de 50 ans. Cet écart est surtout présent pour les adhérents de moins de 35 ans. D'autre part, la résiliation est aussi mieux représentée pour les contrats des adhérents les plus âgés, ces résiliations sont probablement dues aux décès plus fréquents sur cette tranche d'âge.

Nombre de bénéficiaires et d'enfants souscrits

Les distributions du nombre de bénéficiaires et enfants par contrat en fonction de la résiliation se présentent comme suit :



(a) Distribution du nombre de bénéficiaires par contrat selon la résiliation

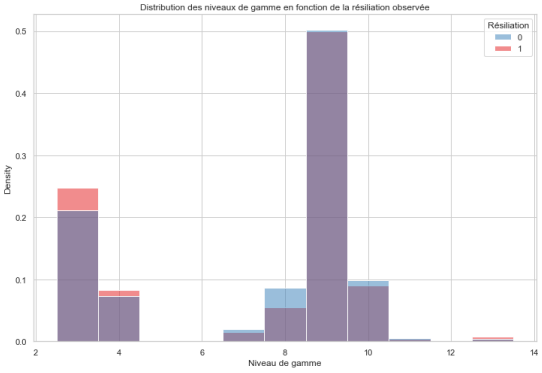


(b) Distribution du nombre d'enfants par contrat selon la résiliation

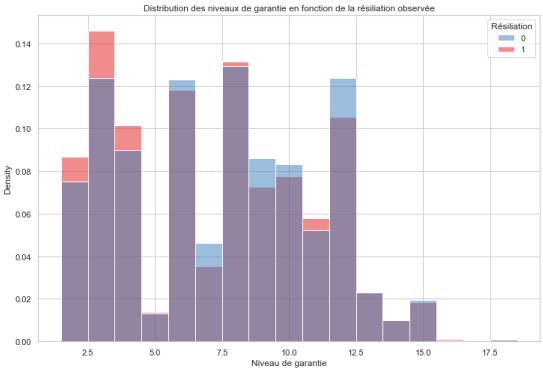
Les contrats couvrants au moins 3 bénéficiaires et ceux comprenant des enfants sont, en proportion, les plus concernés par la résiliation. Le risque de résiliation serait donc croissant en fonction de ces deux variables.

Niveaux de gammes et garanties

Du coté des niveaux de gammes et garanties, les distributions de ces derniers en fonction de la résiliation se présente comme suit :



(a) Distribution des niveaux de gamme en fonction de la résiliation



(b) Distribution des niveaux de garantie en fonction de la résiliation

Les contrats les moins concernés par la résiliation appartiennent généralement aux niveaux intermédiaires de gammes et garanties. Les contrats appartenant aux niveaux basiques et de luxe connaissent plus de résiliations pour notre portefeuille.

Durée avant éligibilité

La durée avant éligibilité est, à la date d'observation, la durée après laquelle un contrat peut résilier. Elle est calculée en fonction de l'ancienneté après l'entrée en vigueur de la résiliation infra-annuelle et en fonction des dates d'anniversaire des contrats avant l'entrée en vigueur de cette réglementation.
La distribution de cette variable, sur toute la durée de couverture du portefeuille, en fonction de la résiliation des contrats est représentée comme suit :

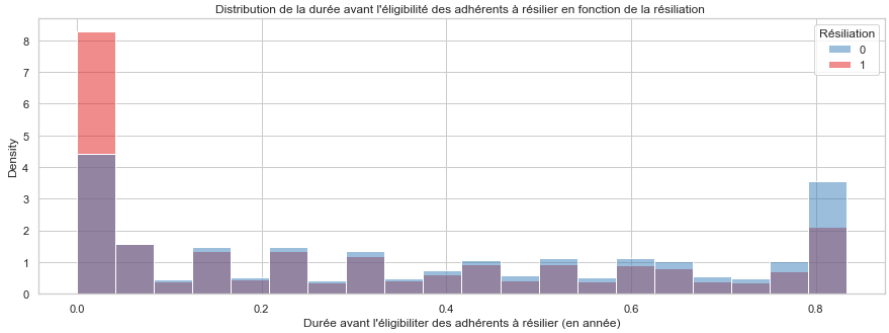
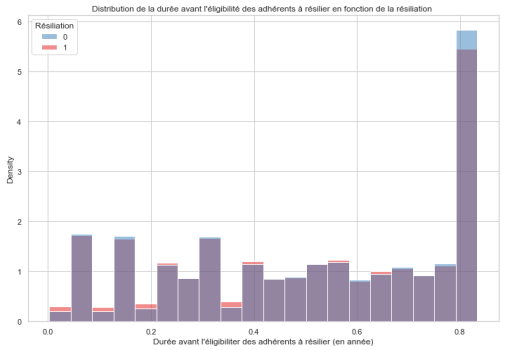


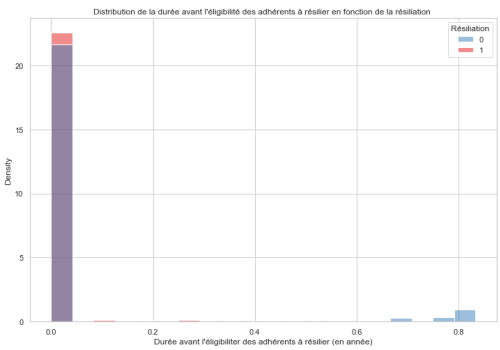
FIGURE 4.28 – Distribution de la durée avant éligibilité en fonction de la résiliation du contrat

Une durée d'éligibilité de zéro correspond à un contrat qui est libre de résilier (ancienneté supérieur à un an après l'entrée en vigueur de la résiliation infra-annuelle), il est donc naturel que la majorité des contrats ayant résilié aient la liberté de résilier à tout moment. De plus, la proportion des contrats résiliés diminue avec l'augmentation de la durée d'éligibilité car sur ces contrats ayant une longue durée avant éligibilité, la période de résiliation est très courte sur notre horizon d'un an.

Cette durée a connu un avant et un après entrée en vigueur de la résiliation infra-annuelle en fin 2020. Les distributions de cette variable relatives aux observations de 2019 et 2021 se présentent comme suit :



(a) Distribution de la durée avant éligibilité en 2019



(b) Distribution de la durée avant éligibilité en 2021

Contrairement à l'année 2019, la majorité des contrats sont éligibles à résilier en 2021 grâce à la nouvelle réglementation.

Ancienneté des contrats

La distribution de l'ancienneté des contrats est représentée comme suit :

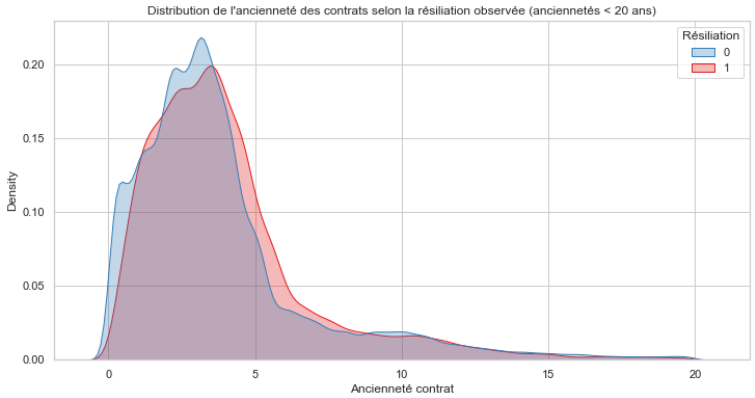


FIGURE 4.30 – Distribution de l'ancienneté des contrats en fonction de la résiliation

D'après ce graphe, les contrats auront plus tendance à résilier entre 4 et 8 ans d'ancienneté. Sur cette distribution nous pouvons voir que l'ancienneté minimale avant la résiliation est respectée. Cette ancienneté n'est pas affectée par un changement de produit au sein de la mutuelle, une variable retournant la durée depuis la dernière souscription prend en considération cet effet.

Durée depuis souscription

Cette variable retourne la durée depuis la souscription à la gamme active à la date d'observation. Sa distribution en fonction de la résiliation est la suivante :

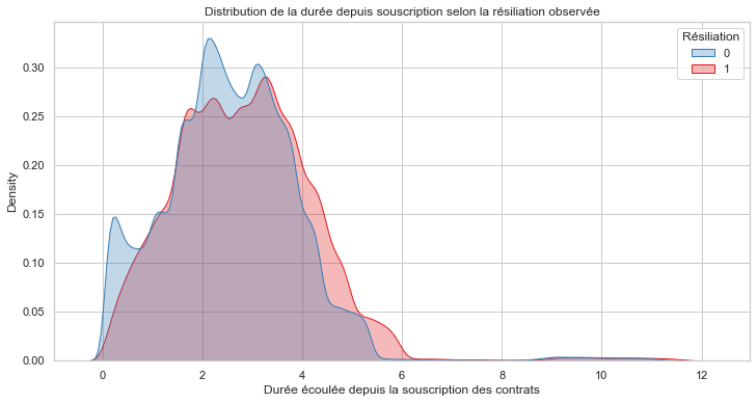


FIGURE 4.31 – Distribution de la durée depuis souscription en fonction de la résiliation

La quasi-totalité des assurés changent de gamme au bout de 4 à 6 ans après la souscription. Par suite, le risque de résiliation évoluerait dans le sens positif de cette variable.

Cotisations annuelles

Les cotisations annuelles représentent ici les cotisations payées pour l'ensemble des bénéficiaires du contrat sur l'année d'observation. La distribution de ces dernières en fonction de la résiliation est la suivante :

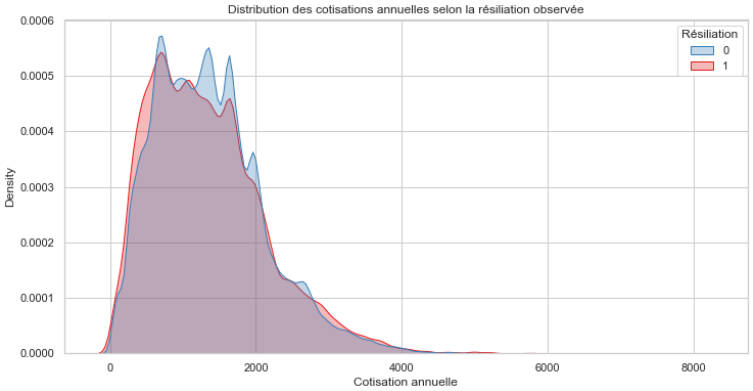
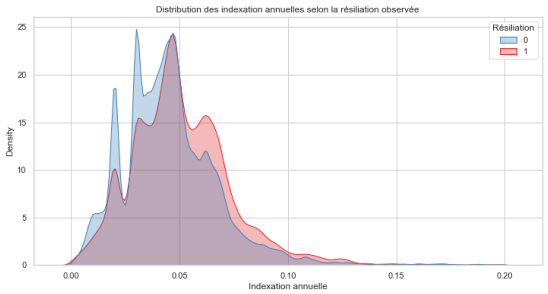


FIGURE 4.32 – Distribution des cotisations annuelles en fonction de la résiliation

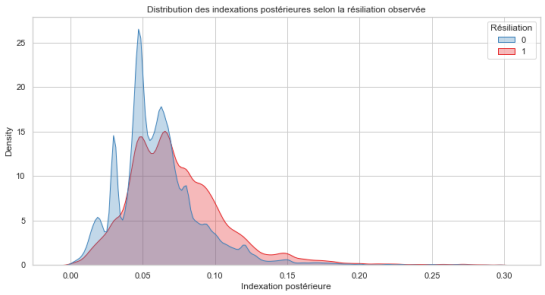
Les contrats les plus concernés par la résiliation sont ceux qui payent les cotisations les moins chères. Cependant, ces distributions restent très proches.

Indexation antérieure et postérieure

Pour une observation l'année N, l'indexation antérieure représente l'indexation appliquée au début de l'année N. Cependant, l'indexation postérieure est celle qui s'appliquerait au début de l'année N+1. Leurs distributions en fonction de la résiliation se présentent comme suit :



(a) Distribution de l'indexation antérieure



(b) Distribution de l'indexation postérieure

Les indexations de niveau élevé, qu'elles soient antérieures ou postérieures, favorisent le risque de résiliation des contrats d'après ces deux graphes.

Consommation antérieure

Les fréquences et coûts moyens des actes antérieurs aux observations en fonction de la résiliation se présentent comme suit :

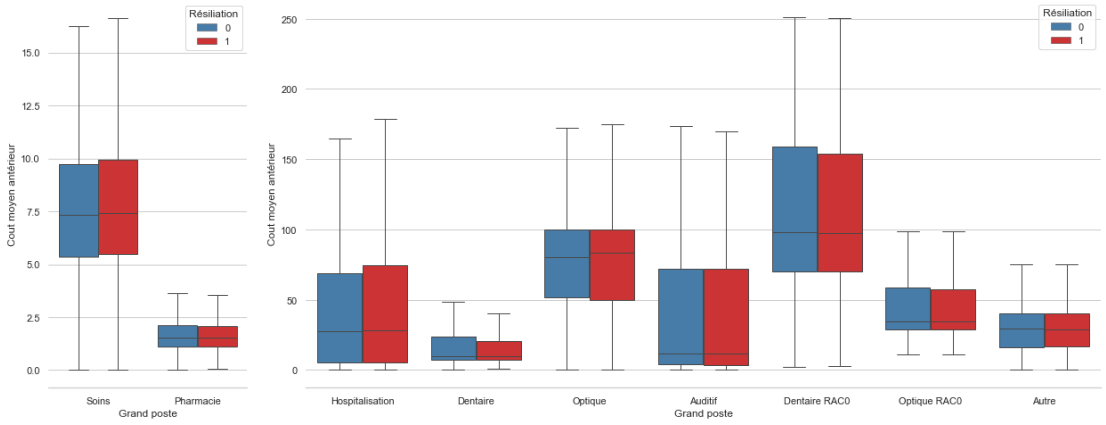


FIGURE 4.34 – Coûts moyens antérieurs par poste

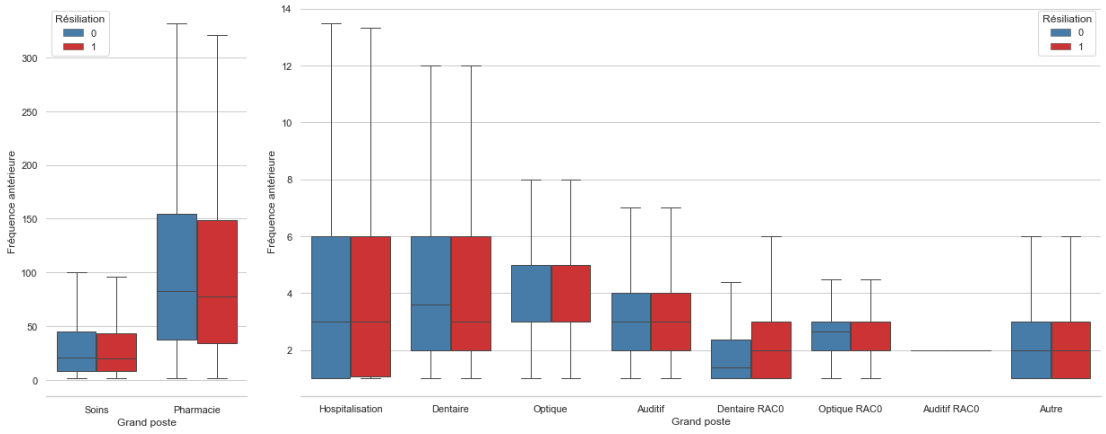


FIGURE 4.35 – Fréquences antérieures par poste

Ces graphiques prennent en compte uniquement les fréquences et coûts moyens antérieurs strictement supérieurs à zéro. Les soins de villes et la pharmacie ont été séparés des autres postes vu leurs fortes fréquences et faibles coûts moyens.

Le coût moyen des actes en auditif étant très élevé n'est pas représenté sur le graphe. Il est aux alentours de 710 € pour 84% des actes et de 550 € pour 13% des actes et aux alentours de 380 € pour les 3% restant.

La fréquence des actes en soins de ville, pharmacie et dentaire est en moyenne inférieure chez les contrats résiliant postérieurement.pour les contrats résiliés, la fréquence antérieure est supérieure pour le poste dentaire RAC0.

De l'autre coté, le coût moyen antérieur des actes en hospitalisation, soins de ville et optique est en moyenne supérieur pour les cas de résiliation. Des coûts moyens antérieurs inférieur pour les résiliations sont observés pour les postes pharmacie, dentaire et dentaire RAC0.

Reste à charge antérieur

Le reste à charge antérieur d'un adhérent représente un élément important dans la résiliation des contrats.

Nous avons regardé le cas des restes à charges, sur une année antérieure aux dates d'observation, élevés pour les postes hospitalisation, dentaire et optique.

La distribution des restes à charges antérieurs, supérieurs à 1000 €, pour les postes hospitalisation, dentaire et optique en fonction de la résiliation postérieure des contrats se présente comme suit :

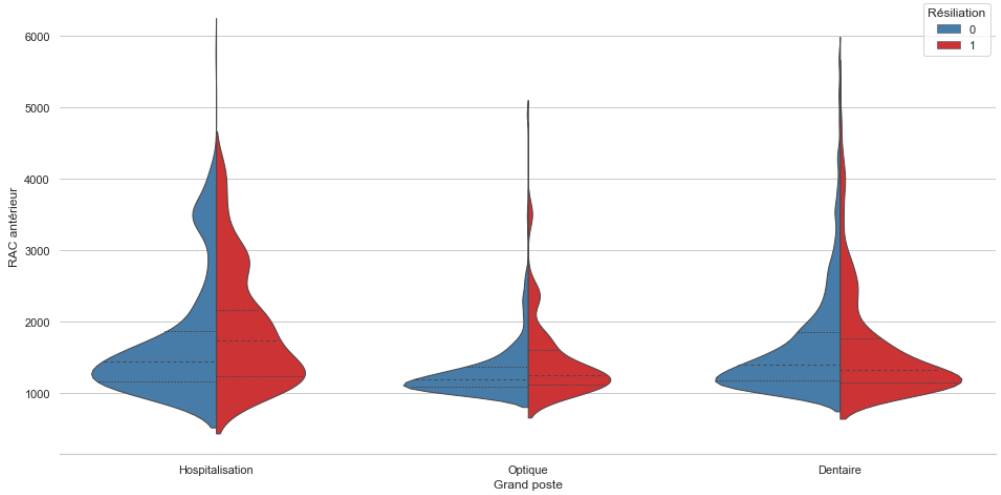


FIGURE 4.36 – Distribution des restes à charge antérieurs (>1000€) en fonction de la résiliation

Les cas de résiliation sont plus représentés pour les restes à charge les plus élevés pour les trois postes. Un reste à charge antérieur élevé augmenterait donc le risque de résiliation.

4.4 Analyse axée sur la durée de couverture restante pour les contrats à risque de résiliation

La durée de couverture restante concerne uniquement les contrats à risque de résilier (cette variable est égale à 1 dans le cas de non résiliation), cette étude est donc basée sur les cas de résiliation.

La durée de couverture restante, pour les cas de résiliation dans un horizon d'un an postérieur à la date d'observation, dépendra de la date d'entrée en vigueur de la résiliation infra-annuelle. Ci-dessous, la distribution de cette variable pour chacune des années d'observation 2019, 2020 et 2021.

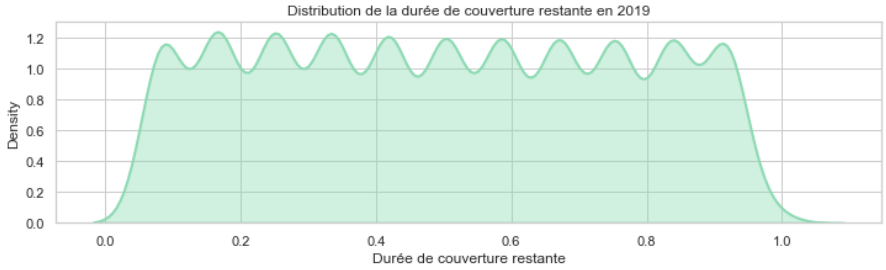


FIGURE 4.37 – Distribution de la durée de couverture restante pour les observations 2019

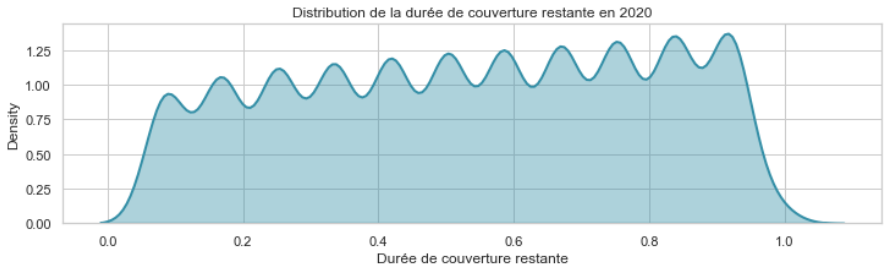


FIGURE 4.38 – Distribution de la durée de couverture restante pour les observations 2020

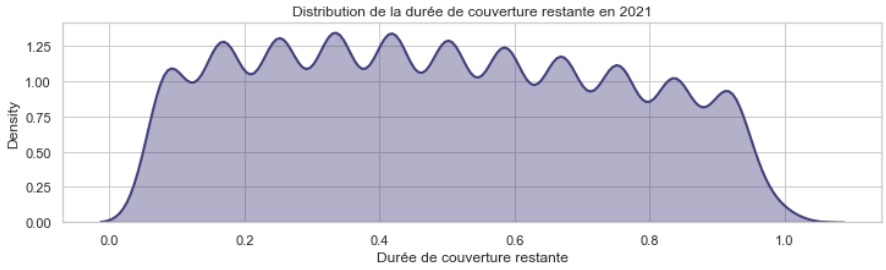


FIGURE 4.39 – Distribution de la durée de couverture restante pour les observations 2021

Les pics présents dans les distributions proviennent de la rencontre de deux faits. Le premier est que la majorité des résiliations s'effectuent en janvier de chaque année. La deuxième est que les observations sont mensuelles (le premier jour de chaque mois).

Par exemple, pour une observation en mars d'une année N, la majorité des résiliations auront une durée de couverture restante égale au nombre de jours restant dans l'année divisé par le nombre de jours dans l'année $\approx 10/12 \approx 0,83$ d'où l'avant dernier pic de chaque distribution.

Mois d'observation

Pour une meilleure interprétation, la décomposition de chacune de ces distributions par mois d'observation se présente comme suit :

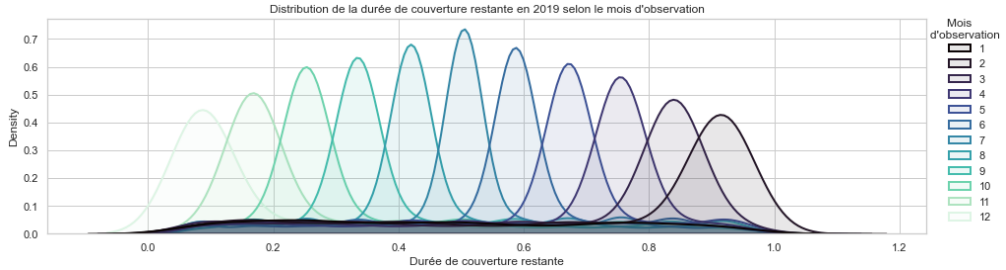


FIGURE 4.40 – Distribution de la durée de couverture restante pour les observations 2019

En 2019 la distribution est équirépartie dû au fait que la résiliation infra-annuelle n'est pas encore entrée en vigueur dans l'horizon de nos observations.

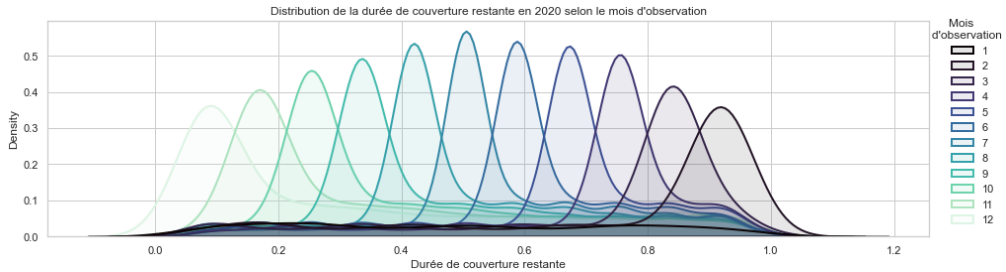


FIGURE 4.41 – Distribution de la durée de couverture restante pour les observations 2020

En 2020, la résiliation entre en vigueur au milieu de chaque horizon d'observation, la proportion des durées de couverture restantes augmente avec l'augmentation de cette durée. En effet, pour les derniers mois d'observation (relatifs aux premiers pics) la résiliation infra-annuelle est entrée tôt dans l'horizon postérieur d'observation et les résiliations pouvaient se faire sur une plus large période ce qui atténue les pics correspondant à la date d'anniversaire des contrats.

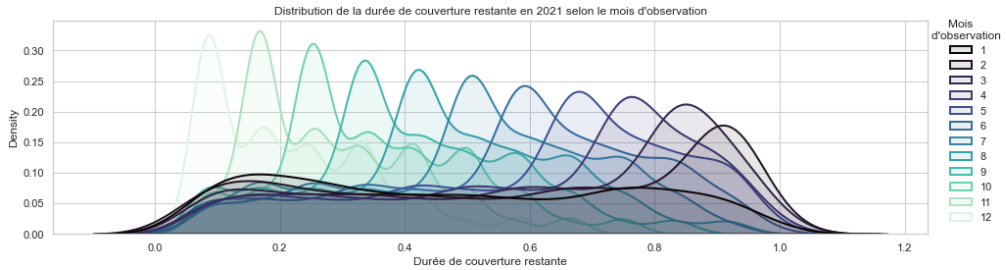


FIGURE 4.42 – Distribution de la durée de couverture restante pour les observations 2021

Une fois la résiliation infra-annuelle totalement installée en 2021, la distribution prend une forme un peu moins régulière. Cependant, il y a toujours les pics relatifs aux résiliations effectués majoritairement au mois de janvier, la résiliation infra-annuelle pourrait être ignorée par une grande partie des adhérents qui n'ont pensé à résilier qu'à la date d'anniversaire de leurs contrats.

Ainsi, la variable mois d'observation jouera un rôle très important pour la prédiction de la durée de couverture restante pour les contrats à risque de résilier.

Année d'observation

Afin de mieux appréhender la distribution de la durée de couverture restante, nous traçons les distributions par an pour des observations effectués le mois de mars puis le mois d'octobre. La distribution de la durée de couverture restante pour une observation du mois de mars se présente comme suit :

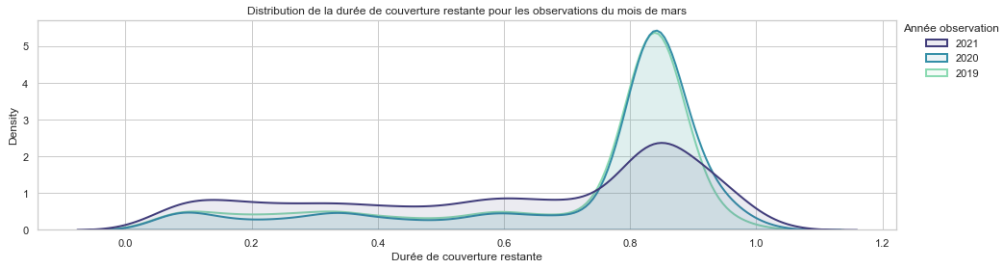


FIGURE 4.43 – Distribution de la durée de couverture restante pour les observations du mois de **mars**

Les pics observés sur la courbe sont relatifs au mois de janvier. La distribution de la durée de couverture restante est quasi identique pour les observations de 2019 et 2020. En effet, la résiliation infra-annuelle n'entre en vigueur qu'à partir du mois de décembre 2020 donc il n'y a pas une grande différence en terme d'éligibilité à résilier pour une observation en mars.

En revanche, la distribution de 2021 est mieux répartie en 2021 grâce à la résiliation infra-annuelle. Les résiliations observées entre mars et janvier qui suit (durée entre 0 et 8,33) ont presque doublées par rapport aux deux années précédentes.

La distribution de la durée de couverture restante pour une observation du mois d'octobre se présente comme suit :

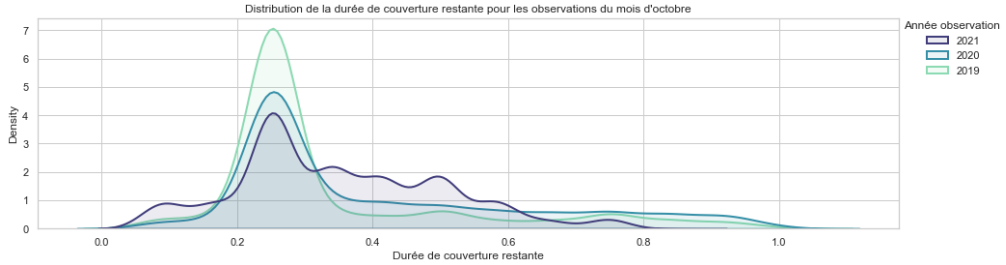


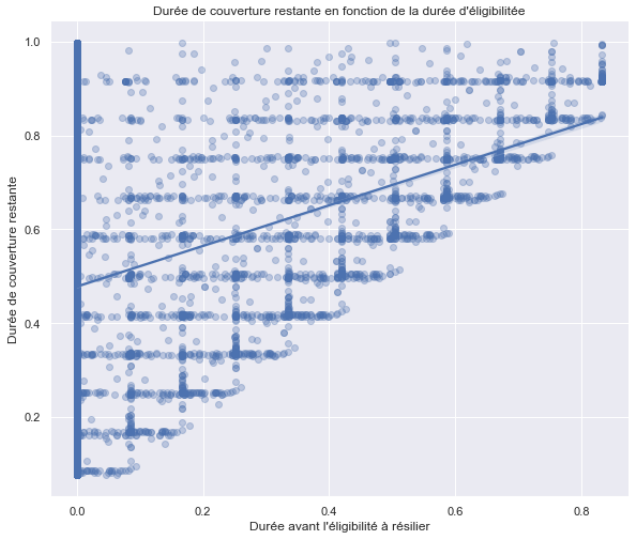
FIGURE 4.44 – Distribution de la durée de couverture restante pour les observations du mois d'**octobre**

Pour les observations de 2020, l'impact de la résiliation infra-annuelle est visible cette fois. En effet, la réglementation est entrée en vigueur en décembre seulement deux mois après. Par la suite, les résiliations sont moins concentrées en janvier et plus réparties sur la période postérieure.

En 2021, la distribution prend en compte seulement les résiliations datées d'avant le 6 mai 2022 (notre horizon de données postérieures). Cette distribution est plus répartie sur l'ensemble de l'année postérieure à l'observation.

Durée avant éligibilité

La durée de couverture restante est encadrée par la durée avant éligibilité. En effet, les contrats ne peuvent résilier qu'après que la durée avant laquelle ils seront éligibles à résilier s'écoule. La durée de couverture restante en fonction de la durée avant éligibilité est représentée ci-dessous :



La durée de couverture restante est bien toujours supérieure à la durée avant éligibilité. Cette variable sera donc d'une grande importance pour la prédiction.

4.5 Analyse de la consommation postérieure des assurés à la date d'observation

La modélisation de l'évolution de la consommation postérieure du portefeuille se fait par la prédiction du coût moyen et fréquence pour chaque bénéficiaire durant sa période de couverture restante et pour chaque grand poste d'actes.

La distribution des coûts moyens et fréquences postérieurs et antérieurs aux dates d'observation par poste d'actes se présente comme suit :

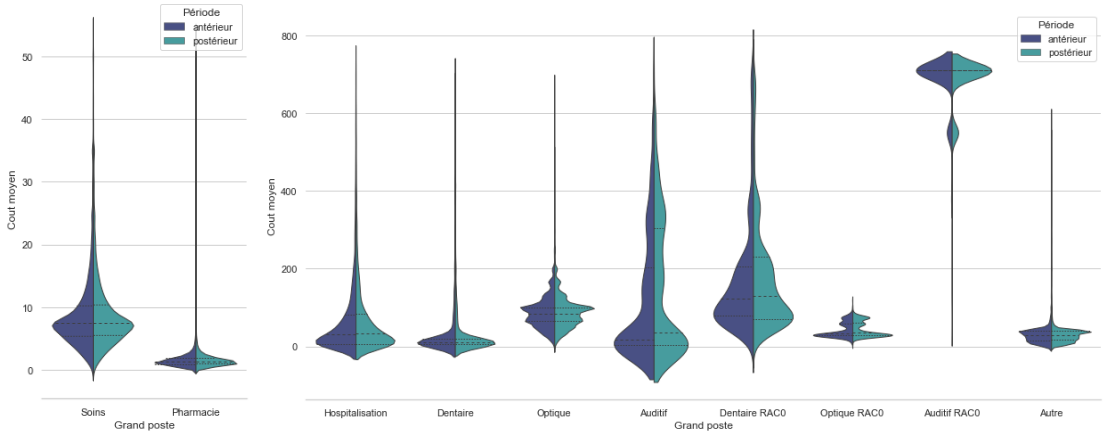


FIGURE 4.45 – Coûts moyens postérieurs et antérieurs par poste

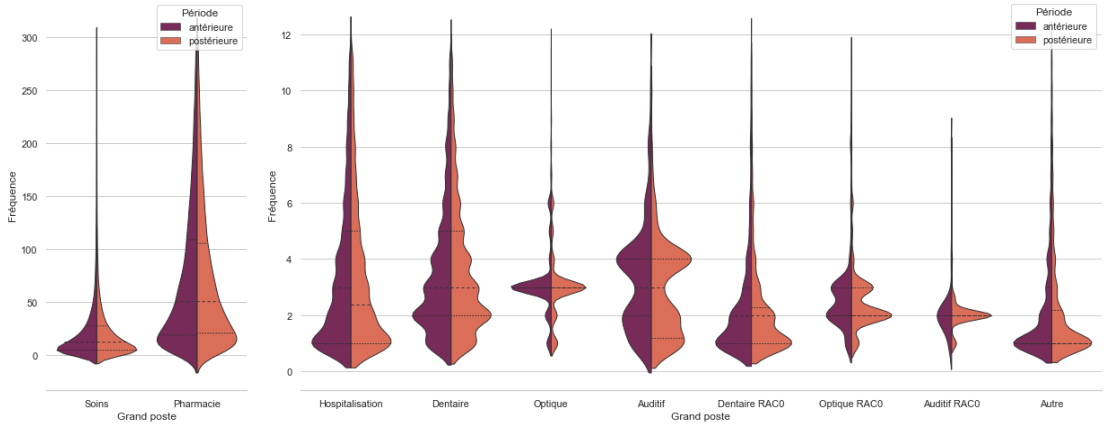


FIGURE 4.46 – Fréquences postérieures et antérieures par poste

Pour ne pas double compter les coûts moyens et fréquences antérieurs et postérieurs à cause des observations successives et pour prendre compte du 100% santé sur tous les postes concernés, ce graphe est issu d'une unique observation d'avril 2021.

Les distributions des coûts moyens et fréquences postérieures à cette date d'observation et par poste d'actes restent très proches de celle observés sur l'année antérieure à la date d'observation.

Âge des bénéficiaires

En santé, l'âge du bénéficiaire représente l'indicateur le plus important sur sa consommation. Dans ce qui suit, nous représentons les coûts moyens et fréquences postérieurs et antérieurs pour une sélection de grands postes d'actes par tranche d'âge :

Les coûts moyens et fréquences en **Hospitalisation** sont représentés comme suit :

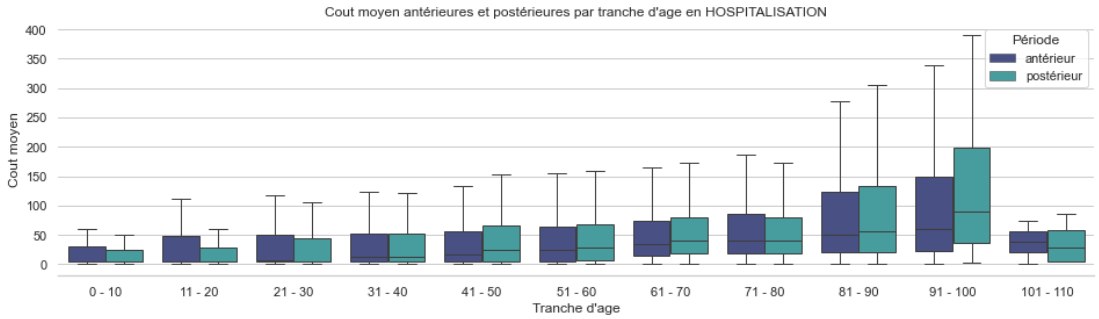


FIGURE 4.47 – **Coûts moyens** antérieurs et postérieurs en **Hospitalisation**

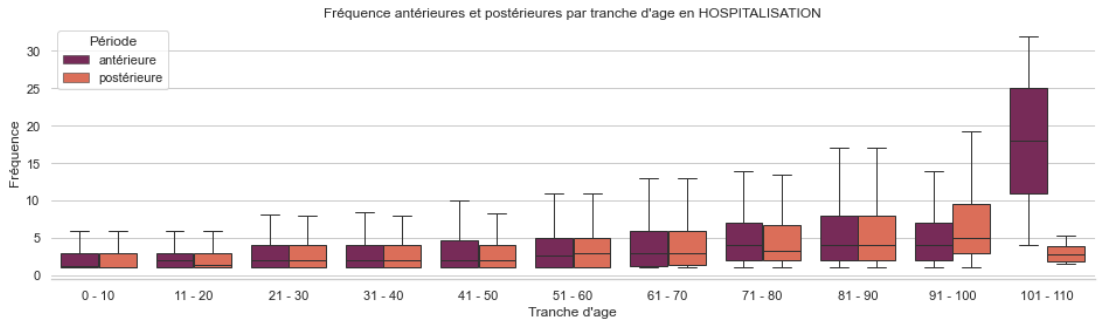


FIGURE 4.48 – **Fréquences** antérieures et postérieures en **Hospitalisation**

En hospitalisation, les coûts moyens et fréquences augmentent généralement avec l'âge des assurés. Une différence au niveau des actes antérieurs et postérieurs à la date d'observation est induite par cette évolution croissante. en effet, pour une date d'observation, l'assuré aura en moyenne un an de plus durant sa consommation postérieure que celle antérieure. Pour les plus jeune, le coût moyen d'un acte décroît entre sa consommation

antérieure et postérieure pour une date d'observation jusqu'à l'âge jusqu'aux alentours de ses 30 ans.

Les coûts moyens et fréquences en **soins de ville** sont représentés comme suit :

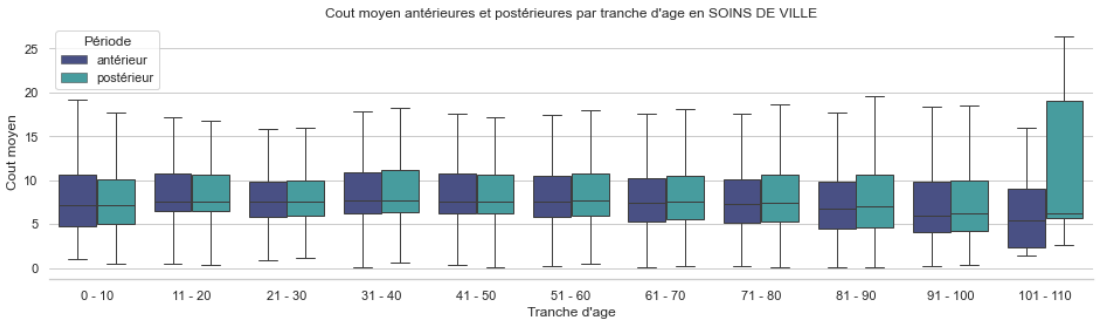


FIGURE 4.49 – **Coûts moyens** antérieurs et postérieurs en **soins de ville**

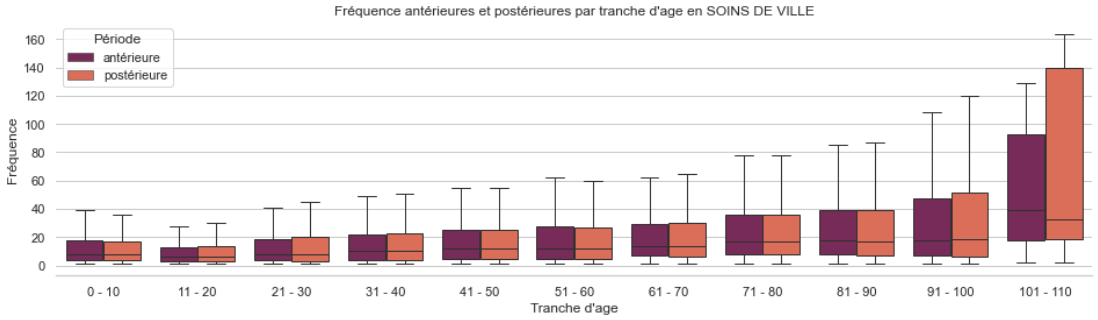


FIGURE 4.50 – **Fréquences** antérieures et postérieures en **soins de ville**

En soins de ville, le coût moyen d'un acte est maintenu sauf pour les âges supérieurs à 100 ans, cette différence est issue du nombre très faible d'assurés pour cette tranche d'âges. Une légère hausse entre les antérieurs et postérieurs à l'observation est repérée pour ce poste.

Concernant les fréquences, celles-ci augmentent de manière exponentielle en fonction des tranches d'âges. Pour les plus âgés, un écart entre les fréquences antérieures et postérieures est induit par cette hausse.

Les actes en soins de ville sont souvent accompagnés par de la pharmacie. Les coûts moyens et fréquences en **pharmacie** sont représentés comme suit :

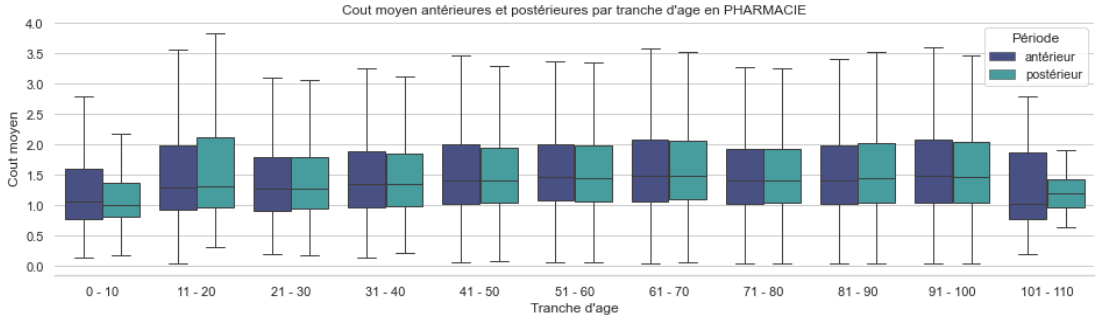


FIGURE 4.51 – **Coûts moyens** antérieurs et postérieurs en **pharmacie**

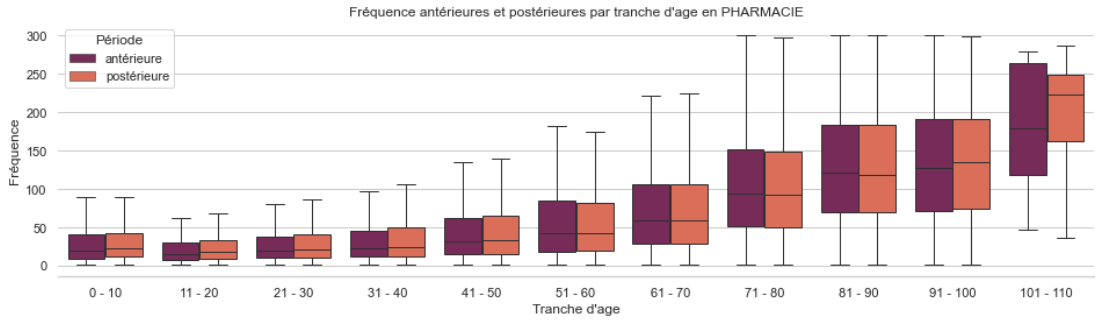


FIGURE 4.52 – **Fréquences** antérieures et postérieures en **pharmacie**

De la même façon que les soins de ville, la fréquence des actes en pharmacie connaît une hausse en fonction de l'âge accompagnée par un élargissement du spectre des fréquences.

En **dentaire hors 100% santé**, les coûts moyens et fréquences se présentent comme suit :

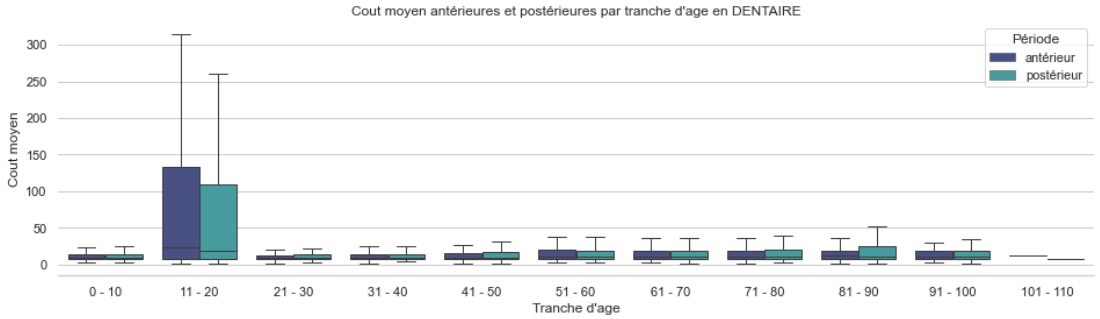


FIGURE 4.53 – **Coûts moyens** antérieurs et postérieurs en **dentaire**

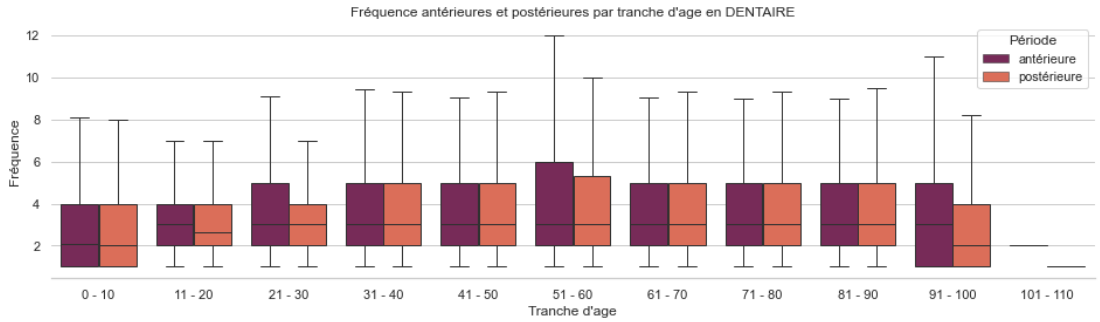


FIGURE 4.54 – **Fréquences** antérieures et postérieures en **dentaire**

Au niveau des coûts moyens, les âges entre 11 et 20 peuvent présenter des coûts moyens très hauts. En effet, cette tranche d'âge concerne l'évolution dentaire de l'enfant à l'adulte. Une légère hausse continue des coûts moyens est observée pour le reste des âges.

Au niveau des fréquences, celles-ci restent en moyenne entre 1 et 6 actes en cas de consommation.

En **dentaire 100% santé**, les coûts moyens et fréquences se présentent comme suit :

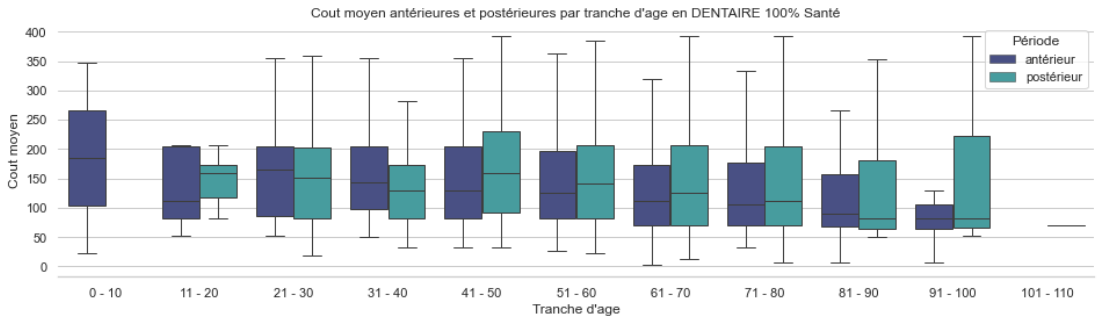


FIGURE 4.55 – **Coûts moyens** antérieurs et postérieurs en **dentaire 100% santé**

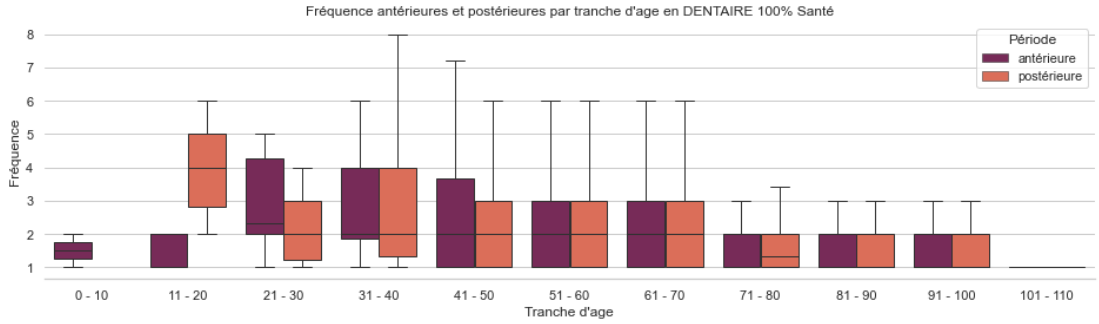


FIGURE 4.56 – **Fréquences** antérieures et postérieures en **dentaire 100% santé**

Les actes en 100% santé, représentant des prothèses dentaires, coûtent beaucoup plus qu'en dentaire hors 100% santé.

La fréquence de ces actes double entre la période antérieure et postérieure à l'observation pour les moins de 20 ans. Cette fréquence est en moyenne inférieure pour les assurés agés de plus de 70 ans.

En **optique**, les coûts moyens et fréquences se présentent comme suit :

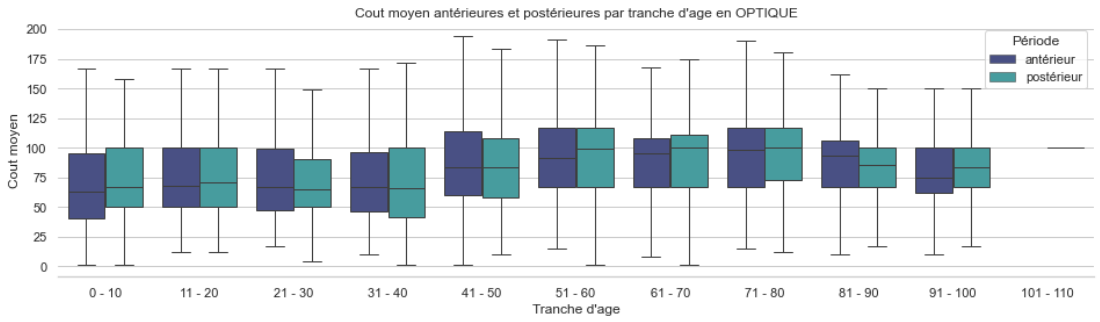


FIGURE 4.57 – **Coûts moyens** antérieurs et postérieurs en **optique**

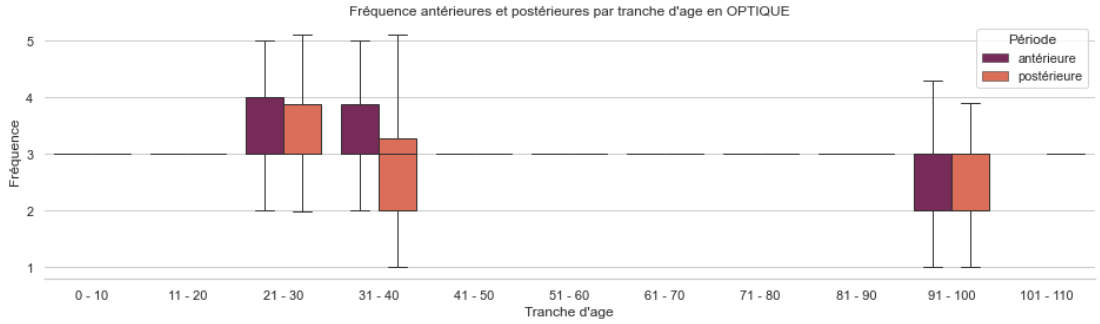


FIGURE 4.58 – **Fréquences** antérieures et postérieures en **optique**

Le coût moyen des actes en optique est supérieur chez les assurés de plus de 40 ans. Quant à la fréquence des actes, elle est principalement égale à 3 actes relatives aux deux verres et monture d'une paire de lunettes.

Enfin, pour les actes en **optique 100% santé**, les coûts moyens et fréquences se présentent comme suit :

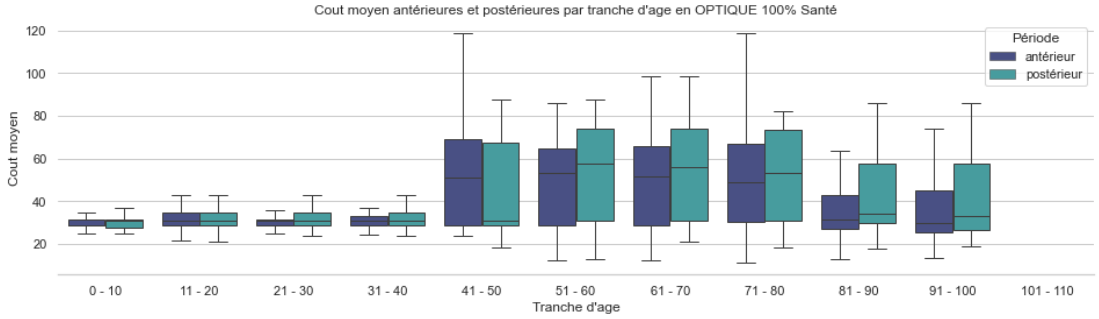


FIGURE 4.59 – **Coûts moyens** antérieurs et postérieurs en **optique 100% santé**

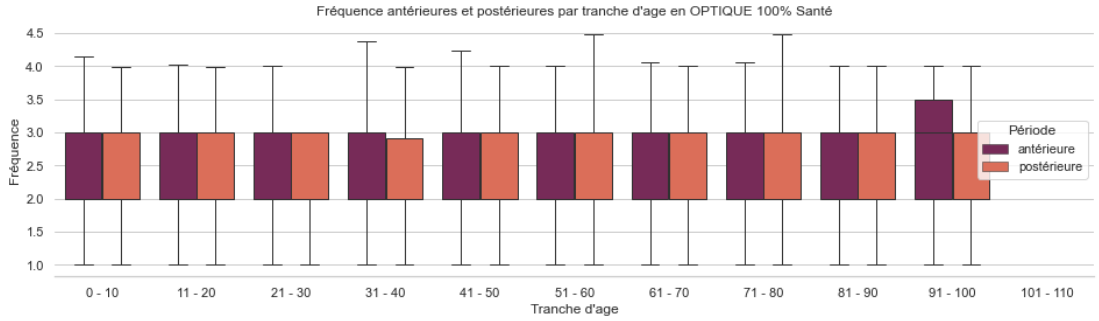


FIGURE 4.60 – **Fréquences** antérieures et postérieures en **optique 100% santé**

L'écart du coût moyen s'agrandit pour les assurés de plus de 40 ans avec un spectre beaucoup plus large. Ces coûts moyens se distinguent aussi par une hausse entre les actes antérieurs et postérieurs à l'observation. Cette hausse pourrait être impliquée par la hausse du prix des dispositifs.
Cependant la fréquence reste entre 2 et 3 actes en moyenne.

En optique 100% santé, la consommation est limitée à une paire de lunette tous les deux ans. Par la suite, la majorité des assurés ayant consommés durant la période antérieure à l'observation ne pourront pas consommer durant la période postérieure et les assurés n'ayant pas consommés durant la période antérieure sont les plus risqués à consommer durant la période postérieure. La consommation antérieure impactera considérablement la prédiction des coûts moyens et fréquences en optique.

Impact de la résiliation sur la consommation postérieure

Les coûts moyens et fréquences des actes en fonction de l'âge des assurés et selon la résiliation postérieure se présentent comme suit :

FIGURE 4.61 – Coût moyen postérieur en fonction de l'âge et la résiliation des contrats

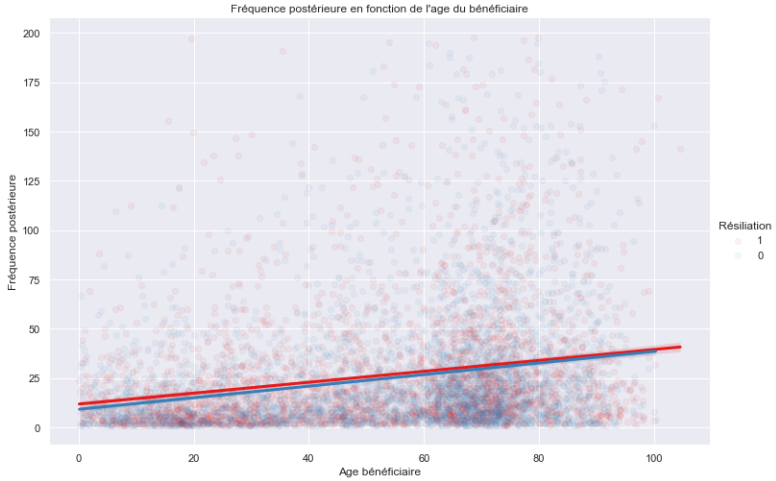


FIGURE 4.62 – Fréquence postérieure en fonction de l'âge et la résiliation des contrats

Le coût moyen postérieure des actes relatifs aux contrat résiliés croie en fonction de l'âge plus rapidement que celle des cas de non résiliation. Ce dernier est en moyenne plus important pour les assurés âgés de plus de 35 ans et inférieur pour les plus jeunes.

De l'autre coté, la fréquence postérieures suit la même évolution pour les deux cas. Cependant, cette dernière est légèrement supérieure pour les cas de résiliation sur tous les âges.

4.6 Conclusion

L'analyse descriptive nous a permis en premier lieu de découvrir la composition du portefeuille en terme de nombre de contrats et d'assurés, d'âges et d'anciennetés ainsi que les caractéristiques de leurs contrats en termes de cotisations et niveaux de gammes et garanties. Ensuite, nous avons observé l'évolution des remboursements mutuelle et les facteurs impactant la consommation tels que la saisonnalité annuelle des remboursements, l'entrée en vigueur du 100% santé et la crise covid.

Concernant la modélisation, nous avons constaté que les résiliations étaient surtout impactées par les variables telles que l'âge des adhérents, le nombre de bénéficiaires des contrats, les niveaux des gammes et des garanties, la durée avant l'éligibilité à résilier, les indexations antérieures et postérieures aux observations et la consommation antérieure à la date d'observation.
Pour la durée de couverture restante, nous reconnaissons un énorme changement de la distribution de cette variable suite à l'entrée en vigueur de la résiliation infra-annuelle. Cependant, une partie importante des résiliation continue à se produire en fin d'année, date d'anniversaire de la plupart des contrats. Le mois d'observation reste une variable importante dans ce cas pour la prédiction de la durée de couverture restante à coté de la durée avant éligibilité qui encadre cette dernière dans notre horizon postérieur d'un an.
Enfin, la consommation postérieure diffère d'un poste d'actes à un autre. Cette dernière dépend fortement de la consommation antérieure et de l'âge des assurés. D'autre part, nous avons vu que qu'une résiliation postérieure d'un contrat peut avoir une influence sur la consommation postérieure de ses bénéficiaires.

La prochaine étape consiste à la modélisation de ces variables par le biais de modèles de machine learning et de modèles statistiques.

Chapitre 5

Modélisations

Après avoir pris connaissance des différentes variables à expliquer et étudié les différentes variables explicatives à disposition, nous passons aux méthodes employées pour les prédictions à effectuer dans le processus de modélisation mis en place.

Avant d'entamer la modélisation, nous allons retirer une portion des contrats du portefeuille, présents au 1 janvier 2021, pour évaluer le modèle obtenu à l'aide d'une application au cours du prochain chapitre.

Un total de 28 060 contrats, éligibles à la résiliation, sont concernés sur toute la période de couverture étudiée. Parmi ces contrats, 19 883 sont présents au 1 mai 2021. À cette date d'observation, ces contrats se présentent comme suit :

Nombre contrats présents	Cas de non-résiliation	Futures résiliations
19 883	17 332	2 551
	87,17 %	12,83 %

Table 5.1 – Contrats observés le 1 janvier 2021

Nous effectuons un tirage aléatoire de 30% de ces contrats pour la partie "application". Nous obtenons la répartition suivante :

	Nombre contrats	Non-résiliation	Résiliation
Application	5 965	5 214	751
		87,40 %	12,59 %
Modélisation	13 918	12 118	1 800
		87,07 %	12,93 %

Table 5.2 – Répartition des contrats observés le 1 janvier 2021

En retirant la liste des contrats, présents au 1 janvier 2021 et dédiés à l'application ultérieure du modèle, de l'ensemble des contrats observés sur la période de couverture étudiée, nous obtenons une liste de **22 095 contrats** qui serviront à l'apprentissage et la validation des modèles.

5.1 Présentation des modèles employés

Pour rappel, la modélisation de l'évolution de notre portefeuille se déroule en deux étapes principales.

La première étape consiste à modéliser l'évolution du flux sortant de contrats du portefeuille dans un horizon d'un an. Dans cette étape, nous allons avoir recours à une classification pour prédire la résiliation des contrats dans un horizon d'un an puis une régression pour la prédiction des durées de couverture restantes pour les contrats à risque de résilier. Pour mener à bien cette première étape primordiale pour notre modélisation, nous avons opté pour le célèbre algorithme **eXtreme Gradient Boosting (XGBoost)**. En haut du podium dans la majorité des compétitions d'apprentissage automatique sur la plateforme populaire *Kaggle* pour sa performance et sa rapidité d'apprentissage, XGBoost nous permettra, du fait qu'il est constitué d'arbres de décision et qu'il relève des techniques d'apprentissage ensembliste, nous permet d'exploiter l'ensemble de nos variables explicatives sans nous soucier de la corrélation entre ces dernières et du sur-apprentissage.

Pour l'évolution de la consommation dans un horizon d'un an, nous procéderons par la méthode coût-fréquence pour le calcul de la prime pure de chaque bénéficiaire pour chaque poste d'actes en tenant compte de sa probabilité de résiliation et de sa durée d'exposition future dans notre horizon. Les **Modèles Linéaires Généralisés (GLMs)** restent les algorithmes les plus prudents pour approximer les distributions de coûts moyens et de fréquences des actes par grand poste.

5.1.1 Introduction à eXtreme Gradient Boosting (XGBoost)

XGBoost est un algorithme d'apprentissage supervisé. Il fait partie des algorithmes d'apprentissage ensembliste et est basé sur la technique de **Boosting**, qui consiste à entraîner de manière successive des modèles (qui représentent ici des arbres de décision) de façon à ce que chaque modèle prennes en compte les erreurs des modèles précédents.

XGBoost est basé plus précisément sur le modèle Gradient Boosting, il ajoute constamment de nouveaux arbres de décision pour ajuster une valeur Y et améliorer l'efficacité et la performance des modèles d'entraînement. Contrairement au modèle gradient boosting, XGBoost utilise une expansion de Taylor pour approximer la fonction de perte, et le modèle présente un meilleur compromis entre biais et variance, utilisant généralement moins d'arbres de décision pour obtenir une meilleure précision.

Une description formalisée de XGBoost, issue du document [3], se présente comme suit :

Soit n le nombre d'observations et m le nombre de variables explicatives formant le vecteur X tel que Y est la valeur observée. La valeur prédite finale $\widehat{Y}$ est par la suite obtenue via la sommation des résultats de K arbres de décision qui composent le modèle,

ce qui nous donne l'expression suivante :

$$\widehat{y}_i = \sum_{k=1}^{K} f_k\left(x_i\right), \quad f_k \in F$$

Avec F désigne l'espace des arbres de décision tel que :

$$F = \left\{f(x) = w_{q(x)}\right\} \quad \left(q : R^m \to T, w \in R^T\right)$$

q représente la structure de chaque arbre de décision, T le nombre de feuilles dans l'arbre et w le poids des feuilles. De cette façon, chaque f_k correspond à une structure indépendante q ayant des poids de feuilles w.

L'optimisation du modèle obtenu est assurée par la minimisation d'une fonction objective qui se présente comme suit :

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right),$$

Cette fonction de compose d'une fonction perte l, qui mesure l'écart entre les observations et les prédictions correspondantes et une fonction Ω pénalisant la complexité du modèle pour éviter le sur-apprentissage.

La fonction perte l doit être convexe et différentiable pour que son minimum puisse être facilement atteint. Elle pourrait correspondre à la méthode des moindres carrés (mean-squared) dans le cas d'une régression ou bien de l'entropie croisée (cross-entropy) dans le cas d'une classification.

Quant à la fonction Ω, elle s'exprime comme suit :

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2.$$

Ici, γ et λ sont des hyper-paramètres du modèle. γ agit sur le nombre de feuilles des arbres de décision et λ contrôle la précision de l'arbre. Plus ces paramètres sont élevés, moins le modèle sera complexe.

L'optimisation de la fonction objective développée dans le document [3] nous permet d'avoir, pour une feuille j, le poids optimal correspondant ci-dessous :

$$w_j = -\frac{\sum_{i \in j} g_i}{\sum_{i \in j} h_i + \lambda}$$

Avec :

— g_i : gradient de la fonction objective L.
— h_i : dérivée seconde de la fonction L.
— j : index de la feuille de l'arbre.

5.1.2 Introduction aux modèles linéaires généralisés (GLMs)

Le modèle linéaire généralisé (GLM) est une généralisation du modèle linéaire simple, tel que les résidus du modèle peuvent avoir une distribution différente d'une loi normale. Le GLM permet principalement de modéliser une variables à **expliquer** Y en fonction d'une ou de plusieurs **variables explicatives** X et est caractérisé par :

— Une variable à expliquer $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$ provenant d'une distribution qui appartient à la famille exponentielle définie par la densité :

$$f_{Y_i}(y_i, \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right)$$

.

— Une composante déterministe : $\eta = \mathbf{X}\beta = \beta_0 + \sum_{i=1}^{p} X_i\beta_i$, avec $X_i = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$ et $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix}$ est un vecteur de coefficient inconnus à estimer par le modèle.

— Une fonction de liaison monotone, inversible et différentiable $g(.)$ reliant le prédicteur linéaire η à la variable à expliquer Y. Étant donné que $g(.)$ est inversible, il existe une fonction inverse telle que : $\mu = g^{-1}(\eta_i) = \mathrm{E}(y_i)$

L'équation générale du modèle est la suivante :

$$g(\mathrm{E}(Y|X)) = \eta$$

Le tableau suivant répertorie les modèles GLM que nous utiliserons par la suite :

Modèle	Loi	Support	Fonction de lien
Log-Normale	Normale	$\mathbb{R}^+$	$g(x) = ln(x)$
Log-Gamma	Gamma	$\mathbb{R}^+$	$g(x) = ln(x)$
Log-Poisson	Poisson	$\mathbb{N}$	$g(x) = ln(x)$

5.2 Outils d'évaluation des modèles

Afin, d'évaluer les modèles prédictifs déjà présentés, des indicateurs de performance ont été présentés. En voici quelques-uns :

Matrice de confusion

La matrice de confusion est un outil standard pour évaluer les performances d'un modèle de classification. Elle peut être appliquée à la classification binaire ainsi qu'aux problèmes de classification multi-classes. Le tableau 5.3 présente un exemple de matrice de confusion pour une classification binaire.

	Négative prédit	Positive prédit
Négative actuel	TN	FN
Positive actuel	FN	TP

TABLE 5.3 – Matrice de confusion pour une classification binaire

À partir de la matrice de confusion, les termes suivants peuvent être définis :

— **Sensibilité :** correspond au taux de vrais positifs :

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

— **Spécificité** correspond au taux de vrai négatif :

$$\frac{TN}{TN + FP}$$

— **Précision** est le taux de prédictions correctes parmi les prédictions positives :

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

Courbe ROC et le critère AUC

Dans l'analyse des résultats d'une classification binaire, la courbe ROC (*receiver operator characteristic*) est très utilisée pour montrer la performance d'un modèle. La courbe ROC donne des informations sur les performances pour une série de seuils de décision et peut être résumée par l'aire sous la courbe (AUC).

L'AUC (*area under the ROC curve*) est un indicateur plus global pour évaluer et comparer la performance des modèles de classification, indépendamment du seuil de décision choisi. L'AUC est comprise entre 0 et 1. Un score proche de 0,5 signifie que le modèle est équivalent à un modèle de classification aléatoire, ensuite plus le score est élevé, meilleure est la classification du modèle.

Erreur quadratique moyenne (RMSE) et erreur absolue moyenne (MAE)

Ce sont deux indicateurs régulièrement utilisées pour évaluer les modèles et ils sont donnés par les formules suivantes :

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

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

avec : $\hat{y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{pmatrix}$ sont les valeurs prédites par le modèle et $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$ sont les valeurs observées.

5.3 Modèle 1 : Prédiction de la résiliation des contrats à horizon 1 an de la date d'observation

La base de données à disposition pour la modélisation de la résiliation des contrats dans un horizon d'un an se compose d'observations effectuées mensuellement de chaque contrat du portefeuille liant ses données relatives à une année antérieure de chaque observation et les évènements à venir, tel qu'une prochaine indexation, avec sa résiliation éventuelle. Les observations du portefeuille sont effectuées du 1 janvier 2019 jusqu'au 1 mai 2021. Certaines observations, dans le cas de résiliation, sont effectuées après cette date.

Notre base de données se compose comme suit :

Nombre d'observations	Non-résiliation	Résiliations
479 819	382 936	96 883
100 %	79,81 %	20,19 %

Table 5.4 – Base de données

La base de données est composée des variables suivantes :

- Ancienneté contrat
- Gamme
- Niveau de gamme
- Niveau de garantie
- Motif adhésion
- Département tarifaire
- Age adhérent
- Présence conjoint
- Age conjoint
- Nombre enfants
- Age moyen enfants
- Année d'observation
- Mois d'observation
- Durée depuis souscription
- Durée d'observation antérieure
- Durée avant éligibilité
- Durée avant indéxation
- Cotisation annuelle
- Indexation annuelle
- Indexation postérieure
- Hospi covid antérieures
- Confinement antérieur
- Hospi covid postérieures
- Confinement postérieur
- Durée soins - Hospitalisation
- Fréquence antérieure - Soins
- Fréquence antérieure - Hospitalisation
- Fréquence antérieure - Dentaire
- Fréquence antérieure - Optique
- Fréquence antérieure - Auditif
- Fréquence antérieure - Pharmacie
- Fréquence antérieure - Dentaire RAC0
- Fréquence antérieure - Optique RAC0
- Fréquence antérieure - Auditif RAC0
- Fréquence antérieure - Autre
- Cout moyen antérieur - Soins
- Cout moyen antérieur - Hospitalisation
- Cout moyen antérieur - Dentaire
- Cout moyen antérieur - Optique
- Cout moyen antérieur - Auditif
- Cout moyen antérieur - Pharmacie
- Cout moyen antérieur - Dentaire RAC0
- Cout moyen antérieur - Optique RAC0
- Cout moyen antérieur - Auditif RAC0
- Cout moyen antérieur - Autre
- RAC antérieur - Hospitalisation
- RAC antérieur - Optique
- RAC antérieur - Dentaire
- Résiliation

La matrice de corrélation des variables se présente comme suit :

FIGURE 5.1 – Matrice de corrélation de la base de données complète

Le modèle utilisé dans cette partie étant non-paramétrique, cette matrice a été construite avec les coefficients de corrélation de **Kendall** (source).

Le modèle utilisé tel qu'il est construit n'est pas impacté par les quelques corrélations présentes dans notre base de données.

Dans ce qui suit, nous allons essayer plusieurs classifications. L'entraînement et le test de chacune de ces classifications est basée sur une portion donnée du portefeuille afin de repérer la partie la mieux adaptée à nos prédictions.

5.3.1 Apprentissage des modèles et évaluation des performances

Prise en compte de l'ensemble des observations

Dans un premier temps, nous allons entraîner et tester un modèle XGBoost avec les données provenant de l'ensemble des observations à disposition. Pour ce faire, nous séparons la base en deux parties destinées à l'entraînement et le test du modèle.

Pour éviter d'avoir des lignes correspondant à un même contrat dans les deux bases, la séparation des deux parties est basée sur une sélection de contrats. En effet, le nombre de lignes référant à un seul contrat dans la base est égale au nombre de mois de couverture de ce contrat sur la période étudiée.

Avec un tirage aléatoire de 30% des contrats pour la partie "test", la décomposition obtenue des contrats se présente comme suit :

Nombre contrats	Contrats pour l'entraînement	Contrats pour le test
22 095	15 466	6 629
	70 %	30 %

Les observations relatives à ces contrats se présentent comme suit :

	Nombre d'observations	Non-résiliation	Résiliation
Entraînement	335 613	267 319	68 294
		79,65 %	20,35 %
Test	144 206	115 617	28 589
		80,17 %	19,83 %

Suite à l'entraînement et le test du modèle, avec un score d'entraînement (pourcentage de bons classements) respectivement de test de 87,3 % respectivement 84,49 %, les résultats obtenus se présentent comme suit :

Importance des variables : F-score

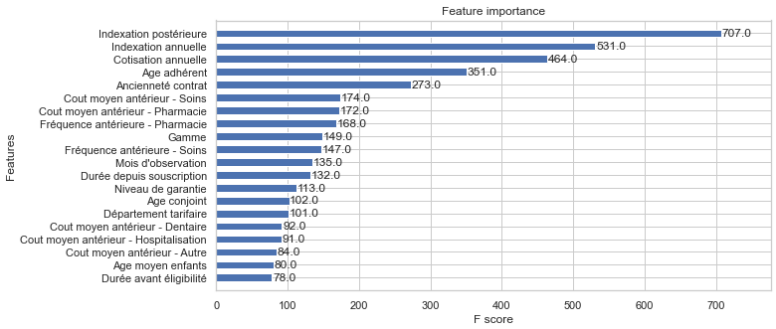


FIGURE 5.2 – F-score des variables

Le F1-score de chaque variable est calculé depuis la fréquence avec laquelle cette dernière est sollicitée par l'arbre de décision dans les différentes prédictions du modèle. Ce score reflète l'implication de chaque variable dans les prédictions obtenues.

Parmi les variables les plus impactantes du modèle, nous retrouvons en première position les variables relatives aux cotisations et indexations antérieures et postérieures à l'observation. Ensuite l'âge de l'adhérent, l'ancienneté du contrat et la gamme, puis la consommation liée au contrat pour les postes pharmacie et soins de ville.

Matrice de confusion

La matrice de confusion des prédictions de la base test se présente comme suit :

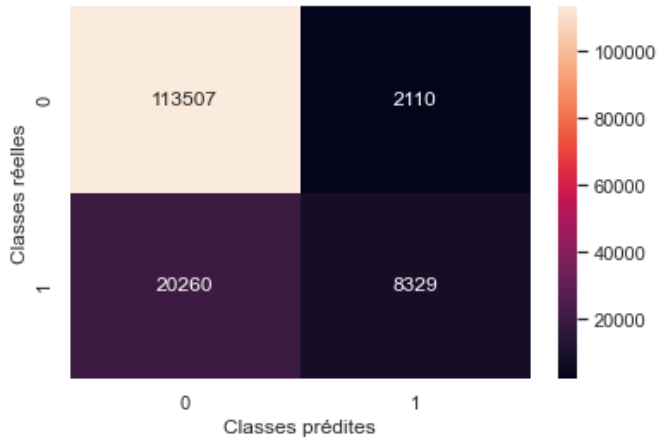


FIGURE 5.3 – Matrice de confusion pour la base test

	precision	recall	F1-score	Support
Non-résiliation	0,85	0,98	0,91	115 617
Résiliation	0,80	0,29	0,43	28 589

Du coté de la sensibilité du modèle, nous parvenons à classer 29 % des résiliations de la base test avec une précision de 80%. Du coté de la spécificité, le modèle parvient à classer 98% des cas de non résiliation avec une précision de 85%.

Courbe ROC

La courbe ROC de l'entraînement et du test du modèle se présente comme suit :

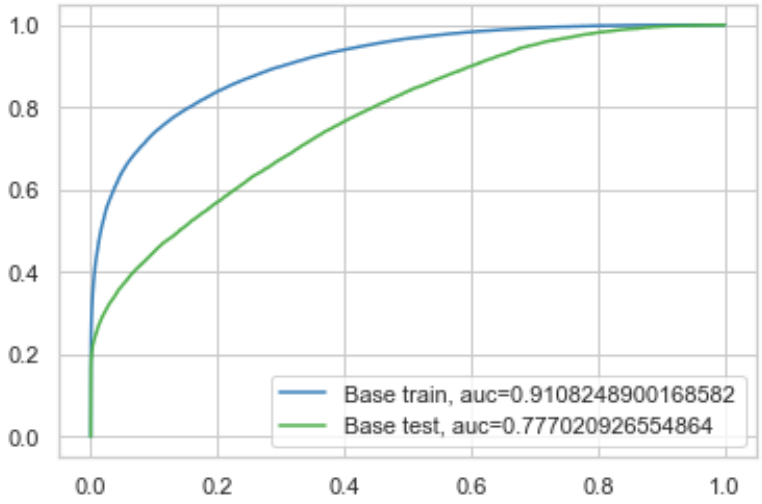


FIGURE 5.4 – Courbe ROC

Nous obtenons un AUC de 77,7 % pour le test de ce modèle et 91,08 % pour l'entraînement.

Avec une précision à 80%, la part des résiliations que le modèle parvient à repérer reste peu satisfaisante (29% des résiliations), la part des résiliations mal classées ont des profils très proches à la majorité des cas de non résiliation. Certaines observations de cette partie peuvent êtres relatives aux contrats observés avant l'entrée en vigueur de la résiliation.

Prise en compte des contrats présents à partir de 2020

Pour cette partie, nous allons prendre en compte uniquement les contrats qui sont présents en 2020 (qui n'ont pas résiliés en 2019), soit 18 889 contrats. Les contrats liées au test du modèle représentent toujours 30% des contrats tirés aléatoirement et présentent comme suit :

Nombre contrats	Contrats pour l'entraînement	Contrats pour le test
18 889	13 222	5 667
	70 %	30 %

Les observations relatives à ces contrats se présentent comme suit :

	Nombre d'observations	Non-résiliation	Résiliation
Entraînement	315 479	267 024	48 455
		84,64 %	15,36 %
Test	134 610	114 135	20 475
		84,79 %	15,21 %

Les données prises en compte incluent des observations en 2019 des contrats présents après 2020. De ce fait, ces observation de contiennent pas de cas de résiliation. Afin de ne pas biaiser le modèle, toutes les variables en lien avec l'année d'observation n'ont pas été prises en compte ("Année d'observation","Hospi covid antérieures","Hospi covid postérieures","Confinement antérieur" et "Confinement postérieur"). La prise en compte des observations de 2019 est dans le but de permettre au modèle la détection du changement de comportement entre les cas de non résiliation et les cas de résiliation pour un même contrat.

Suite à l'entraînement et le test du modèle, avec un score d'entraînement respectivement de test de 91,68 % respectivement 88,67 %, les résultats obtenus se présentent comme suit :

Importance des variables : F-score

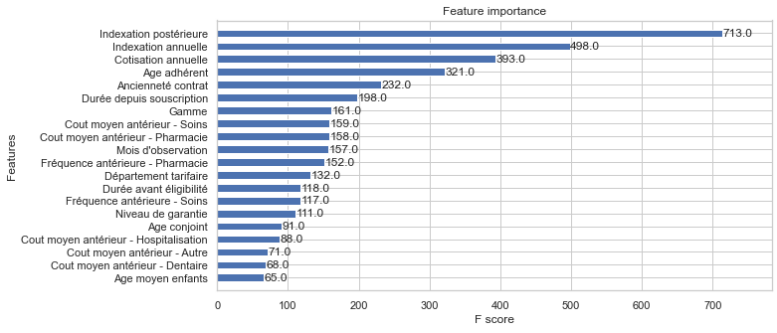


FIGURE 5.5 – F-score des variables

Les cinq variables les plus sollicitées dans la prise de décision restent les mêmes pour ce modèle. En revanche, les variables relatives à la consommation antérieure des contrats passent derrière la "gamme" et la "durée depuis souscription" du contrat (cette durée est différente de l'ancienneté du contrat car prend en compte la dernière souscription dans le cas d'un changement de gamme ou de niveau de garantie).

Matrice de confusion

La matrice de confusion des prédictions de la base test se présente comme suit :

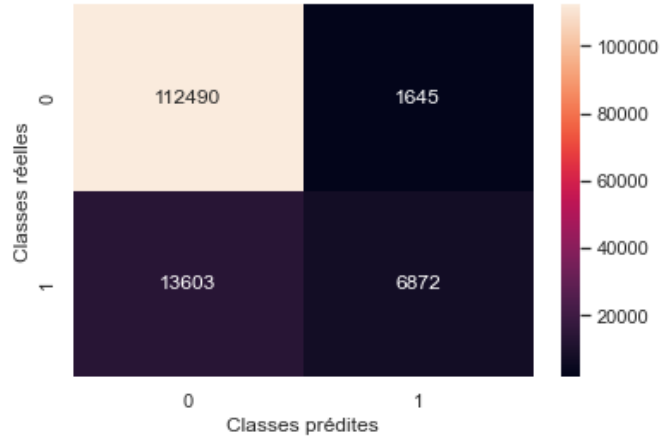


FIGURE 5.6 – Matrice de confusion pour la base test

	precision	recall	F1-score	Support
Non-résiliation	0,89	0,99	0,94	114 135
Résiliation	0,81	0,34	0,47	20 475

En ignorant les contrats résiliés avant 2020, nous obtenons une meilleure sensibilité du modèle à détecter les cas de résiliation (nous passons de 29% à 34%) avec une meilleure précision à hauteur de 81%.

Concernant les cas de non résiliations, elles sont à leur tour mieux prédites avec une spécificité de 99% pour une précision à 89%.

Courbe ROC

La courbe ROC de l'entraînement et du test du modèle se présente comme suit :

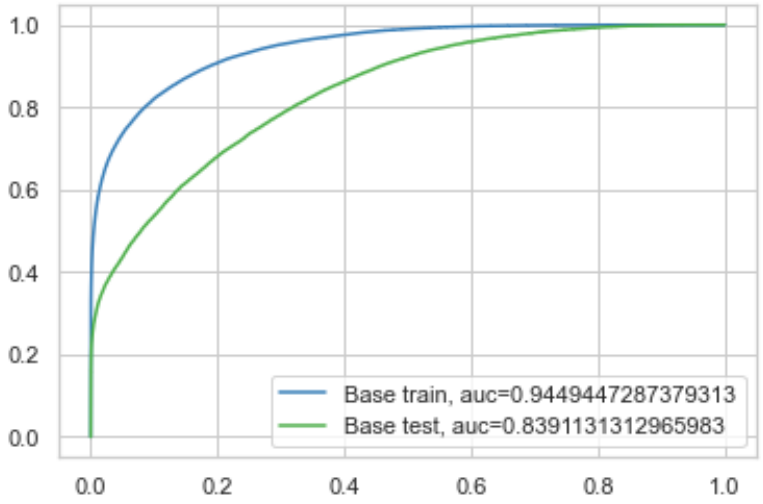


FIGURE 5.7 – Courbe ROC

Du coté de la courbe ROC, nous avons une amélioration du critère AUC qui passe de 77,7% à 83,9% pour la base "test", suivie d'un meilleur apprentissage du modèle avec AUC d'entraînement à 94,49%.

Avec l'exclusion des résiliations datant de 2019, nous obtenons une meilleur prédiction. Cela peut être l'effet de l'entrée en vigueur de la résiliation infra-annuelle ce qui rend la décision des adhérents plus cohérentes avec leurs données antérieures.

Dans le même esprit, nous allons nous restreindre aux contrats présents à partir de 2021 une fois que la résiliation infra-annuelle est déjà installée.

Prise en compte des contrats présents à partir de 2021

Maintenant, nous allons prendre en compte uniquement les contrats qui sont présents en 2021 (qui n'ont pas résiliés avant 2021), soit 15 927 contrats. Les contrats liées au test du modèle représentent toujours 30% des contrats tirés aléatoirement et se présentent comme suit :

Nombre contrats	Contrats pour l'entraînement	Contrats pour le test
15 927	11 148	4 779
	70 %	30 %

Afin de ne pas avoir une base trop déséquilibrée (pas de cas de résiliations avant 2021), les observations, relatives aux contrats choisis, ne sont prises en compte qu'à partir de 2020. Ces observations se présentent comme suit :

	Nombre d'observations	Non-résiliation	Résiliation
Entraînement	174 788	148 942	25 846
		85,21 %	14,79 %
Test	74 769	64 248	10 521
		85,93 %	14,07 %

L'apprentissage du modèle se termine avec un score d'entraînement, respectivement de test, de 95,47%, respectivement 92,73%. Les résultats obtenus se présentent comme suit :

Importance des variables : F-score

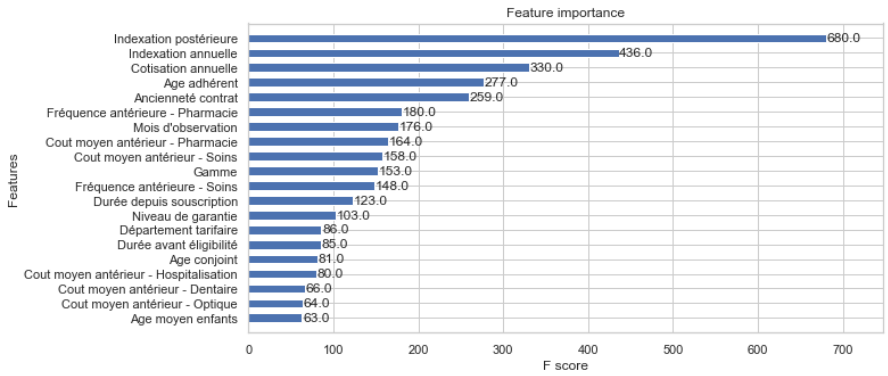


FIGURE 5.8 – F-score des variables

Les cinq variables les plus impactantes restent toujours les mêmes. Cependant, la consommation antérieure en soins de ville et en pharmacie sont de nouveaux en tête pour le reste des variables.

Matrice de confusion

La matrice de confusion des prédictions de la base test se présente comme suit :

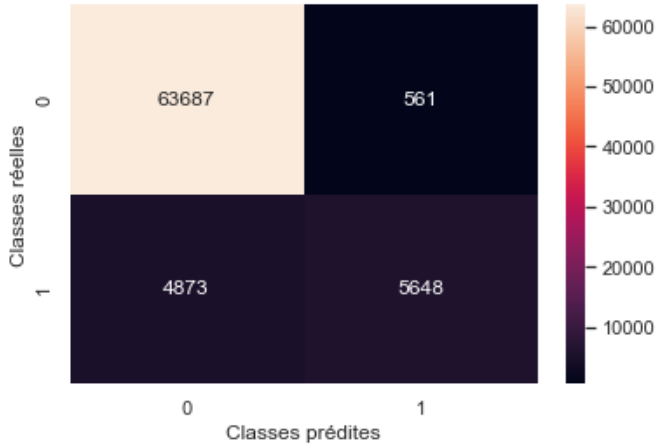


FIGURE 5.9 – Matrice de confusion pour la base test

	precision	recall	F1-score	Support
Non-résiliation	0,93	0,99	0,96	64 248
Résiliation	0,91	0,54	0,68	10 521

La sensibilité du modèle a atteint 54% avec une précision de 91%. La prise en compte des contrats présents à l'entrée en vigueur de la résiliation infra-annuelle nous a permis de détecter plus de résiliations avec une meilleure précision.

Au niveau des cas de non résiliations, nous reconnaissons également une amélioration avec une spécificité de 99% avec une précision de 91%.

Courbe ROC

La courbe ROC de l'entraînement et du test du modèle se présente comme suit :

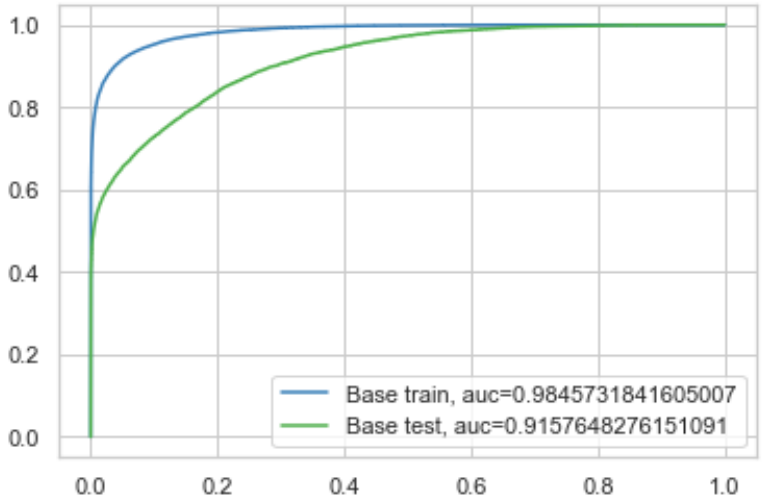


FIGURE 5.10 – Courbe ROC

Le critère AUC de ce modèle s'élève à 91,58% pour la base de test et est de 98,46% pour la base d'entraînement. Les capacités discriminatoires de ce modèle sont bien meilleures que celles des deux modèles précédents.

Le F-Score ne reflète pas vraiment l'impact de la valeur que prend une variable pour la prédiction mais plutôt la fréquence moyenne avec laquelle cette variable a servi pour la classification. Afin de mieux comprendre l'impact des variables explicatives sur la classification du modèle, nous avons eu recours à la méthode SHAP.

5.3.2 Interprétation de la classification par SHAP

La méthode SHAP[6] nous permettra dans cette partie d'interpréter la contribution de chaque variable, en fonction de la valeur qu'elle prend, sur la prédiction des résiliations.

Interprétation de la contribution des variables

En calculant les valeurs de Shapley des toutes les variables pour chaque observation de la base "test", nous obtenons le graphique suivant :

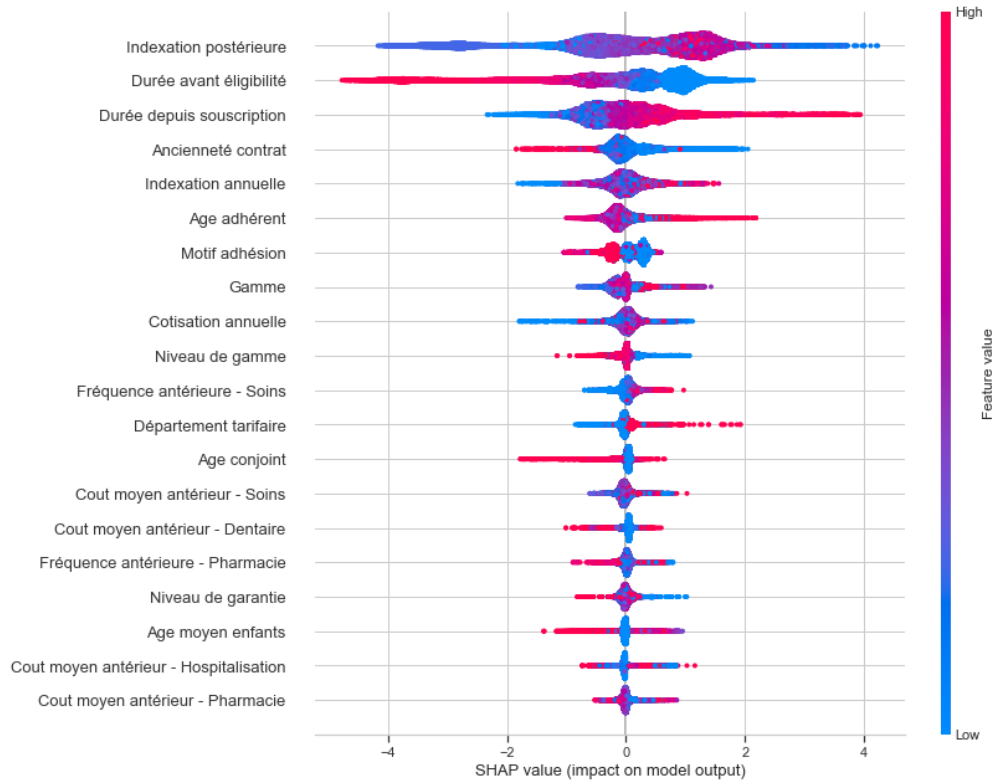


FIGURE 5.11

Ce graphique contient l'interprétation des 20 variables les plus impactantes sur les prédictions. Les couleurs des points varient en fonction de la valeur que prend la variable en question (du bleu pour les valeurs les plus faibles au rose pour les valeurs les plus hautes). Sur le graphe, les points sont représentés en fonction de la valeur de Shapley. Les valeurs positives signifient que la valeur prise par la variable en question augmente la probabilité de résiliation.

D'après les valeurs de Shapley, la variable la plus influente sur la prédiction est l'**indexation postérieure**. La représentation montre que la majorité des indexations postérieures élevées ont fait augmenter la probabilité de résiliation des contrats concernés.

La **durée avant éligibilité** représente la deuxième variable la plus influente sur les prédictions du modèle. On voit qu'une durée avant éligibilité haute diminue la probabilité de résiliation d'un contrat. En effet, plus la valeur que prend cette variable est grande plus la période pendant laquelle un contrat peut être résilié dans l'horizon postérieur est faible. De la même façon, les valeurs faibles de cette variable représentants des contrats qui sont éligibles à la résiliation tout au long de l'année sont représentés avec des valeurs de Shapley positives qui augmentent la probabilité de résiliation.

La troisième variable est la **durée depuis souscription**. Nous pouvons voir que cette variable est croissante en fonction de la valeur de Shapley. En effet, nous avons vu dans le chapitre précédent que la majorité des contrats changent de gamme ou de niveau de garantie au bout de 5 ans, cette variable réfère à la date de souscription à la gamme et au niveau de garantie actuels. Ainsi une valeur élevée de cette durée augmente la probabilité de résiliation.

De l'autre coté, l'**ancienneté des contrats** est décroissante en fonction de la valeur de Shapley. Cela reflète la fidélité des anciens adhérents.

Afin d'interpréter l'influence de l'**âge de l'adhérent** sur la résiliation des contrats, nous traçons la valeur de Shapley en fonction de l'âge de l'adhérent (nous prendrons en compte la présence d'un conjoint ou non dans le contrat pour une meilleure interprétation) :

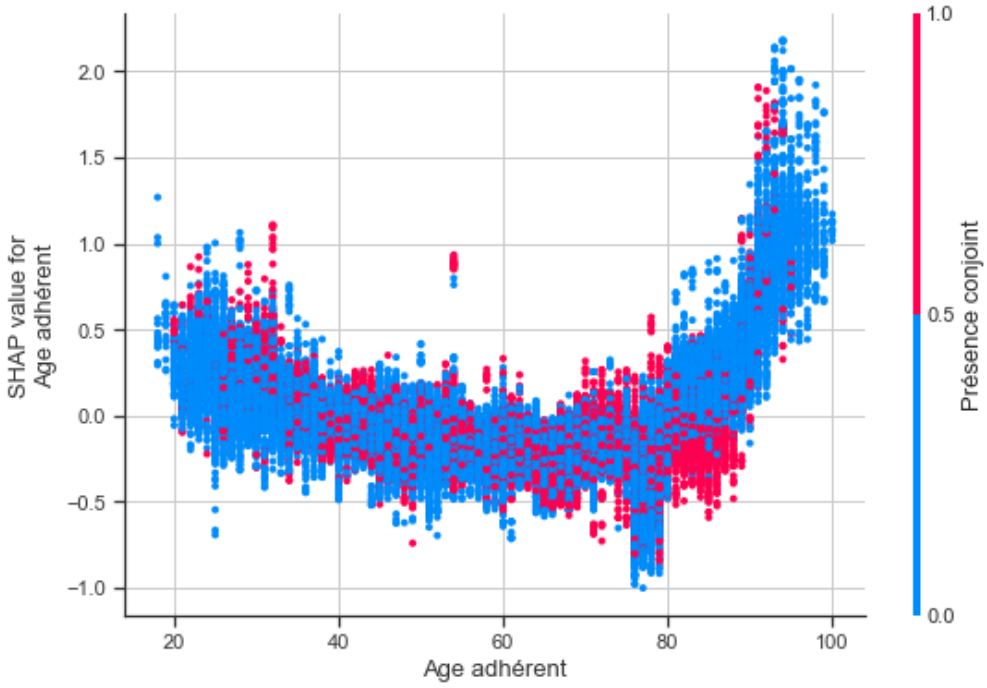


FIGURE 5.12 – Valeur de Shapley en fonction de l'âge des adhérents

Comme nous l'avons remarqué en comparant les distributions de l'âge des adhérents en fonction de la variable résiliation, les résiliations concernent surtout les jeunes de moins de 40 an et les plus âgés (probablement à cause des décès). Les valeurs de Shapley nous le confirment par le biais de cette représentation, les valeurs de Shapley décroient pour les plus jeunes et croient pour les plus âges. En particuliers, nous retrouvons que la présence d'un conjoint pour les adhérents âges de plus de 80 ans baisse la probabilité de résiliation.

Enfin, une moyenne des valeurs absolues des valeurs de Shapley observées pour chaque variable nous donne une mesure de l'importance des variables dans la prédiction de notre modèle qui se représente comme suit :

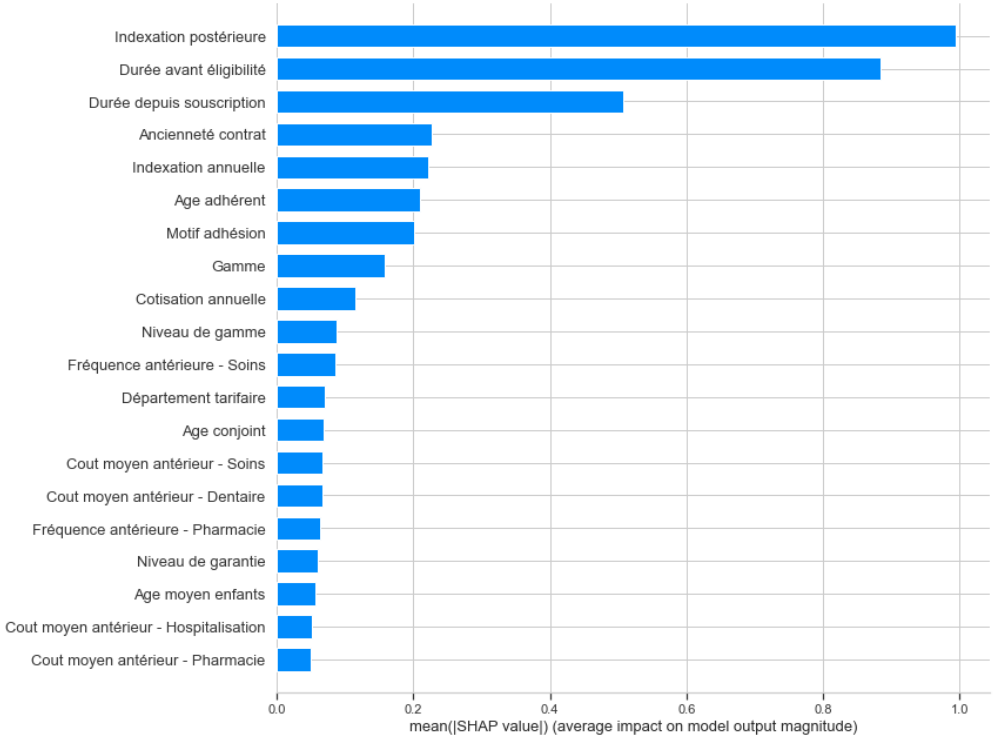


FIGURE 5.13 – Importance des variables par la méthode Shapley

5.3.3 Conclusion

Dans cette partie nous avons construit un modèle de prédiction de la résiliation des contrats qui réussit à prédire prêt de 50 % des résiliations à l'horizon d'un an avec une précision de prêt de 90%. Ce modèle est construit à l'aide des données de contrats qui ont résiliés après l'entrée en vigueur de la résiliation infra-annuelle. Les variables les plus importantes dans la classification obtenue sont l'indexation postérieure suivie de la durée avant éligibilité et la durée depuis la souscription, ces variables possèdent des valeurs de

grande sensibilité telles qu'une durée avant souscription et une indexation postérieure très élevées qui résultent une résiliation dans la plupart des cas, et une spécificité élevée pour la variable durée avant éligibilité quand cette dernière est élevée.

D'autre part, nous avons vu que plus les résiliations prises en compte dans la construction du modèle sont avancées dans le temps, plus le modèle est performant. Cela indique que la décision de résiliation des contrats se base de plus en plus sur les informations antérieures des contrats des contrats. Cependant, la résiliation infra-annuelle n'est pas encore connue par l'ensemble de la population et certains adhérents ont gardés l'habitude de ne penser à la résiliation qu'à la date d'anniversaire du contrat.

Enfin, plusieurs variables influant sur les résiliations ne sont prises en compte dans notre modèle telles que l'exposition des assurés aux offres de la concurrence. Ces informations peuvent donner une meilleure performance du modèle.

5.4 Modèle 2 : Prédiction de la durée de couverture restante avant la résiliation des contrats à risques à horizon un an de la date d'observation

Dans cette partie, nous allons construire un modèle de prédiction de la durée de couverture restante, cette durée est entre 0 et 1 et les variables utilisées pour la prédiction sont les mêmes que celles utilisées pour la prédiction de la résiliation. Nous allons donc utiliser un XGBoost afin pour la modélisation de cette variable.

Comme vu précédemment, la distribution de la variable à expliquer a subi un grand changement après l'entrée en vigueur de la résiliation infra-annuelle, nous allons donc nous baser sur les résiliations effectuées à partir de l'année 2021 pour l'apprentissage et le test du modèle.

5.4.1 Apprentissage des modèles et évaluation des performances

Notre modèle se base uniquement sur les cas de résiliation, le nombre de contrats résiliés après 2021 est de 6 523 contrats, la répartition de ces contrats sur la base d'entraînement et de test et les observations liées aux contrats se présentent comme suit :

	Entraînement	Test
Nombre de contrats	5 218	1 305
	80 %	20 %
Nombre d'observations	55 169	13 761
	80 %	20 %

Suite à l'entraînement du modèle, les résultats du test obtenues se présentent comme suit :

Score d'entraînement	MAE	RMSE
0,67	0,16	0,21

Les tests de performance du modèle semblent satisfaisants.

Nous avons vu que les durées de couvertures restantes dépendent encore des mois d'observation. Nous allons donc essayer d'entraîner le modèle avec des observations du mois de janvier. Les nouvelles bases d'entraînement se présentent comme suit :

	Entraînement	Test
Nombre de contrats	2 489	623
	80 %	20 %
Nombre d'observations	27 372	6 849
	80 %	20 %

Suite à l'entraînement du modèle, les résultats du test obtenues se présentent comme suit :

Score d'entraînement	MAE	RMSE
0,78	0,16	0,21

Nous obtenons un meilleur score d'entraînement mais pas d'amélioration sur le niveau de la MAE et RMSE. Les tests de performance du modèle semblent satisfaisants.

5.4.2 Conclusion

Les résultats obtenus montrent la capacité de l'XGBoost à segmenter les données, ce dernier a su tenir compte de la différence des distributions de la variable à expliquer en fonction du mois d'observation.

La modélisation de la durée de couverture restante complète la modélisation des résiliations afin de prédire l'évolution du portefeuille en termes d'effectifs à horizon un an.

Dans ce qui suit, nous allons modéliser les coûts moyens et fréquences d'actes pour chaque groupe de poste afin de modéliser la consommation postérieure du portefeuille.

5.5 Modélisation de la consommation postérieure à la date d'observation par bénéficiaire

La prédiction de la consommation postérieure des assurés à un horizon d'un an se fait par la prédiction de leurs coûts moyen et fréquences. La fréquence et coût moyen d'un acte diffèrent selon le poste d'actes, nous allons donc modéliser ces derniers pour chaque poste.

Comme lors des derniers modèles, afin d'avoir des bases d'entraînement et de test totalement disjointes, la répartition des bases de la modélisation repose sur une sélection de contrats.

Afin de prendre en compte les actes en 100% santé, les contrats sélectionnés pour cette partie sont ceux présents à partir de 2021. La décomposition des contrats et leurs bénéficiaires pour la partie entraînement et test du modèle se présente comme suit :

	Entraînement	Test
Nombre de contrats	11 148	4 779
	70 %	30 %
Nombre de bénéficiaires	16 752	7 265
	69,75 %	30,25 %
Nombre d'observations	488 827	213 031
	69,65 %	30,35 %

La liste des postes d'actes pour lesquelles nous construirons les modèles se présente comme suit :

— Soins de ville
— Hospitalisation
— Dentaire
— Optique
— Auditif
— Pharmacie
— Dentaire RAC0
— Optique RAC0
— Auditif RAC0
— Autre

Pour chacun de ces actes nous allons dans la suite établir un modèle de prédiction du coût moyen d'un acte puis de la fréquence des actes.

5.5.1 Modèle 3 : Prédiction des coûts moyens postérieurs

Pour la prédiction du coût moyen postérieur pour chaque poste d'acte, nous allons choisir entre un modèle GLM-LogGamma et un modèle GLM-LogNormale en fonction de leurs performances pour chaque poste d'acte.

La matrice de corrélation, basée sur les coefficients de corrélation de **Pearson**, des variables utilisées pour ces modèles se présente comme suit :

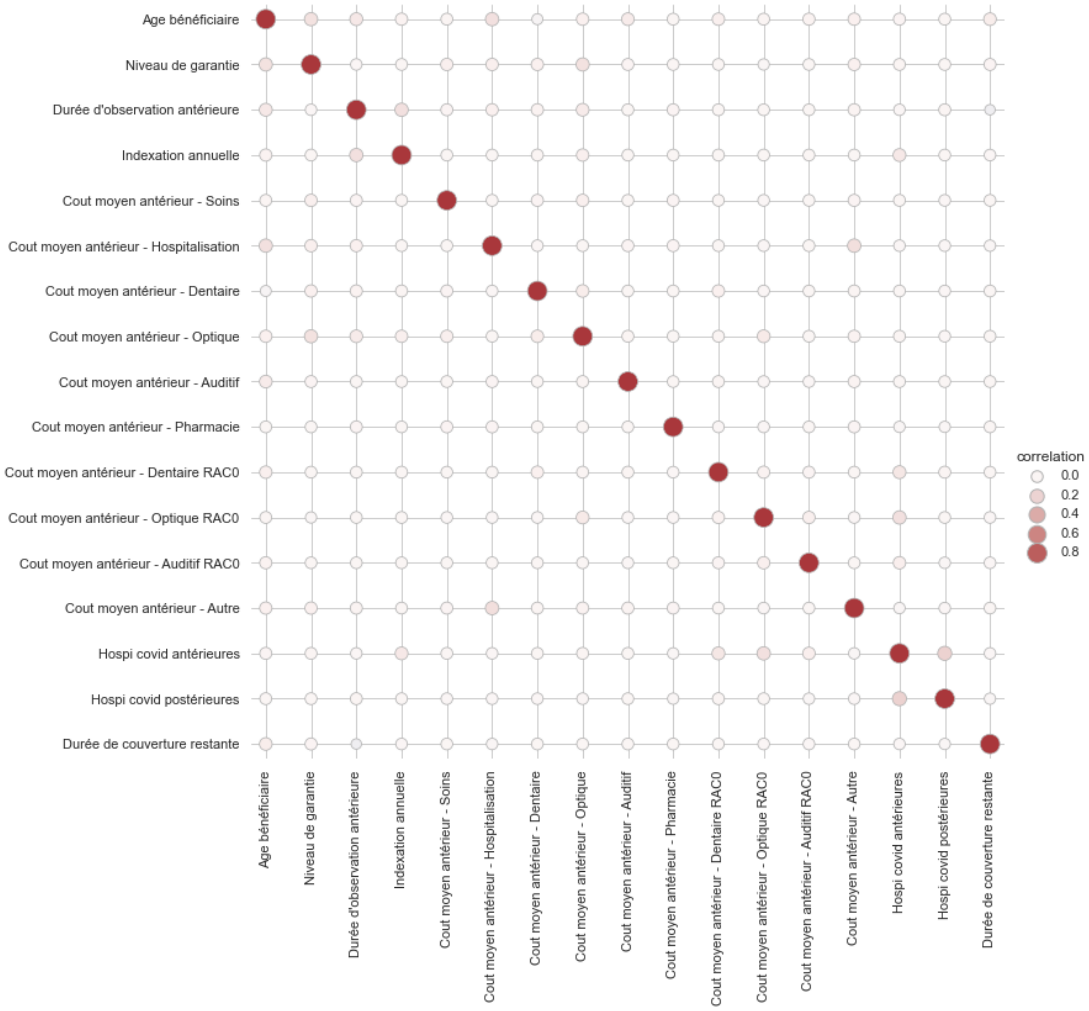


FIGURE 5.14 – Matrice de corrélation de la base destinée à la prédiction des coûts moyens

Ces variables ont été sélectionnées pour minimiser les corrélation au sein de la base. Les variables explicatives obtenues ne présentent pas de corrélations significatives entre elles.

Pour un poste d'acte "P" donné, les variables explicatives prises en compte pour la prédiction du "Coût moyen postérieur - P" se présentent comme suit :

— Âge bénéficiaire
— Niveau de garantie
— Département tarifaire
— Mois d'observation
— Année d'observation
— Type bénéficiaire
— Durée d'observation antérieure
— Indexation annuelle
— Coût moyen antérieur - P
— Hospi covid antérieures
— Hospi covid postérieures
— Durée de couverture restante

Par poste, les résultats des modèles obtenus se présentent comme suit :

Soins de ville

Modèle GLM	MAE	RMSE
Log-Gamma	3,914	6,523
Log-Normale	3,792	6,499

Hospitalisation

Modèle GLM	MAE	RMSE
Log-Gamma	63,730	93,967
Log-Normale	59,667	92,509

Dentaire

Modèle GLM	MAE	RMSE
Log-Gamma	35,744	63,443
Log-Normale	34,493	61,195

Optique

Modèle GLM	MAE	RMSE
Log-Gamma	34,996	45,545
Log-Normale	29,244	38,660

Auditif

Modèle GLM	MAE	RMSE
Log-Gamma	155,641	202,695
Log-Normale	104,661	143,665

Pharmacie

Modèle GLM	MAE	RMSE
Log-Gamma	1,383	2,789
Log-Normale	1,353	2,793

Dentaire RAC0

Modèle GLM	MAE	RMSE
Log-Gamma	472,709	767,647
Log-Normale	248,888	413,348

Optique RAC0

Modèle GLM	MAE	RMSE
Log-Gamma	18,446	23,582
Log-Normale	16,759	20,432

Auditif RAC0

Modèle GLM	MAE	RMSE
Log-Gamma	531,046	796,337
Log-Normale	104,41	218,542

Autre

Modèle GLM	MAE	RMSE
Log-Gamma	14,244	19,941
Log-Normale	13,686	19,406

Sur tous les modèles obtenus, les modèles Log-Normales présentent de meilleurs résultats. En effet, les valeurs de RMSE et de MAE pour ce modèle sont toujours inférieures à celle des modèles Log-Gamma. Le modèle Log-Normale est donc mieux adapté à nos distributions. Ainsi, nous retenons ce modèle pour la prédiction des coûts moyens postérieurs des actes.

5.5.2 Modèle 4 : Prédiction des fréquences postérieures

Pour cette prédiction de fréquence, nous utiliserons un **GLM Poisson**, la matrice de corrélation, basée sur les coefficients de corrélation de Pearson, des variables utilisées pour les prédictions se présente comme suit :

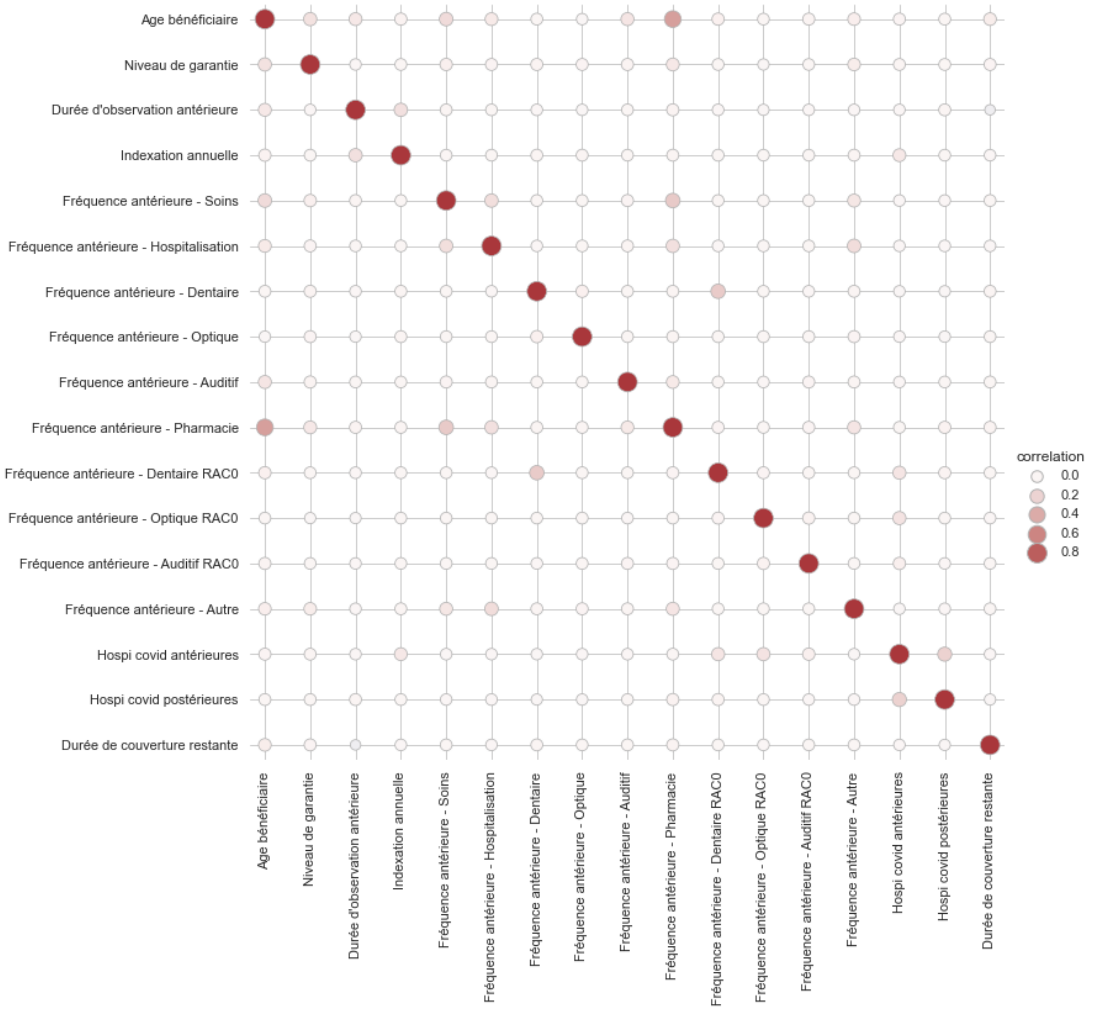


FIGURE 5.15 – Matrice de corrélation de la base destinée à la prédiction des fréquences

Cette base de données est choisie de la même façon que la base destinée aux coûts moyens afin de minimiser les corrélations entre les variables. Nous n'observons pas de corrélations significatives entre les variables, nous procédons par la suite à la modélisation de la fréquence postérieure des actes.

Pour un poste d'actes "P" donné, les variables prises en compte dans le modèle pour la prédiction de la fréquence sont les suivantes :

— Âge bénéficiaire
— Niveau de garantie
— Département tarifaire
— Mois d'observation
— Année d'observation
— Type bénéficiaire
— Durée d'observation antérieure
— Indexation annuelle
— Fréquence antérieure - P
— Hospi covid antérieures
— Hospi covid postérieures
— Durée de couverture restante

Par poste, les résultats des modèles obtenus se présentent comme suit :

Poste d'actes	**MAE**	**RMSE**
Soins de ville	16.960	33.672
Hospitalisation	1.501	2.720
Dentaire	1.283	1.956
Optique	0.995	1.351
Auditif	0.034	0.199
Pharmacie	32.456	52.046
Dentaire RAC0	0.133	0.350
Optique RAC0	0.0141	0.028
Auditif RAC0	13,686	19,406
Autre	0.424	0.809

La précision des modèles obtenues par rapport à chaque poste d'actes est acceptable. Les modèles obtenues semblent performants.

5.5.3 Conclusion

Dans cette partie, notre modélisation de la consommation postérieure d'un bénéficiaire est obtenue par la prédiction de son coût moyen et de sa fréquence pour chaque poste d'actes. Cette prédiction prend en compte la durée de couverture restante d'un bénéficiaire et le mois d'observation afin de tenir compte de l'impact de la saisonnalité sur la consommation.

Chapitre 6

Application de l'évolution des effectifs pour une portion du portefeuille

Dans cette partie, nous allons mettre en oeuvre le modèle de résiliation des contrats au cours d'un cas d'application sur une portion de contrats retirée du portefeuille au début de la modélisation. Cette portion concerne 5 965 contrats tous présents au 1 mai 2021.

Le cas d'application consiste à prédire l'évolution du nombre de contrats présents à l'horizon d'un an à partir d'une observation effectuée le 1 mai 2021 pour les contrats en question. L'évolution de la pyramide des âges via l'implémentation du modèle se présente comme suit :

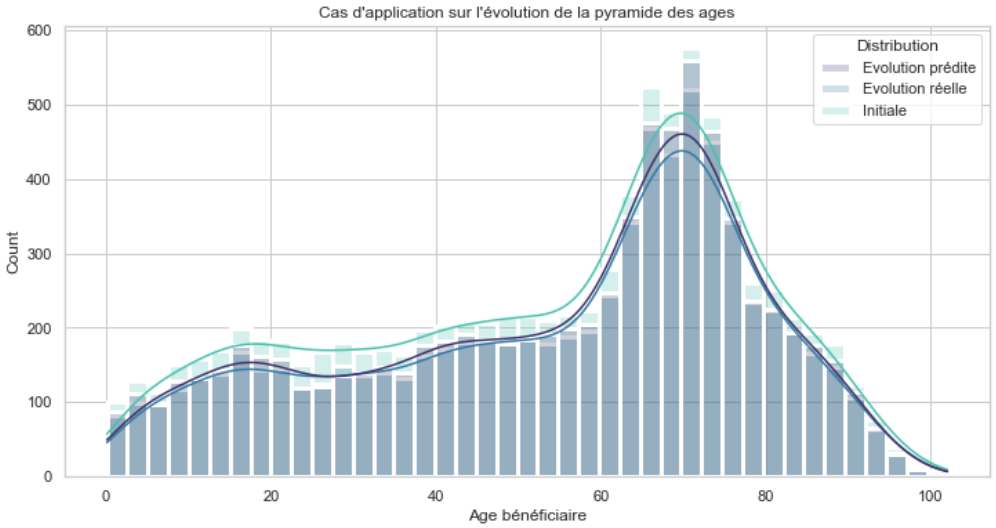


FIGURE 6.1 – Prédiction de l'évolution de la pyramide des âges

La prédiction du modèle de l'évolution de la pyramide des âges est très proche de la réalité. Cependant, cette dernière reste légèrement supérieure à l'observation réelle de la pyramide au bout d'un an, le modèle a donc prédit moins de résiliations pour certains âges.

Pour avoir une idée sur l'impact de l'évolution de la composition des effectifs sur la consommation du portefeuille, nous traçons les distributions des âges obtenues.

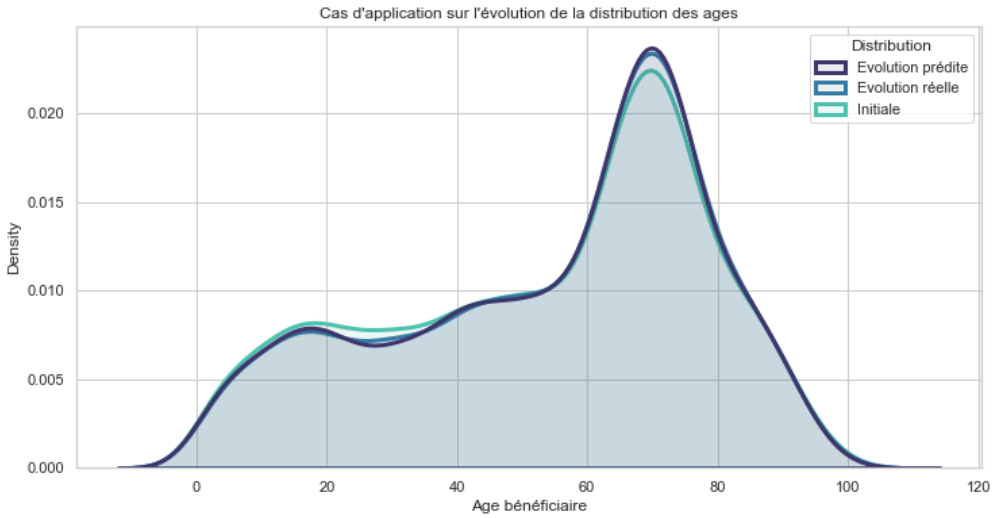


FIGURE 6.2 – Prédiction de l'évolution de la distribution des âges

La distribution des âges obtenue est très proche de la réalité. Le XGBoost a particulièrement su détecter les résiliation concernant les jeunes âgés de 20 à 40 ans ce qui a permis de prédire le creux observé sur cette tranche d'âge.

Cette évolution de distribution permet de mieux estimer la consommation du portefeuille vu que l'âge des bénéficiaires est le principal indicateur sur la consommation de ces derniers.

L'application des modèles de prédiction des coûts moyens et fréquences par poste nous permet d'obtenir les primes pures pour l'année postérieure à l'observation que nous traçons en fonction de l'âge des assurés comme suit :

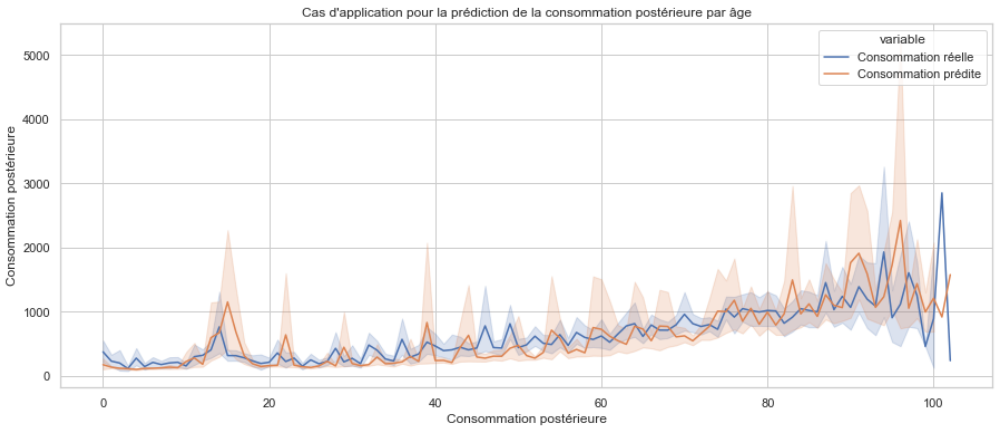


FIGURE 6.3 – Prédiction de l'évolution des primes pures par âge

Malgré que les modèles GLMs n'étaient pas excellents en terme de précision, vu l'irrégularité des distributions due aux différents forfaits et systèmes de remboursement définis dans les grilles de garanties de la mutuelle, nous arrivons à approcher les primes pures réelles par âge avec une prudence assez bonne. Le fait de modéliser la consommation par grand poste nous a permis de détecter le pic de consommation pour les jeunes de 17 à 19 ans dû aux actes en orthodontie comme vu précédemment.

Chapitre 7

Conclusion

Suite aux différentes évolutions réglementaires dans le secteurs de l'assurance santé individuelle telles que l'entrée du 100% santé et de la résiliation infra-annuelle, les mutuelles santé ont de plus en plus de mal à maintenir leurs équilibres au cours de ce développement. De ce fait, il peut être utile de mettre en place un système de prédiction de l'évolution de la consommation et des effectifs au sein des portefeuilles en assurance santé individuelle afin de maintenir leurs équilibres dans un tel contexte.

Cette étude nous a permis la mise en place d'un exemple de système permettant la prédiction de l'évolution sur 1 an des flux sortants d'effectifs d'un portefeuille en assurance santé individuelle, ainsi que leurs consommations à venir dans l'horizon d'un an. Disposant des données de janvier 2018 à mai 2022, nous nous sommes basés sur des observations mensuelles reliant les données antérieures d'un an aux données postérieures à prédire sur une durée d'un an. Avec les modèles mis en place, nous sommes arrivé à prédire plus de la moitié des résiliations à l'horizon d'un an avec une précision de plus de 90%. D'autre part, les modèles utilisés pour la prédiction de la durée de couverture restante des contrats à risque de résilier et de la consommation postérieure des assurés présentent des résultats de performance prometteurs.

Cependant ce modèle ne tient pas compte de certaines variables pouvant impacter considérablement les résiliations au sein du portefeuille telles que l'exposition des adhérents aux offres de la concurrence. De plus, le portefeuille étudié est concentré géographiquement sur deux départements voisins, il n'était donc pas possible d'exploiter certaines données en libre accès pouvant améliorer la modélisation. D'autre part, plusieurs points restent à améliorer dans cette modélisation, telle que la segmentation des données et le repérage des profils risqués mais aussi de pouvoir séparer les résiliations dues au décès d'un adhérent et celle issues d'une décision personnelle.

Enfin, cette étude permet la prédiction de l'évolution des flux sortants de contrats sur l'horizon d'un an. Cependant, la réelle évolution du portefeuille doit aussi tenir compte des flux entrants des contrats. Une modélisation de ces flux viendrait donc compléter le modèle.

Bibliographie

[1] AKAFFOU, D. Méthode alternative de tarification santé : Glm / xgboost. *Mémoire Institut des Actuaires* (2020).

[2] BUCCI, S. Étude et implémentation de techniques d'analyse de sensibilité dans les modèles de tarification non-vie. application à la tarification à l'adresse. *Mémoire Institut des Actuaires* (2021).

[3] Chen, T., and Guestrin, C. Xgboost : A scalable tree boosting system.

[4] conseil du financement et de la protection sociale, H. Évolution de la structure des recettes finançant la protection sociale.

[5] Drees. Rapport 2021 - sur la situation financière des organismes complémentaires assurant une couverture santé.

[6] et Su-In Lee, S. M. L. A unified approach to interpreting model predictions.

[7] le secrétaire général de la commission des comptes de la sécurité sociale. Les comptes de la sÉcuritÉ sociale.

[8] Maladie, L. Notre environnement : la sécurité sociale.

[9] MARIE-ROSE, J. Application sur r en tarification non-vie. *Support de cours M2 EURIA* (2021-2022).

[10] P.AILLIOT. Modèle linéaire généralisé. *Support de cours M1 EURIA* (2020-2021).

[11] VERMET, F. Apprentissage statistique : Une approche connexionniste, 2021. *Support de cours M1 EURIA* (2020-2021).