Mots clés :

Copule, Quotité, Risques compétitifs, Kaplan Meier, Nelson-Allen, XGBoost, LightGBM

RESUME

Traditionnellement, l'étude des risques en assurance-vie et en prévoyance se fait tête par tête. Toutefois, l'assurance prévoyance et en particulier l'assurance des emprunteurs implique très souvent la couverture de plusieurs têtes et de différents événements.

En effet, lors d'une souscription à une telle assurance, les co-emprunteurs peuvent non seulement choisir le prorata global du prêt à couvrir, mais également la quotité de couverture de chacun des deux assurés. Par conséquent, un décès sur une tête couverte à 100% au sein d'un couple engendrera le remboursement du capital restant dû et la clôture du contrat, mais celui-ci continuera pour une couverture inférieure.

La prise en compte de ces spécificités dans la tarification est souvent faite partiellement. Il est possible par exemple d'observer une réduction en pourcentage de la cotisation sur tous les contrats d'assurance contractés en couple. La complexité croissante des produits d'assurance et l'augmentation de la compétitivité sur le marché de la prévoyance, soulignent l'importance de la prise en compte de la structure de dépendance des données liées à plusieurs assurés, mais aussi aux couvertures pour déterminer une tarification plus robuste.

Pendant longtemps, la distribution normale multivariée a été le seul outil utilisé en statistique, en science financière et en science actuarielle pour rendre compte de la corrélation au sein des populations étudiés. Cependant, les techniques qui reposent sur des distributions normales multivariées ont été récemment remises en cause, et des alternatives plus réalistes ont été développées pour modéliser les corrélations entre les variables aléatoires de diverses distributions.

C'est dans ce contexte qu'une étude a été menée pour améliorer la modélisation technique d'un produit d'assurance emprunteur qui, à la base, néglige cette structure de dépendance. Ce mémoire s'attache à créer des modèles multivariés basés sur des outils mathématiques récents pour modéliser le risque d'entrée en incapacité temporaire de travail et le taux de décès instantané au sein d'un contrat emprunteur sur 2 têtes.

Après avoir présenté dans la première partie de ce travail, le marché de l'assurance emprunteur et son contexte juridique au travers de multiples lois permettant une forte compétitivité des assureurs, nous introduirons dans le second chapitre les concepts et les outils sous-tendant ces nouvelles approches. En particulier, les copules qui mesurent la structure de dépendance au sein des couples emprunteurs, le Machine Learning qui permet de prendre en compte la parcimonie de notre variable réponse et les risques compétitifs quantifiant de manière exacte les risques étudiés lorsque ceux-ci s'inhibent entre eux.

Dans la troisième partie, nous aborderons la mise en œuvre de ses outils et leur adaptation au portefeuille étudié. Nous nous attarderons notamment sur la mise en œuvre de modèles de survie en fonction du ratio de couverture, appelé également quotité, de chaque individu au sein du couple.

Enfin, nous mesurerons la sensibilité et la robustesse de nos différents modèles, afin de juger de leur qualité ainsi que de leurs limites.

Ce mémoire suggère la possibilité de prendre en compte la structure de dépendance au sein des co-emprunteurs. La mise en exergue de cette relation est cruciale dans l'adaptation des études en cas de changement de la loi de mortalité, comme le vieillissement du portefeuille, ou de son changement structural.

Key words:

Copula, Quota, Competitive Risks, Kaplan Meier, Nelson-Allen, XGBoost, LightGBM

ABSTRACT

Traditionally, actuarial development has been based on the assumption of independence of the random variables involved. The increasing complexity of insurance products and the need to cover events previously excluded from coverage, emphasize the importance of taking into account the dependence structure of the data.

For a long time, the multivariate normal distribution was the only tool used in statistics, financial science and actuarial science to account for the correlation within the specimens studied. However, techniques based on multivariate normal distributions have recently been challenged, and more realistic alternatives have been developed to model correlations between random variables of various distributions.

It is in this context that a study was launched to improve the coverage of a loan insurance product that basically neglects this dependency structure. This dissertation focuses on creating multivariate models based on recent mathematical tools to model the risk of disability and the instantaneous death rate in a joint life insurance policy.

After having presented in the first part of this work the new loan insurance market and its legal context through the multiple laws allowing a strong competitiveness of the insurers, we will introduce in the second chapter the concepts and the tools underlying these new approaches. In particular, copulas that measure the structure of dependence within borrower couples, machine learning that allows us to take into account the parsimony of our response variable and competitive risks that quantify the risks studied in an exact manner when they inhibit each other.

In the third part, we will discuss the implementation of these tools and their adaptation to the studied portfolio. In particular, we will focus on the implementation of survival models based on the coverage ratio of each individual within the couple.

Finally, we will measure the sensitivity and robustness of our different models, in order to judge their quality as well as their limitations.

This dissertation highlights the possibility of segmenting the risk of disability and death while taking into account the structure of dependence within the co-borrowers. The highlighting of this relationship is crucial in the adaptation of studies in case of changes in the mortality law (linked to climatic risks for example), the aging of the portfolio or its structural change.

SYNTHESE

I) INTRODUCTION

Depuis quelques années, les changements réglementaires de l'assurance des emprunteurs plongent le marché dans un environnement de plus en plus concurrentiel.

Le monde de l'assurance est en perpétuel mouvement. Avec l'arrivée du Big Data et les multiples changements réglementaires, les assureurs sont de plus en plus soumis au challenge et à la difficulté de trouver des solutions innovantes :

- Des méthodes toujours plus efficaces sans que cela soit trop coûteux.
- Des solutions personnalisées sans mettre en péril la mutualisation.

C'est dans ce contexte que BNP Cardif a mis en place une étude dans le but de trouver des outils adaptés afin d'améliorer sa tarification. Plus particulièrement, l'estimation de la sinistralité décès et d'entrée en incapacité de travail, à savoir le taux de hasard sur deux têtes.

Contrairement à ce qui est fait historiquement, la tarification ne sera plus calibrée en fonction de l'individu. Nous prendrons en compte, de manière plus innovante, la structure de dépendance au sein d'un dossier couple.

Ainsi, les méthodes utilisées consistent à trouver la loi jointe modélisant la sinistralité par dossier, c'est-à-dire, sur deux têtes.

Après avoir présenté les données, et utilisé des méthodes d'analyse de survie classique, nous proposerons tout d'abord une méthode qui consiste à agir sur la modification de la base donnée en entrée du modèle. Nous pourrons à partir du modèle déjà utilisé prendre en compte l'interdépendance des caractéristiques au sein d'un couple.

Par la suite, nous étudierons une seconde méthode plus statistique : les copules ; un outil qui permettrait de donner la loi jointe du taux de hasard au sein d'un couple.

Enfin, nous utiliserons les risques compétitifs afin de tenir compte de l'inhibition des risques au sein d'une paire d'assurées.

a. LES OUTILS DISPONIBLES : LES DONNEES

L'application de cette analyse concerne le périmètre de la Prévoyance. Il s'agit de suivre la sinistralité d'un jeune portefeuille en assurance emprunteur qui est proposé depuis 2012 à nos jours. Les observations de notre étude s'arrêtent au 31 décembre 2019 pour éviter que la tendance soit biaisée par la surmortalité causée par la Covid-19, ou encore l'excès d'arrêts-maladies qui en ont découlé.

Dans ce produit, la tarification se fait tête par tête. Néanmoins, une réduction est proposée lorsque les emprunts sont contractés à deux. Cela est justifiable en partie. En effet, lorsque nous couvrons deux têtes sur le même emprunt, nous faisons de la diversification donc nous prenons ainsi moins de risques.

Cependant, nous souhaiterions étudier l'adaptabilité de cette réduction.

Chaque année i, nous pouvons voir le risque porté par un couple comme une allocation d'actifs sur deux actifs risqués.

$$R_{contrepartie(i)} = (quot_1 \mathbb{P}(DC_1) + quot_2\ \mathbb{P}(DC_2))\ \times CRD_i$$

Avec la variable $R_{contrepartie}$ qui fait référence au risque défaut de remboursement de l'emprunt, et la variable $quot_i$, la part d'assurance affectée à chaque co-emprunteur.

b. MODELE D'ASSURANCE

En général, lorsque nous souhaitons étudier la dépendance au sein de deux risques, c'est la corrélation linéaire qui est prise en compte. Cependant, dans notre cas, cette corrélation est nulle. Contrairement aux idées reçues, une absence de corrélation linéaire entre deux variables X et Y ne suggère pas l'indépendance de X par rapport à Y. Il existe en effet d'autres types de liens comme les copules. Cette structure de dépendance peut ne pas être linéaire, c'est pourquoi ce type de lien n'est pas capté par le coefficient de corrélation de Pearson. Elle peut également ne pas être monotone, ou être symétrique ce qui ne permet pas non plus de la mesurer en fonction des coefficients de concordance et de discordance, i.e. le coefficient de corrélation de Kendal.

Une tarification par tête est interprétable et facile à mettre en place. Cependant, cette idée de segmentation par individu néglige fortement le lien qu'il pourrait y avoir au sein des variables explicatives conjointes. Ceci pourrait même être dangereux si la mortalité des uns entraîne inéluctablement celle des autres.

L'introduction des copules comme seconde couche du modèle pourrait présenter un bon compromis entre la complexité de la modélisation et la rentabilité de cette modification. Il s'agit en effet d'outils mathématiques plus développés, mais facile à implémenter. D'autre part, ces outils permettent de s'adapter à tous les types de lois dès lors que nous avons un échantillon exhaustif.

La méthode des risques compétitifs permet quant à elle de quantifier le risque réel porté par l'individu dès la souscription. En effet, elle permet une modélisation plus juste du risque porté par chacun des assurés en couple grâce à la prise en compte de la structure du contrat. En effet, sous certaines conditions définies par le choix des quotités, cette structure inhibe l'observation de sinistres subits par le conjoint survivant (entré en incapacité de travail, décès, …). Ce sont des évènements que nous aurions pu observer si le décès du conjoint assuré à 100% n'avait pas eu lieu par exemple, ou si la personne décédée était couverte à moins de 100%. La modélisation du risque dans ces conditions conduirait à sous-estimer le risque du conjoint survivant, du fait que son exposition ait été tronquée.

De plus, la disparition de ces individus du portefeuille empêche également de capter la vraie structure de dépendance qu'il peut y avoir au sein des conjoints. En effet, si deux décès ont lieu au sein d'un même couple à des dates proches mais différentes, la clôture du premier décès empêchera l'assureur d'accéder à cette information.

c. LE MODELE DE BASE

Dans ce mémoire nous partons sur la base du modèle dit individuel. Il s'agit ici de la mise à disposition du modèle light-GBM. Il est dit « individuel » car l'apprentissage est seulement effectué sur les variables explicatives de chaque assuré sans tenir compte de la relation qu'il peut avoir avec les autres individus du portefeuille.

Nous supputons ainsi qu'il y a indépendance au sein des individus du portefeuille.

Nous avons donc à notre disposition une base d'assurés qui ne respecte pas les hypothèses de modélisation.

d. METHODE D'AJOUT DES VARIABLES DU CONJOINT

Le modèle LightGBM tel qu'il est implémenté ne tient pas compte des interactions qu'il peut y avoir au sein des individus de la même base. Effectivement, la base de données qui constitue l'entrée du modèle présente un individu par ligne. Toutefois, si l'on modifie cette base en y ajoutant les informations essentielles, telles qu'on peut le voir sur la transformation qui suit, nous pouvons déjà observer une réelle amélioration de la prédiction.

Clef_couple	ID_Ind	Age_Ind	CSP_Ind	Taille_Ind	Variable_cible
1	1	45	CSP1	173	173
2	2	51	CSP2	165	165
3	3	28	CSP3	162	162
4	4	19	CSP4	178	178
5	5	21	CSP3	189	189
6	6	35	CSP1	155	155
7	7	54	CSP2	190	190
8	8	19	CSP1	176	176
9	9	33	CSP1	169	169
10	10	38	CSP3	156	156
11	11	26	CSP2	167	167
12	12	34	CSP1	186	186
...					
24	24	65	CSP1	187	187

periode	Clef_coupl	ID_Ind_1	ID_Ind_2	Age_Ind_1	Age_Ind_2	CSP_Ind_1	CSP_Ind_2	IMC_Ind1	IMC_Ind2	variable_cible_1	variable_cible_2
2012	1	1	13	45	43	CSP1	CSP1	18	19	0	0
2013	1	1	13	46	44	CSP1	CSP1	18	19	0	0
2014	1	1	13	47	46	CSP1	CSP1	18	19	1	0
2013	4	4	16	19	22	CSP4	CSP3	26	24	0	0
2014	4	4	16	20	23	CSP4	CSP3	26	24	0	0
2015	4	4	16	21	24	CSP4	CSP3	26	24	0	0
2016	4	4	16	22	25	CSP4	CSP3	26	24	0	1
2019	8	8	20	19	21	CSP1	CSP1	24	23	0	1
2015	9	9	21	33	31	CSP1	CSP4	25	29	0	0
2016	9	9	21	34	32	CSP1	CSP4	25	29	1	1
2012	11	11	23	26	20	CSP2	CSP2	25	21	0	0
2013	11	11	23	27	21	CSP2	CSP2	25	21	0	0

periode	Clef_couple	ID_Ind_2	ID_Ind_1	Age_Ind_2	Age_Ind_1	CSP_Ind_2	CSP_Ind_1	IMC_Ind2	IMC_Ind1	variable_cible_1	variable_cible_2
2012	1	13	1	43	45	CSP1	CSP1	19	18	0	0
2013	1	13	1	44	46	CSP1	CSP1	19	18	0	0
2014	1	13	1	46	47	CSP1	CSP1	19	18	0	1
2013	4	16	4	22	19	CSP3	CSP4	24	26	0	0
2014	4	16	4	23	20	CSP3	CSP4	24	26	0	0
2015	4	16	4	24	21	CSP3	CSP4	24	26	0	0
2016	4	16	4	25	22	CSP3	CSP4	24	26	1	0
2019	8	20	8	21	19	CSP1	CSP1	23	24	1	0
2015	9	21	9	31	33	CSP4	CSP1	29	25	0	0
2016	9	21	9	32	34	CSP4	CSP1	29	25	1	1
2012	11	23	11	20	26	CSP2	CSP2	21	25	0	0
2013	11	23	11	21	27	CSP2	CSP2	21	25	0	0

Figure 1 - Passage de la base de données individuelle à la base de données des couples

En effet, nous pouvons voir que le biais est amélioré de 6% tandis que la variance affiche une valeur inférieure de 13%.

Nous pouvons donc observer que le modèle LightGBM arrive à capter une structure de dépendance au sein des variables des conjoints.

Cette méthode améliore la fonction de coût ainsi que la variance du modèle, cependant, elle suggère une gestion plus lourde du portefeuille. En effet, le traitement pour les couples et les individus seuls doivent se faire séparément. Cela engendrera un coût de gestion plus important.

e. UNE APPROCHE PLUS STATISTIQUE

Au lieu d'ajouter des variables, nous réutilisons le modèle actuel sans engager de modification ni de traitements lourds sur les bases de données actuelles. Il s'agit simplement de modéliser le taux de hasard de chacun des individus constituant un couple, puis d'ajouter une couche supplémentaire à cette modélisation afin d'obtenir la fonction de répartition de la loi jointe.

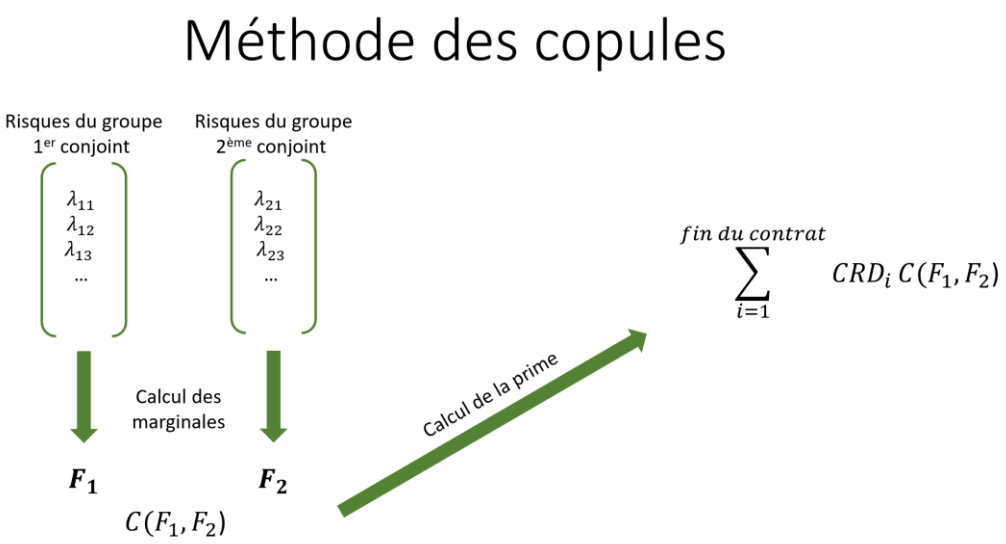


Figure 2 - Méthode des copules

Les résultats de l'adéquation montrent que la copule optimale est celle de Student. Dans ce cas, la loi jointe des taux de hasard s'écrit comme suit :

$$C(t_1, t_2) = \int_{-\infty}^{T_\nu^{-1}(u_1)} \int_{-\infty}^{T_\nu^{-1}(u_2)} \frac{1}{\pi\nu\sqrt{1-r^2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{{t_1}^2 - 2rt_1t_2y + {t_2}^2}{\nu(1-r^{2)}}\right)^{\left(-\frac{\nu}{2}+1\right)} dt_1 dt_2$$

L'équation se réécrit de la manière suivante :

$$C(t_1, t_2) = \int_{-\infty}^{T_\nu^{-1}(u_1)} \int_{-\infty}^{T_\nu^{-1}(u_2)} \frac{1}{\pi\nu\sqrt{1-r^2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{{t_1}^2 - 2rt_1t_2y + {t_2}^2}{\nu(1-r^{2)}}\right)^{\left(-\frac{\nu}{2}+1\right)} dt_1 dt_2$$

Nous pouvons d'autre part visualiser cette copule qui semble capter une dépendance au niveau des valeurs extrêmes, à savoir en (0,0), (1,1), (0,1) et (1,0). D'autre part, cette copule est symétrique par rapport à la diagonale, ce qui signifie que la dépendance entre les variables est la même dans les deux sens. Comme pour la copule Gaussienne, il s'agit d'une copule elliptique, mais elle est beaucoup plus flexible que celle-ci.

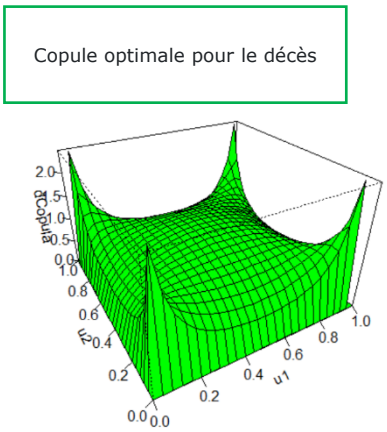


Figure 4 - Copule gaussienne pour le risque arrêt de travail

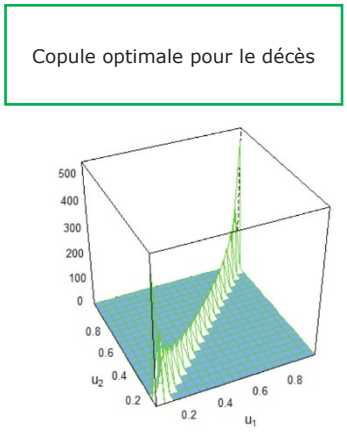


Figure 3 - Copule de Student pour le risque décès

Lorsque nous essayons d'obtenir ses paramètres après l'adéquation, nous n'arrivons pas à estimer le paramètre de forme υ car l'estimateur n'est pas assez précis. Cette étude donne exactement les mêmes résultats pour le décès et pour l'arrêt maladie. La copule qui en résulte alors ne permet pas d'avoir une bonne adéquation.

En somme, l'étude de la dépendance entre les taux de hasard au sein d'un couple affirme l'existence d'un lien entre le comportement des conjoints face à leur sinistralité. Cependant la parcimonie des données ne nous permet pas de conclure sur la nature exacte de ce lien.

f. UNE APPROCHE PAR LES RISQUES COMPETITIFS

Dans cette section, nous abordons les erreurs d'estimation du risque qui peuvent être commises si l'on ne tient pas compte de la relation entre les quotités couvertes au sein d'un couple.

En effet, lors de la souscription à notre assurance emprunteur, les individus ont le choix (sous réserve d'accord avec la banque) de prendre un taux de couverture qui leur convient. En effet, nous comprenons ainsi qu'une personne n'est pas obligée de couvrir entièrement son capital restant dû. C'est ce qui définit la notion de quotité assurée.

De ce fait, lorsqu'un couple de personnes contracte une assurance en commun pour couvrir leur prêt, ils peuvent choisir leurs quotités conjointes librement.

Nous pouvons par exemple avoir des assurés qui couvrent ensemble 100% du risque de l'emprunt, mais ils peuvent également le couvrir à 200%.

Dans ce dernier cas, le remboursement du capital restant dû se fait au premier décès. Nous observerons alors la clôture du contrat et la sortie de l'individu survivant de notre base d'observation. Ce qui nous empêchera d'estimer la probabilité du décès des deux têtes, ou l'entrée en incapacité du conjoint survivant. Les couples couverts à 200% représentent une forte proportion du portefeuille.

Cette répartition est homogène quels que soient les sexes au sein des couples. Ce qui peut fortement biaiser notre analyse si nous ne tenons pas compte de cette forme de « censure » à droite.

Lorsque la quotité totale assurée est entre 100% et 190% nous observons un risque compétitif partiel. Cela veut dire que l'observation du couple s'arrêtera seulement si le décès concerne la personne assurée à 100%. Dans le cas contraire, le taux de hasard de l'individu survivant ne sera pas impacté. Nous observerons cependant une diminution de son capital sous risque étant donné que la partie assurée par son conjoint a été remboursée.

Nous pouvons distinguer plusieurs cas de figure dans le tableau ci-dessous :

Les quotités		Etat du contrat en cas de sinistre décès d'un des deux individus			
Individu 1	Individu 2	Ratio de couverture	Ratio d'indemnisation maximal	Etat du contrat	Remarque
50%	50%	100%	100%	Continuité	
80%	20%	100%	100%	Continuité	
100%	100%	200%	100%	Clôture du contrat	
80%	50%	130%	100%	Continuité	Sur couverture
20%	30%	50%	50%	Continuité	
30%	50%	80%	80%	Continuité	
100%	50%	150%	100%	Dépend de l'individu sinistré	

Figure 5 - Etat du contrat en cas de sinistre décès d'un des individus

Imaginons deux assurés couverts respectivement à 40% et 70%, alors dans ce cas : lorsque l'assuré couvert à 70% décède, nous remboursons 70% du capital restant dû. Dans ce cas, le contrat d'assurance continuera de courir, cependant le capital sous risque sera de 30% et non de 40%, ce qui a été estimé

lors de la souscription puisque la tarification a été effectuée indépendamment sur la tête de chacun des deux assurés au sein du couple.

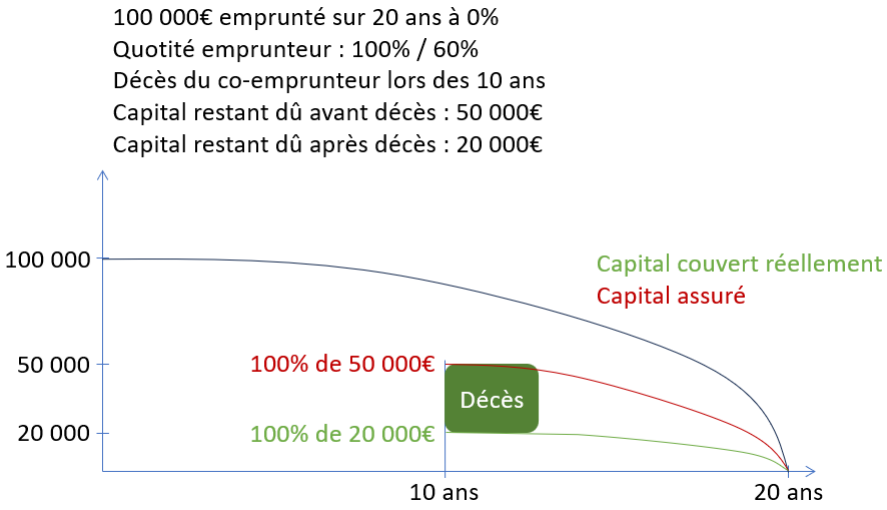


Figure 6 - Exemple de risques compétitifs

Après la mise en place de cette méthode, nous obtenons une tarification différente par tête. Nous remarquerons en particulier la diminution de la prime des personnes ayant un taux de couverture supérieur à 100%. D'autre part, nous confirmons l'ordre de réduction appliqué à la population couple, cependant celui-ci est distribué de manière plus homogène en fonction du risque encouru par chaque client.

II) CONCLUSION

La prise en compte de la structure de dépendance qui compose une partie de la sinistralité en assurance emprunteur sur deux têtes permet de montrer que la méthode traditionnelle peut, dans certains cas, être mal calibrée et conduire à biaiser l'estimation du risque réel. De manière globale, l'approximation du risque par tête ne permet pas d'intégrer l'influence que pourrait avoir un co-emprunteur sur la santé de l'autre. Par exemple, l'exposition d'un non-fumeur au tabac consommé par son conjoint peut augmenter son risque de cancer du poumon de 25%, rapporte le ministère de la Santé et de la solidarité.

Effets du tabagisme passif sur l'adul	Colonne1
Accidents cardiaques	Augmentation de 27%
Cancer du poumon	Augmentation de 25%
Cancer des sinus de la face	le tabagisme passif fait plus que doubler ce risque.
Accidents vasculaires cérébraux :	le tabagisme passif altère les parois des artères et double le risque d'accident vasculaire cérébral.

Figure 7 - Etude menée conjointement par le ministère chargé de la santé et la sécurité sociale

Cette étude démontre de manière scientifique, que le tabagisme passif comporte des risques réels pour la santé.

Nous pouvons ainsi penser que l'estimation de la survie d'une personne, individuellement, ne suffit pas pour estimer son risque réel. Il est, en effet, conseillé de prendre en compte les facteurs environnementaux qui pourraient exercer une influence non-négligeable sur sa santé.

D'autre part, la structure et la flexibilité du produit profère une grande liberté de choix pour les assurés, mais ceci suppose une gestion plus complexe quant au risque encouru par l'assureur. C'est ce problème que nous tenterons de résoudre par la méthode des risques compétitifs.

A travers cette étude, nous avons mis en place une proposition de prise en compte de ces interactions en utilisant les copules et en intégrant les variables explicatives du conjoint dans la modélisation.

EXECUTIVE SUMMARY

I) INTRODUCTION

In recent years, regulatory changes in creditor insurance have plunged the market into an increasingly competitive environment.

The world of insurance is constantly changing. With the advent of Big Data and multiple regulatory changes, insurers are increasingly challenged to find innovative solutions:

- Increasingly efficient methods without being too costly.
- Customised solutions without jeopardising pooling.

It is against this backdrop that BNP Cardif has set up a study with the aim of finding suitable tools to improve its pricing. In particular, the estimation of death and disability claims, i.e. the rate of chance on two lives.

Unlike in the past, pricing will no longer be calibrated according to the individual. Instead, in a more innovative way, we will consider the dependency structure within a couple's file.

The methods used consist of finding the joint law modelling the claims experience per case, i.e. over two heads.

After presenting the data and using traditional survival analysis methods, we will first propose a method that consists of modifying the base given as input to the model. Using the model already in use, we will be able to consider the interdependence of characteristics within a couple.

We will then look at a second, more statistical method: copulas, a tool that gives the joint distribution of the rate of chance within a couple.

Finally, we will use competitive risks to take account of risk inhibition within a pair of insureds.

a. AVAILABLE TOOLS : DATA

The application of this analysis concerns Personal Risk. The aim is to monitor the claims experience of a young loan insurance portfolio that has been offered since 2012. The observations in our study stop at 31 December 2019 to avoid the trend being skewed by the excess mortality caused by Covid-19, or the excess sick leave that resulted.

This product is priced on a per capita basis. However, a reduction is offered when the loans are taken out by two people. This is partly justifiable. Indeed, when we cover two lives on the same loan, we diversify and therefore take less risk.

However, we would like to study the adaptability of this reduction.

Each year i, we can see the risk borne by a couple as an asset allocation on two risky assets.

$$R_{contrepartie(i)} = (quot_1 \mathbb{P}(DC_1) + quot_2 \, \mathbb{P}(DC_2)) \times CRD_i$$

With the variable $R_{contrepartie}$ which refers to the risk of default on repayment of the loan, and the variable $quot_i$, the share of insurance allocated to each co-borrower.

b. INSURANCE MODEL

In general, when we want to study the dependency between two risks, we take the linear correlation into account. However, in our case this correlation is zero. Contrary to popular belief, an absence of linear correlation between two variables X and Y does not suggest that X is independent of Y. There are other types of link, such as copulas. This dependency structure may not be linear, which is why this type of link is not captured by the Pearson correlation coefficient. It may also be non-monotonic, or

symmetrical, which means that it cannot be measured by the concordance and discordance coefficients, i.e. Kendal's correlation coefficient.

A per capita rate is easy to interpret and implement, but this idea of segmentation by individual strongly neglects the link that could exist within the joint explanatory variables. This could even be dangerous if the mortality of some inevitably leads to that of others.

The introduction of copulas as the second layer of the model could represent a good compromise between the complexity of the modelling and the cost-effectiveness of this modification. These are mathematical tools that are more developed but easy to implement. Furthermore, this tool can be adapted to all types of laws as long as we have an exhaustive sample.

The competitive risk method quantifies the real risk borne by the individual at the time of underwriting. It provides a more accurate model of the risk borne by each policyholder in a couple by taking into account the structure of the contract. In fact, under certain conditions defined by the choice of quota-shares, this structure inhibits the observation of claims suffered by the surviving spouse: becoming unable to work, death, etc. These are events that we could have observed if the death of the spouse insured at 100% had not occurred, for example, or if the deceased was covered at less than 100%. Modelling the risk under these conditions would lead to an underestimation of the surviving spouse's risk, because his or her exposure has been truncated.

In addition, the disappearance of these individuals from the portfolio also prevents the true dependency structure that may exist within spouses from being captured. If two deaths occur within the same couple on close but different dates, the closure of the first death will prevent the insurer from accessing this information.

c. METHOD FOR ADDING SPOUSE VARIABLES

The LightGBM model, as implemented, does not take into account the interactions that can occur between individuals in the same database. Indeed, the database that forms the input to the model presents one individual per line. However, if we modify this database by adding the essential information, as can be seen in this transformation, we can already see a real improvement in the prediction.

Clef_couple	ID_Ind	Age_Ind	CSP_Ind	Taille_Ind	Variable_cible
1	1	45	CSP1	173	173
2	2	51	CSP2	165	165
3	3	28	CSP3	162	162
4	4	19	CSP4	178	178
5	5	21	CSP3	189	189
6	6	35	CSP1	155	155
7	7	54	CSP2	190	190
8	8	19	CSP1	176	176
9	9	33	CSP1	169	169
10	10	38	CSP3	156	156
11	11	26	CSP2	167	167
12	12	34	CSP1	186	186
...					
24	24	65	CSP1	187	187

periode	Clef_coupl	ID_Ind_1	ID_Ind_2	Age_Ind_1	Age_Ind_2	CSP_Ind_1	CSP_Ind_2	IMC_Ind1	IMC_Ind2	variable_cible_1	variable_cible_2
2012	1	1	13	45	43	CSP1	CSP1	18	19	0	0
2013	1	1	13	46	44	CSP1	CSP1	18	19	0	0
2014	1	1	13	47	46	CSP1	CSP1	18	19	1	0
2013	4	4	16	19	22	CSP4	CSP3	26	24	0	0
2014	4	4	16	20	23	CSP4	CSP3	26	24	0	0
2015	4	4	16	21	24	CSP4	CSP3	26	24	0	0
2016	4	4	16	22	25	CSP4	CSP3	26	24	0	1
2019	8	8	20	19	21	CSP1	CSP1	24	23	0	1
2015	9	9	21	33	31	CSP1	CSP4	25	29	0	0
2016	9	9	21	34	32	CSP1	CSP4	25	29	1	1
2012	11	11	23	26	20	CSP2	CSP2	25	21	0	0
2013	11	11	23	27	21	CSP2	CSP2	25	21	0	0
periode	Clef_couple	ID_Ind_2	ID_Ind_1	Age_Ind_2	Age_Ind_1	CSP_Ind_2	CSP_Ind_1	IMC_Ind2	IMC_Ind1	variable_cible_1	variable_cible_2
2012	1	13	1	43	45	CSP1	CSP1	19	18	0	0
2013	1	13	1	44	46	CSP1	CSP1	19	18	0	0
2014	1	13	1	46	47	CSP1	CSP1	19	18	0	1
2013	4	16	4	22	19	CSP3	CSP4	24	26	0	0
2014	4	16	4	23	20	CSP3	CSP4	24	26	0	0
2015	4	16	4	24	21	CSP3	CSP4	24	26	0	0
2016	4	16	4	25	22	CSP3	CSP4	24	26	1	0
2019	8	20	8	21	19	CSP1	CSP1	23	24	1	0
2015	9	21	9	31	33	CSP4	CSP1	29	25	0	0
2016	9	21	9	32	34	CSP4	CSP1	29	25	1	1
2012	11	23	11	20	26	CSP2	CSP2	21	25	0	0
2013	11	23	11	21	27	CSP2	CSP2	21	25	0	0

Figure 8 – Switch from individual to couple database

We can see that the bias is improved by 6%, while the variance is 13% lower.

We can therefore observe that the lightGBM model manages to capture a dependency structure within the spouses' variables.

This method improves the cost function and the variance of the model, but it does suggest a heavier management of the portfolio. Indeed, the treatment for couples and single individuals must be done separately. This will result in a higher management cost.

d. A MORE STATISTICAL APPROACH

Instead of adding variables, in this section we can reuse the current model without making any changes or carrying out any heavy processing on the current databases. All we need to do is model the hazard rate of each of the individuals making up a couple, then add an additional layer to this modelling to obtain the distribution function of the joint distribution.

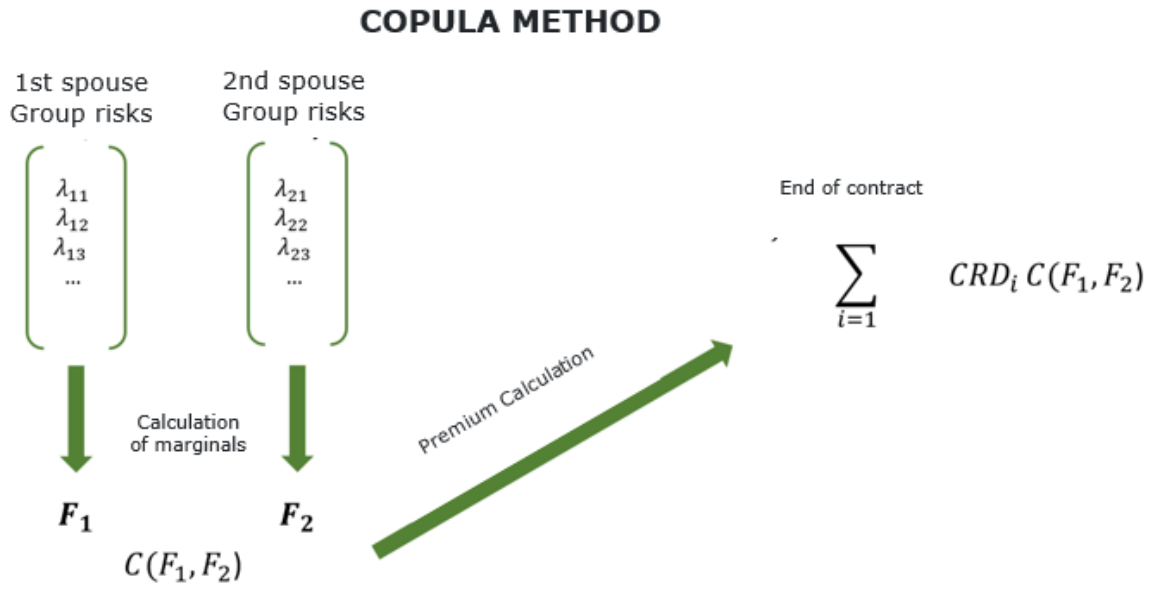


Figure 9 – Copula method

The matching results show that the optimal copula is the Student's copula. In this case, the joint distribution of hazard rates is written as follows:

$$C(t_1, t_2) = \int_{-\infty}^{T_\nu^{-1}(u_1)} \int_{-\infty}^{T_\nu^{-1}(u_2)} \frac{1}{\pi\nu\sqrt{1-r^2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{{t_1}^2 - 2rt_1t_2y + {t_2}^2}{\nu(1-r^2)}\right)^{\left(-\frac{\nu}{2}+1\right)} dt_1 dt_2$$

The equation can be rewritten as follows:

$$C(t_1, t_2) = \int_{-\infty}^{T_\nu^{-1}(u_1)} \int_{-\infty}^{T_\nu^{-1}(u_2)} \frac{1}{\pi\nu\sqrt{1-r^2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{{t_1}^2 - 2rt_1t_2y + {t_2}^2}{\nu(1-r^2)}\right)^{\left(-\frac{\nu}{2}+1\right)} dt_1 dt_2$$

We can also visualise this copula, which seems to capture a dependency at the extreme values, i.e. at (0,0), (1,1), (0,1) and (1,0). Furthermore, this copula is symmetrical with respect to the diagonal, which means that the dependence between the variables is the same in both directions. Like the Gaussian copula, this is an elliptical copula, but it is much more flexible than the Gaussian copula.

Optimal Copula for temporary disability risk

Optimal Copula for death risk

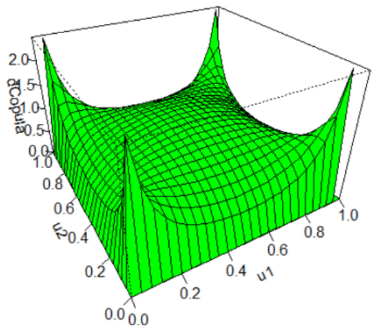


Figure 11 – Gaussian copula

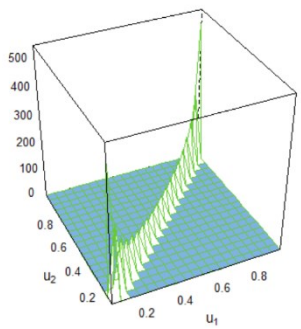


Figure 10 – Student copula

When we try to obtain its parameters after matching, we are unable to estimate the shape parameter u because the data are very unbalanced. This study gives exactly the same results for death as for sick leave.

In short, the study of the dependence between hazard rates within a couple confirms the existence of a link between the behaviour of spouses with regard to their claims experience. However, the parsimony of the data does not allow us to conclude on the exact nature of this link.

e. A COMPETITIVE RISK APPROACH

In this section, we discuss the errors in estimating risk that can be made if we do not take account of the relationship between cover ratios within a couple.

When taking out our loan insurance, individuals have the choice (subject to agreement with the bank) of taking out a rate of cover that suits them. This means that people are not obliged to cover their entire outstanding capital. This is what defines the notion of insured percentage.

So when a couple take out joint insurance to cover their loan, they are free to choose their joint percentage cover.

We can, for example, have policyholders who together cover 100% of the risk of the loan, but they can also cover it at 200%.

In the latter case, the outstanding capital is repaid on the first death, so we will observe the closure of the contract and the removal of the surviving individual from our observation base. This will prevent us from estimating the probability of both heads dying, or of the surviving spouse becoming incapacitated. Couples covered at 200% represent a large proportion of the portfolio.

This distribution is homogeneous, irrespective of the sex of the couple. This can strongly bias our analysis if we do not consider this form of 'censorship' on the right.

When the total insured percentage is between 100% and 190%, we observe a partial competitive risk. This means that the observation of the couple will stop only if the death concerns the person insured at 100%. Otherwise, the hazard rate of the surviving individual will not be affected. However, we will observe a reduction in the capital at risk, given that the part insured by the spouse has been reimbursed.

There are several possible scenarios in the table below:

Quotas		**Status of the contract in the event of the death of one of the two individuals**			
Individual 1	**Individual 2**	**Coverage ratio**	**Maximum compensation ratio**	**Contract status**	**Note**
50%	50%	**100%**	100%	Continuity	
80%	20%	**100%**	100%	Continuity	
100%	100%	**200%**	100%	Close of contract	
80%	50%	**130%**	100%	Continuity	**Over-coverage**
20%	30%	**50%**	50%	Continuity	
30%	50%	**80%**	80%	Continuity	
100%	50%	**150%**	100%	Depends on the individual	

Figure 12 – Contract status in case of death of one of the individuals

Let's imagine two insureds covered at 40% and 70% respectively. In this case, when the insured covered at 70% dies, we reimburse 70% of the outstanding capital. In this case, the insurance contract will

continue to run, but the capital at risk will be 30% and not 40%, which was estimated when the policy was taken out, since the rates were calculated independently for each of the two insured persons in the couple.

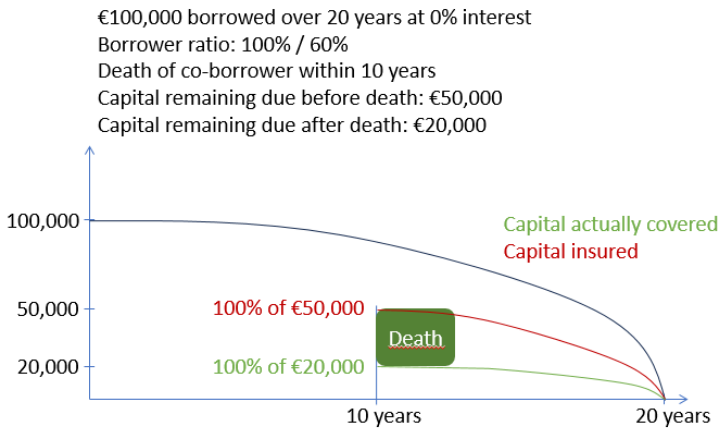


Figure 13 – Competitive risks example

After the implementation of this method, we obtain a different pricing per head. We will note in particular the reduction in the premium for people with a coverage rate above 100%. Moreover, we confirm the order of reduction applied to the couple population; however, it is distributed more evenly according to the risk incurred by each customer.

II) CONCLUSION

Taking into account the structure of dependency, which accounts for part of the claims experience in double life loan insurance, shows that the traditional method can, in certain cases, be poorly calibrated and lead to a biased estimate of the real risk. Overall, the approximation of risk per head does not allow for the influence that one co-borrower could have on the health of the other. For example, a non-smoker's exposure to the tobacco consumed by his or her spouse can increase his or her risk of lung cancer by 25%, reports the Ministry of Health and Solidarity.

Effect of passive smoking on adults	
Cardiac accidents	27% increase
lung cancer	25% increase
Facial sinus cancer	passive smoking more than doubles this risk
cerebrovascular accidents	passive smoking damages artery walls and doubles the risk of cerebrovascular accident

Figure 14 - Study conducted jointly by the Ministry of Health and the Social Security Department

This study provides scientific evidence that passive smoking entails real health risks.

We can therefore assume that estimating the survival of an individual is not enough to estimate the real risk. It is advisable to take into account environmental factors that could have a significant influence on health.

On the other hand, the structure and flexibility of the product gives policyholders a great deal of freedom of choice, but this implies more complex management of the risk incurred by the insurer. It is this problem that we will attempt to address using the competitive risk method.

Through this study, we have put in place a proposal to take these interactions into account by using copulas and integrating the spouse's explanatory variables into the modelling.

TABLE DES MATIERES

1 REMERCIEMENTS

Je souhaite tout d'abord remercier ma responsable de stage Madame Delphine De Fournoux La Chaze, pour sa disponibilité, son écoute, ses précieux conseils avisés et son investissement afin que mon stage se déroule dans les meilleures conditions.

Je remercie également l'ensemble de l'équipe du département "Souscription" BNP Cardif, dirigé par David Antonetti, pour leur accueil, leur soutien et leur aide.

Je tiens plus particulièrement à remercier Monsieur Thomas Fuks, Monsieur Thomas Bluteau, Monsieur Aymane Berradi.

Ma reconnaissance va également à Monsieur Boris Noumedem ainsi que Monsieur Quang Do Xuan pour leur aide et leurs conseils.

Je n'oublierai jamais mes professeurs de l'ENSAE, qui m'ont beaucoup apporté durant ma formation.

Ma gratitude va également à l'égard de Monsieur Nicolas Barradel, mon tuteur académique, pour son aide et sa disponibilité tout au long de mes travaux.

La réalisation de ce témoignage a été possible grâce à ma famille et amis qui m'ont soutenue tout au long de mon parcours.

2 INTRODUCTION

L'assurance emprunteur est une protection que les banques proposent invariablement lorsqu'elles prêtent de l'argent. Bien que ce ne soit pas obligatoire, c'est en fait une exigence établie par l'organisme prêteur afin d'accepter l'accord d'un crédit. Ce produit présente un avantage réciproque, il rassure le prêteur et protège les héritiers de l'assuré. En effet, en cas de défaut de remboursement, aucune dette ne leur incombera. Cette assurance peut être collective ou individuelle et peut offrir plusieurs types de protection prévoyance : principalement le décès, mais aussi l'incapacité de travail, l'invalidité permanente totale ou partielle, le chômage…

Parmi ces garanties, seule la couverture décès est obligatoire, elle permet le remboursement de la totalité du capital restant dû au prêteur en cas de sinistre.

Cependant, comme dans toutes les garanties de prévoyance, les risques sont estimés tête par tête. Cette méthode néglige donc la structure de lien qu'il peut y avoir au sein des variables explicatives d'un couple. Or il est important pour l'assurance d'estimer cette dépendance afin de connaitre comment ces risques s'influencent entre eux. Les risques ont-ils tendance à varier dans le même sens ? ou s'opposent-ils l'un à l'autre ?

Dans la majorité des contrats d'assurance emprunteur à deux têtes, les co-emprunteurs sont en couple.

Or un couple n'est pas à proprement dit l'association de deux personnes qui contractent un même contrat. Il s'agit de personnes qui peuvent avoir de nombreux points communs.

En effet, les couples composés de deux personnes appartenant à la même catégorie socioprofessionnelle représentent 30% de l'ensemble des couples en 1999 (Vanderschelden, 2006). Cette proportion est près de deux fois supérieure à celle que l'on observerait si les couples s'étaient formés au hasard.

Ce chiffre est confirmé par notre base dans laquelle 66% des couples ont la même catégorie.

Enfin, d'après l'Insee et comme nous pouvons le voir sur le tableau ci-dessous, il a été montré de manière statistique que vivre en couple diminue le risque de mortalité.

Situation matrimoniale légale, à la date du recensement		
Marié	1	1
Situation inconnue (non-réponse)	1,59***	0,99
Célibataire	1,77***	1,92***
Veuf	2,29***	1,37***
Divorcé	1,93***	1,74***
1. Les résultats sont issus d'un modèle de Cox – modèle de durée à risques proportionnels (cf. encadré 3).		

Figure 15 - Situation matrimoniale légale - Etude INSEE

Ce mémoire constitue une analyse du risque décès et incapacité de travail au sein d'un portefeuille d'assurance des emprunteurs (ADE), agrégé par dossier. Cela veut dire que la quantification de cette sinistralité tiendra compte en sus des relations qu'il peut y avoir dans l'explication de la sinistralité d'un individu en tenant compte des caractéristiques de son conjoint.

Cette analyse pourra servir à améliorer la tarification en trouvant une autre manière d'adapter la modélisation. Les assurances sur deux têtes présentant un meilleur profil de risque pour l'assureur du fait de la diversification, nous avons donc choisi d'étudier cette possibilité afin de savoir si nous pouvions proposer des tarifs plus adéquats pour les couples.

Dans un premier temps, nous présenterons le concept de l'assurance des emprunteurs ainsi que les méthodes de tarification habituellement utilisées. La suite traitera des études, analyses et propositions de modélisation que nous avons effectuées afin de mettre en exergue la structure de dépendance et sa prise en compte. Cette modélisation se base principalement sur l'analyse de durée, les copules et LightGBM, qui est un algorithme de boosting. Nous terminerons enfin, par des tests de sensibilité sur les différents paramètres des modèles.

3 CADRE DE L'ETUDE

Cette étude a été mise en place afin de mesurer l'impact de la prise en compte des variables du conjoint dans l'estimation du risque encouru par l'assuré modélisé « toutes choses égales par ailleurs ». C'est ce qui a motivé le choix *de LightGBM comme l'apprentissage l'algorithme d'apprentissage*.

3.1 Présentation de l'assurance emprunteur

3.1.1 Définition

L'assurance emprunteur est une assurance qui permet la subrogation de l'assureur à l'assuré prenant en charge tout ou partie des échéances de remboursement ou du capital restant dû d'un crédit.

Cette couverture est appliquée en cas de survenance de certains événements, généralement :

- Le décès
- La perte totale et irréversible d'autonomie (PTIA)
- L'invalidité permanente
- L'incapacité temporaire de travail
- La perte d'emploi

Elle s'apparente ainsi à un contrat de prévoyance qui couvre le crédit de l'assuré, en cas de survenance d'un sinistre prévu dans le contrat (décès ou invalidité par exemple).

Dans ce cas, l'assureur se substituera à l'assuré pour le remboursement du prêt.

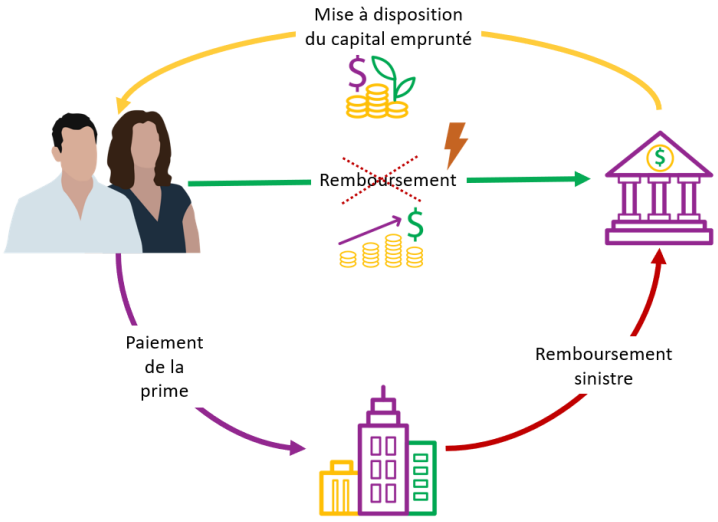


Figure 16 - Mécanisme type d'un contrat d'assurance emprunteur en cas de sinistre

3.1.2 Intérêt

L'assurance emprunteur permet à son contractant de se protéger contre les aléas de la vie. Malgré une situation critique qui empêcherait l'assuré d'honorer ses mensualités, il ne perdra pas son bien. D'autre part, en cas de décès, la dette ne sera pas transmise aux héritiers et le bien entrera directement dans la succession. Enfin, cette couverture permet à l'organisme créditeur de diminuer grandement son risque de contrepartie.

Cette couverture permet l'accès à la propriété aux individus avec peu de ressources, tels que des jeunes couples. Elle représente donc une réelle sécurité financière, qui est parfois indispensable à la bonne poursuite du projet du contractant.

3.1.3 Que dit la législation

L'assurance emprunteur n'est pas une assurance obligatoire : aucune disposition légale n'impose à un emprunteur d'être assuré. Toutefois, les établissements bancaires conditionnent souvent l'octroi d'un prêt immobilier à la souscription d'une assurance emprunteur. En pratique, en France, une personne sollicitant un crédit à la consommation ne sera que rarement contrainte de s'assurer. A l'inverse, celle sollicitant un crédit immobilier se verra presque systématiquement imposer l'obligation de s'assurer contre un ou plusieurs des risques cités ci-dessus.

3.2 Description des différentes garanties d'un contrat d'assurance emprunteur

3.2.1 La garantie Décès

La garantie décès est obligatoire dans tous les contrats d'assurance emprunteur.

En cas de sinistre, l'assureur versera donc le capital restant dû à la date du décès, à l'organisme prêteur. Cette indemnisation se fait sous réserve des exclusions de garantie prévues par le contrat (exclusion du décès durant la première année de souscription par suicide par exemple, sauf si le prêt concerne la résidence principale pour un montant maximal de 120 000 € fixé par décret en 2004).

3.2.2 La garantie incapacité de travail

3.2.2.1 Généralités dans le cadre de l'assurance maladie

Lorsque l'assuré salarié se trouve en arrêt de travail, il bénéficie des indemnités journalières (IJ) octroyées par la sécurité sociale. Cette indemnisation est effectuée à hauteur de 50% du salaire journalier de base (dans la limite du plafond de la sécurité sociale (PASS) soit 189€ pour 2022).

Habituellement, le versement des IJ est soumis à un délai de franchise de 3 jours, tandis que la durée du versement est limitée à trois ans. Au-delà de ces trois ans, l'assuré passe obligatoirement en état d'invalidité, si ce dernier est toujours en incapacité de reprendre son poste. Il est stipulé dans certaines conventions collectives des conditions permettant aux salariés de bénéficier d'indemnités complémentaires aux IJ, en cas d'incapacité de travail. Cela permet donc de maintenir une partie, voire la totalité du salaire pendant l'arrêt de travail du salarié, lui assurant ainsi un confort supérieur. Cette allocation peut être versée par l'employeur en fonction de :

- L'ancienneté du salarié
- De la nature de l'accident (à partir du premier jour en cas d'accident de travail par exemple)
- Du délai de carence (s'il existe)

Cette indemnisation prendra cependant en compte le dédommagement de la sécurité sociale de sorte que l'indemnisation totale ne dépasse jamais le niveau de salaire si l'assuré continuait son activité professionnelle sans interruption. En effet, il est illégal de s'enrichir avec une assurance.

3.2.2.2 La couverture Incapacité en assurance emprunteur

Est considéré en état d'incapacité totale de travail (ITT) tout assuré qui :

- A la suite d'un sinistre ou d'une maladie se trouve dans l'obligation d'interrompre son activité professionnelle de manière totale et temporaire.
- L'assuré n'exerçant plus d'activité professionnelle rémunérée mais qui est dans l'incapacité de reprendre l'exercice faute de sinistre à la suite duquel il se voit médicalement obligé d'observer un repos.

Une fois que cette reconnaissance est actée par l'assurance, l'assuré peut bénéficier du remboursement de 100% de ses mensualités d'emprunt au prorata des quotités choisies par celui-ci dans le contrat.

3.2.3 La garantie Perte Totale et Irréversible d'Autonomie

La garantie Perte Totale et Irréversible d'Autonomie (PTIA), correspond à une situation d'invalidité Absolue et Définitive. Elle doit satisfaire deux conditions :

1. L'assuré doit se trouver dans l'impossibilité totale et définitive de se livrer à une quelconque activité rémunérée pouvant lui procurer gain ou profit y compris une activité de surveillance ou de direction.
2. Il doit être dans l'obligation absolue d'avoir recours à l'assistance totale et constante d'une tierce personne pour effectuer trois ou quatre des actes ordinaires de la vie courante, c'est-à-dire :
   - Faire sa toilette
   - S'habiller
   - Se nourrir
   - Se déplacer

3.2.4 La garantie Invalidité Permanente

L'invalidité prévue par les contrats d'assurance emprunteur correspond à une inaptitude au travail permanente, qui peut être totale ou partielle. On parle :

- D'Invalidité Permanente Totale (IPT) si, du fait d'un accident ou d'une maladie et après consolidation de son état de santé, l'assuré est reconnu soit totalement inapte à l'exercice de toute activité pouvant lui procurer gains et profits, soit totalement inapte à l'exercice de l'activité qu'il exerçait au jour du sinistre. Le plus souvent la mise en jeu de la garantie IPT n'est possible que si l'assuré présente un taux d'incapacité au moins égal à 66%.
- D'Invalidité Permanente Partielle (IPP) si, du fait d'un accident ou d'une maladie et après consolidation de son état de santé, l'assuré est reconnu soit partiellement inapte à l'exercice de toute activité pouvant lui procurer gains et profits, soit partiellement inapte à l'exercice de l'activité qu'il exerçait au jour du sinistre. Le plus souvent la mise en jeu de la garantie IPP n'est possible que si l'assuré présente un taux d'incapacité au moins égal à 33% et inférieur à 66%. Cette garantie ne peut être souscrite qu'en complément d'une garantie IPT. Cette garantie n'a pas fait l'objet d'étude dans ce mémoire.

3.3 Contexte marché et juridique

3.3.1 Types de contrats

On distingue à la souscription deux types de contrats :

- Le contrat collectif : c'est un contrat proposé par l'établissement prêteur. Il s'agit d'une adhésion groupée.
- Le contrat en délégation d'assurance, qui est une police d'assurance souscrite par l'emprunteur, appelé couramment « contrat individuel » (bien qu'il s'agisse juridiquement de contrats collectifs à adhésion facultative et non de contrats individuels, le contrat étant généralement souscrit par une association et non par l'assuré. Pour le produit Cardif, il s'agit de l'UFEP)

Choisir l'adhésion collective est en général plus pratique :

- La souscription est prise en charge par l'organisme prêteur
- La cotisation est unique pour l'emprunt et l'assurance

Cependant, ce type de contrat est standard. Un contrat externe est quant à lui fait sur mesure, il est bien plus personnalisé et pourrait être également moins cher car la tarification est plus segmentée. Pour l'assureur, la gestion est un peu plus lourde car il est par exemple, tenu d'informer le prêteur du non-paiement par l'emprunteur de sa prime d'assurance ou de toute modification substantielle du contrat d'assurance. C'est ce second type de contrat qui sera analysé dans ce mémoire.

3.3.1.1 La Directive européenne

La Directive européenne 2014/17/UE datant du 4 février 2014 sur les contrats de crédit aux consommateurs relatifs aux biens immobiliers à usage résidentiel ouvre le marché de l'assurance des

emprunteurs à la concurrence. L'emprunteur a donc le droit de choisir son assurance de prêt à condition que les garanties proposées par l'assurance de son choix soient équivalentes ou supérieures à celles proposées par la banque.

Cette loi ordonne « s'il est justifié que les prêteurs puissent demander au consommateur de contracter une police d'assurance appropriée pour garantir le remboursement du crédit ou assurer la valeur de la garantie, le consommateur devrait pouvoir choisir son propre assureur, pour autant que sa police d'assurance présente un niveau de garantie équivalent à la police d'assurance proposée ou offerte par le prêteur ». L'emprunteur n'est donc jamais contraint d'adhérer au contrat groupe, et la banque ne peut refuser un contrat externe si la couverture est équivalente à celle qu'elle avait elle-même proposée. Elle ne pourra pas non plus augmenter le taux de l'emprunt en conséquence, ni facturer de frais particuliers pour l'analyse des polices alternatives présentées par l'emprunteur.

3.3.1.2 La législation Française

En vertu des directives européennes, un emprunteur en France peut choisir de s'assurer auprès de la banque ou d'un organisme assureur. Ce choix d'assurance peut être fait indépendamment, à condition que la garantie qu'il apporte soit égale ou supérieure à celle apportée par la banque au travers d'un contrat groupe. Par ailleurs, l'amendement Bourquin avait introduit en 2018 la possibilité pour l'assuré de résilier son contrat d'assurance à sa date d'anniversaire. Il pouvait ainsi choisir un autre contrat qui lui est plus favorable, sous réserve de garanties équivalentes. Cette possibilité de renégociation permettait aux emprunteurs d'ajuster l'assurance à leurs besoins. En effet, ce système donnait la possibilité d'abandonner certaines couvertures qui n'étaient plus d'actualité par exemple. Outre cette possibilité de revue des couvertures annuellement, l'emprunteur avait la possibilité de résilier son contrat douze mois suivant la signature du contrat.

Ces possibilités de résiliation ont été étendues avec la loi Lemoine qui est entrée en application au 1er juin 2022. Cette loi permet désormais la résiliation du contrat d'assurance emprunteur à tout moment, et plus seulement pendant les douze premiers mois du prêt (loi Hamon) ou à la date d'anniversaire du prêt une fois le délai de douze mois écoulé (loi Bourquin). Elle supprime aussi les formalités médicales (y compris les questionnaires de santé) pour les emprunteurs souscrivant un prêt immobilier, lorsque le capital assuré sur leur tête n'excède pas 200 000 € et lorsque l'âge au terme du prêt est inférieur à 60 ans. Ces dispositions s'appliquent aux affaires nouvelles à compter du 1er juin 2022 et aux affaires en stock au 1er septembre 2022.

3.3.1.3 L'obligation d'information de l'assureur envers l'assuré

Lors de la souscription, l'organisme préteur doit fournir à son assuré les conditions générales et particulières sur les coûts ainsi que les garanties proposées par son assurance groupe. D'autre part, l'organisme préteur doit informer son client du caractère facultatif d'assurance concernant les crédits à la consommation et, le cas échéant, sur le droit pour l'emprunteur de choisir un autre organisme pour sa souscription d'une couverture équivalente.

3.3.2 Appréciation de l'état de santé à la souscription

3.3.2.1 Le formulaire médical

Avant d'octroyer un prêt, l'organisme préteur peut demander à son client d'effectuer un questionnaire de santé. Le remplissage d'un tel formulaire peut aider l'assureur à mieux estimer le risque de son assuré.

Cette formule consiste à apporter une réponse brève, voire de type booléen à certaines questions pertinentes. Le souscripteur devra prendre garde à y répondre de manière exacte. Toute fausse déclaration pourra aboutir à la nullité du contrat et à la perte des primes déjà versées.

Ces questions peuvent porter sur :

- Son âge,
- Sa taille
- Son poids
- Ses antécédents familiaux (diabète, maladie de cœur, aliénation mentale, ...)

- Ses affections diverses (rhumatismes, pression artérielle, dépressions nerveuses, tumeurs, handicap, …)
- Ses habitudes de vie (fumeur, consommation d'alcool, …)

3.3.2.2 Visite médicale et convention (AERAS)

Après réception de ce formulaire, le futur assuré pourra se soumettre à une visite médicale. C'est lors de cette visite que son état de santé sera définitivement évalué. Ainsi, si l'état de santé de l'individu correspond aux critères de sélection par l'organisme assureur, celui-ci validera son dossier et le contrat d'assurance emprunteur pourra être signé dans les jours qui suivent.

En revanche, si l'emprunteur a déclaré des antécédents médicaux, une visite médicale lui sera demandée.

D'autre part, si certains seuils font franchis par les caractéristiques de l'emprunt ou de l'emprunteur, une visite médicale sera automatiquement demandée au futur assuré.

Exemples :

- Le montant de l'emprunt dépasse de 300.000 euros
- L'âge de l'assuré dépasse 65 ans …

La forme de la visite médicale sera en fonction des éléments du dossier, à savoir :

- Un simple bilan sanguin
- Une recherche plus spécifiques associés aux pathologies déclarées par le futur adhérent

L'assureur va ainsi se baser sur les résultats de cette visite afin de prendre une décision sur le type de prime à appliquer au contrat. Il aura donc à jauger entre :

- Appliquer une prime habituelle
- Ajouter une surprime
- Exclure le futur assuré de certaines garanties
- Refuser d'octroyer la couverture si le risque dépasse le niveau d'appétence de l'organisme

L'assureur s'octroie ainsi le droit de détenir des informations sensibles et confidentielles concernant ses assurés mais doit en contrepartie garantir la protection de ses informations pour que ceci ne puisse pas porter atteinte à la vie de ses assurés.

Il doit en outre supporter les frais médicaux demandés à ces derniers.

Enfin, les personnes ayant été exclues des contrats d'assurance sous prétexte de présenter un risque aggravé peuvent faire valoir leur droit à l'emprunt grâce à la convention AERAS (s'Assurer et Emprunter avec un Risque Aggravé de Santé).

Cette convention, signée par les banquiers, les pouvoirs publics, les associations de malades et de consommateurs et les assureurs, permet aux personnes exclues des contrats conventionnels d'améliorer leurs chances de pouvoir contracter un emprunt et de le couvrir par une assurance. Cela permet donc d'étendre l'assurance emprunteur aux personnes présentant certaines pathologies de bénéficier du droit à l'oubli, sous certaines conditions.

3.3.3 Les risques non couverts

Les exclusions de garanties sont au nombre de trois :

3.3.3.1 Les exclusions légales

Les exclusions légales font référence aux risques

- Techniquement inassurables

Sont considérés inassurables les risques dont l'assuré pourrait être victime et dont on ne peut pas estimer la fréquence ou la sévérité par une des lois statistiques stables. On pourrait penser à une charge de sinistralité qui suivrait une loi de Pareto dont le maximum n'a pas de limite.

Parmi ces phénomènes, peuvent faire partie les risques de :

- Acte de terrorisme
- Guerre civile ou étrangère
- Explosion nucléaire ou sabotage
- Émeute
- Rixes
- Moralement inassurables

Peut être considéré comme immorale toute assurance destinée à couvrir un crime ou un délit commis par l'assuré. En effet, ceci procurerait une sorte de sentiment d'impunité inhibant ainsi les personnes à répondre de leurs actes. Il sera donc plus aisé d'enfreindre la loi en toute impunité.

- Fraude
- Trouble à l'ordre public
- Le suicide de l'assuré au cours de la première année du contrat sous certaines conditions. Cette couverture est obligatoire pour la suite.
- Résultat de l'état d'ivresse de l'assuré
- Tentative d'escroquerie
- Résultat de consommation de l'usage de stupéfiants, d'hallucinogène de doses ou de médicaments non prescrits par un professionnel de santé…

3.3.3.2 Les exclusions particulières

Les exclusions particulières sont des exclusions liées à l'exécution de chaque contrat. Elles dépendent entièrement du niveau d'appétence de l'assurance par rapport aux risques qu'il souscrit.

Un assureur peut par exemple accepter ou non la souscription d'un individu pratiquant des sports extrêmes, ou s'exposant à un sur-risque de mortalité à cause de sa profession (exemple : démineur, reporter de guerre...)

Il peut également proposer le rachat de cette exclusion moyennant une surprime.

3.3.3.3 Fausse déclaration

Constitue une fausse déclaration toute faute intentionnelle ou dolosive de la part de l'assuré.

De ce fait, il est déclaré la nullité du contrat à partir du moment où l'assuré a volontairement retenu des informations susceptibles de changer la tarification du contrat ou la considération de sa couverture.

A cela peut s'ajouter, la réclamation de dommages et intérêt de la part de l'assureur.

Un exemple des plus importants serait l'oubli intentionnel de déclarer une maladie lors du remplissage du questionnaire de santé. Cela ne concerne évidemment pas les personnes bénéficiant de la loi Lemoine puisqu'elle supprime les formalités médicales sous certaines conditions.

3.4 Les engagements

3.4.1 Tableau d'amortissement

Un emprunt bancaire se traduit par des flux financiers entre l'organisme prêteur et son client.

Afin d'avoir avoir une traçabilité de ces flux, l'organisme prêteur met en place un tableau d'amortissement. Il s'agit d'un échéancier annuel ou mensuel qui récapitule tous les flux effectués par les parties afin de pouvoir avoir une traçabilité des dépenses et des recettes.

Nous commençons ce volet par un rappel général sur les emprunts immobiliers, où nous définissons les notations utilisées, présentons les différentes modalités de remboursement, et plus particulièrement le remboursement à échéances constantes.

3.4.1.1 Composantes du tableau d'amortissement

Dans la suite de ce mémoire, nous noterons ainsi les éléments suivants :

- i : le taux d'intérêt annuel de l'emprunt
- i_m : le taux mensuel de l'emprunt
- n_a : la durée de l'emprunt en années
- n_m : la durée de l'emprunt en mois
- C_0 : le capital emprunté ou initial
- C_k: le capital restant dû à l'échéance k, avec k entier.
- I_k : le montant des intérêts versés à l'échéance k
- A_k : l'amortissement à l'échéance k
- $R_k = A_k + I_k$: le remboursement à l'échéance k.
- a_k : le montant de l'annuité

3.4.1.2 Les types de remboursement

Le remboursement d'un capital d'un emprunt s'effectue :

- Soit, en une seule annuité, c'est l'amortissement **in fine**. Ce type de remboursement a l'inconvénient d'obliger l'emprunteur à rembourser la totalité du capital en une seule fois.
- Soit, en plusieurs annuités, ce sont les **prêts amortissables** (par séries égales, par annuités constantes, …). Chaque versement effectué dans le cadre d'un prêt amortissable contient non seulement des intérêts, mais également une partie du capital amorti. L'emprunteur rembourse sa dette au fur à mesure, jusqu'à l'échéance du prêt.

Un tableau d'amortissement est un tableau permettant de faire figurer simultanément plusieurs informations, de la première à la dernière échéance de l'emprunt

Périodes	**Capital restant dû avant amortissement**	**Termes**	**Intérêts**	**Amortissement**	**Capital restant dû après amortissement**
1	C_0	a_1	$I_1 = i \times C_0$	$A_1 = a_1 - I_1$	$C_1 = C_0 - A_1$
2	C_1	a_2	$I_2 = i \times C_1$	$A_2 = a_2 - I_2$	$C_2 = C_1 - A_2$
3	C_2	a_3	$I_3 = i \times C_2$	$A_3 = a_3 - I_3$	$C_3 = C_2 - A_3$
…	…	…	…	…	…
K	C_{k-1}	a_k	$I_k = i \times C_{k-1}$	$A_k = a_3 - I_3$	$C_k = C_{k-1} - A_k$
…	…	…	…	…	…
N	C_{n-1}	a_n	$I_n = i \times C_{n-1}$	$A_n = a_n - I_n$	$C_n = C_{n-1} - A_n$

Avec

- $C_n = 0$
- $C_0 = A_1 + A_2 + A_3 + \ldots + A_n$

D'un point de vue comptable et juridique, l'intérêt est vu comme une indemnité à payer pour avoir privé l'organisme prêteur d'une certaine somme C_0 durant une certaine période t.

3.4.1.3 Amortissement in fine

Le remboursement du capital d'un emprunt s'effectue en une seule fois en fin de contrat.

Dans ce cas $I = C_0 \times \mathrm{i}$

3.4.1.4 Amortissement par séries égales

Le montant amorti à chaque période est de $M_k = \frac{C_0}{n}$

$$I_k = i \times C_k$$

$$I_k = i \times C_0 \times \frac{n-k+1}{n}$$

3.4.1.5 Amortissement par annuités constantes

Le remboursement du capital d'un emprunt s'effectue en une seule fois en fin de contrat. Le montant d'intérêt versé à chaque échéance prévue par le contrat est égal au montant emprunté par le taux nominal.

Dans ce type de contrat, le montant de l'annuité a_k est constante sur les périodes de remboursement.

De ce fait, nous avons $C_0 = \frac{a}{(1+i)} + \Delta\frac{a}{(1+i)^2} + \ldots + \frac{a}{(1+i)^n} \Rightarrow a = C_0 \times \frac{i}{1-(1+i)^{-n}}$

3.4.1.6 Le co-emprunt

Un groupe de deux ou plus individus (co-emprunteurs) peuvent souscrire ensemble à un même prêt. Dans ce cas, l'ensemble de ces personnes seront engagées de la même manière dans le remboursement de cet emprunt. Elles s'engagent donc ensemble à verser une prime en contrepartie d'une couverture d'assurance afin que le prêt soit validé par l'organisme emprunteur.

Elles peuvent de la même manière que dans le cas d'un emprunteur choisir leur quotité de couverture. C'est-à-dire, le ratio de couverture.

Elles peuvent donc s'assurer chacune à 100%, ce qui leur procure une couverture totale et un remboursement de la totalité du capital restant dû au décès de l'une d'elles. Lors de la souscription, le risque assuré est donc le capital restant dû au premier décès. Cependant, après le décès de la première tête, le remboursement de la totalité de l'emprunt amène le risque de la seconde tête à un risque nul, car nous clôturons indéniablement le contrat. Dans ce cas, nous ne pouvons plus observer le comportement de la seconde tête qui crée ainsi un biais dans notre base de données, donc dans l'estimation de la fréquence.

Elles peuvent également s'assurer chacune à hauteur de 50% sur l'assiette du capital restant dû.

> *Si le premier décès survient à la date t, nous observerons le remboursement de 50% du capital restant dû à la date du sinistre. Il restera alors seulement 50% de ce capital restant dû à couvrir par le co-emprunteur survivant au ratio de 50%. Ceci ramène le risque réellement couvert à 25% du capital restant dû à la date t-Δt*

Ou s'assurer seulement à une quotité moindre.

> *Dans ce cas, en cas de sinistre, les personnes seront remboursées au prorata de la couverture choisie par l'individu sinistré, dont l'assiette de calcul est le capital restant dû.*

Enfin, les co-emprunteurs peuvent s'assurer auprès d'organismes différents.

Nous pouvons ainsi souligner que le risque compétitif partiel impacte le capital sous risque alors que le risque compétitif total impacte la fréquence et capital sous risque car n'assure plus la seconde tête. Nous devons donc prendre en compte ces deux scenarios futurs probables dans l'estimation des provisions.

3.4.2 Les engagements de l'assureur

- **Assurance temporaire décès**

Nous pouvons voir cette garantie comme un contrat d'assurance vie dans le cas d'une temporaire décès.

Le contrat d'assurance temporaire a pour objet de garantir le versement d'un capital au moment du décès de l'assuré, si ce décès survient pendant la durée du contrat. Le risque couvert est le décès "toutes causes", c'est-à-dire consécutif à une maladie ou à un accident. Le risque de perte totale et irréversible d'autonomie (PTIA) est également en règle générale couvert par le contrat.

Comme le nom l'indique, le contrat temporaire décès a une durée limitée et répond donc à des besoins de couverture précis (couverture de prêt, protection de ses proches : conjoint, enfants).

- **Couverture en cas d'arrêt de travail**

Lorsque l'assuré entre en arrêt de travail, l'assureur devra également couvrir le remboursement du prêt durant cette période.

- **Cas spécifique de l'Assurance en couverture de prêt**

Dans le cas de l'assurance en couverture de prêt, l'assureur se subroge à l'emprunteur/assuré. En effet, si celui-ci décède avant le terme du prêt, l'assureur se libère de son engagement vis-à-vis de la banque en une seule fois en versant le capital restant dû. Ces capitaux restants dus année après année sont déterminés à partir d'un **tableau d'amortissement**.

3.4.3 Les engagements de l'assuré

En contrepartie de cette couverture, l'assuré doit honorer chaque mois le versement de sa prime d'assurance calculée comme suit :

$$ {}_{p}P_{x} = \frac{\Pi_{x}}{{}_{p}\ddot{a}_{x}} $$

Avec :

- P la prime Annuelle
- p la durée de couverture
- ä l'annuité à terme d'avance
- π_x La prime pure

3.5 La Tarification actuelle avec le modèle individuel

3.5.1 La garantie décès

3.5.1.1 Tarification des contrats individuels

La tarification actuelle consiste en une méthode hybride.

Tout en se basant sur la théorie de l'assurance vie sous la forme de $\text{Capital assuré} \times Probabilité_{décès}$,

Cette méthode consiste à modéliser la $Probabilité_{décès}$ grâce à un modèle de machine Learning nommé Light GBM.

Elle fait donc abstraction des tables de mortalité utilisées habituellement dans les contrats d'assurance vie. En effet, la population du portefeuille en emprunteur est très différente de la population nationale car il y a plusieurs biais de sélection dès le départ.

Cette population est :

- Financièrement plus solvable (endettement interdit au-delà de 35% des capacités financières, assurance emprunteur incluse)
- Plus jeune (entre 18 et 65 ans)
- En meilleure santé (grâce aux sélections médicales).

Nous pouvons voir cette modélisation de fréquence de décès comme la mise en place d'une table de mortalité dynamique.

3.5.1.2 Tarification des contrats couples

Lors de la souscription, chacun des risques assurés composant un couple est estimé individuellement, puis une réduction de 10% à vie est accordée sur la cotisation du total. Si nous notons P'' la prime commerciale, la cotisation périodique devient alors :

$$ P'' = 0.9(P''_1 + P''_2) $$

3.5.2 La garantie arrêt maladie

La garantie Incapacité totale de travail est tarifée de la même manière que le décès. Il s'agit ici d'estimer la probabilité annuelle $\mathbb{p}_{ITT}$ d'entrer en Incapacité et non celle d'établir une loi de maintien. Pour faire la relation avec l'analyse de survie, nous pouvons le voir à l'image du taux de décès instantané. Dans notre cas, il s'agirait plutôt du taux instantané d'entrer en incapacité totale de travail.

3.6 Analyse de survie

3.6.1 Variable de durée

A l'origine, les modèles de survie ont été développés pour étudier la durée de vie. Cet outil a pu néanmoins être utilisé dans l'étude d'autres phénomènes comme par exemple la durée du chômage, la durée de vie d'une machine ou la durée d'un contrat.

La formalisation de l'analyse de durée date de l'école Anglaise d'Arithmétique. John Granut (1620-1674) et William Petty (1623-1687) l'ont employée pour l'étude de mortalité de la population anglaise au 17ème siècle. C'est à cette époque (cf. Droesbeke et al. (1989), Klein and Moeschberger (2005)) que les notions d'espérance de vie et d'espérance de vie résiduelle ont été définies.

Benjamin Gompertz en 1825 a mis en place un modèle pour la probabilité de décéder à l'âge t

$$\lambda\,(\mathrm{t}) = \mathrm{a}.\,\mathrm{bt}$$

C'est une progression géométrique des taux de décès de raison b. William Makeham viendra compléter cette équation en 1860 :

$$\lambda\,(\mathrm{t}) = \mathrm{c} + \mathrm{a}.\,\mathrm{b}\,\mathrm{t}$$

la composante c, sera indépendante de l'âge, et traduira le décès accidentel.

Enfin, Waloddi Weibull viendra bouleverser l'étude des durées de vie grâce à la théorie de la fiabilité pour les systèmes physiques. Ainsi W. Weibull publiera un article dans un journal de mécanique en 1951 dans lequel il propose la fonction de hasard (ou de risque instantané, ou taux de panne, ou taux de défaillance, ou taux de décès, ...)

$$\lambda\,(\mathrm{t}) = \lambda\alpha t^{\alpha-1}$$

Le mot « hasard » est un anglicisme ("hazard" signifie davantage "risque" que "hasard"). Néanmoins W. Weibull aborde surtout l'une des particularités importantes des données de durée, à savoir la présence de données :

- Tronquées
- Censurées

Deux autres professeurs doivent être citées, en effet en 1958, l'article d'E. Kaplan et P. Meier propose d'utiliser un estimateur non paramétrique permettant d'intégrer les données censurées. C'est ainsi que l'application de l'estimateur de Kaplan Meier a vu le jour dans le domaine médical.

David Cox, en 1972 fait suite à cette découverte et pose les bases d'un cas particulier important de modèle à « hasard proportionnel ». Il suggère l'usage de variables explicatives exogènes X en proposant une fonction de hasard comme

$$\lambda(\mathrm{t}) = \lambda_0(\mathrm{t}).\exp(\mathrm{X}\,\beta)$$

où $\lambda_{0(t)}$ est la fonctionne hasard de base, c'est-à-dire, celui qui correspond au taux de hasard de la population de référence, et β, un vecteur de coefficients. C'est un modèle semi-paramétrique car $\lambda_{0(t)}$ ne dépend pas de paramètres (cf. Cameron and Trivedi (2005)). Ce modèle de référence a donné lieu à de nombreux développements.

C'est désormais ce que nous utilisons en assurance (cf. Charpentier (2014), Johansson and Ohlsson (2010)) dans le but de :

- Créer des tables de mortalité à partir des données démographiques (Insee, Ined)
- Créer des tables de mortalité d'expérience à partir des données spécifiques à chaque compagnie d'assurance
- Estimer les probabilités d'entrée en arrêt de travail

Pour qu'une variable de durée T soit définie comme telle, elle doit être strictement positive.

D'autre part, il faut savoir que T pâtit souvent de problèmes de censure et troncature.

En effet, l'arrêt de la collecte à une date donnée fait que des durées commencées n'ont parfois pas le temps de se terminer et sont ainsi censurées. Nous pouvons toutefois affecter des poids moins importants aux durées incomplètes. On parle de censure à droite.

Inversement, il est possible de commencer l'extraction des données à une date où le processus observé a déjà commencé pour certains individus. La durée est alors censurée à gauche, il faudra donc y affecter des poids de la même manière qu'en cas de censure à droite.

Afin d'estimer le risque étudié au plus juste, il est important de tenir compte de toutes les observations, qu'elles soient censurées ou complètes. De plus, il faudra garder en mémoire que plus une durée est longue, plus elle a de chance d'être censurée, de sorte qu'enlever les durées censurées revient à causer un biais de sélection. Par exemple, si l'on étudie la durée d'un prêt, enlever les données censurées reviendrait à réaliser une étude sans les emprunteurs de longue durée.

Comme pour les variables aléatoires réelles, on définit la loi d'une variable de durée par sa fonction de répartition $F_T(t)$, cependant, pour des raisons pratiques, nous préférons utiliser des concepts spécifiques comme le taux de mortalité noté $\lambda(t)$, la probabilité de survie notée $S(t)$ ou l'espérance de vie à la naissance.

Soit une variable aléatoire de durée $T > 0$. Sa fonction de répartition est définie par :

$$F(t) = \Pr[T \leq t], t \in t = 1 - F(t), t \in \mathbb{R}_*^+$$

Sa fonction de densité donne la fraction d'individus d'une génération ayant survécu jusqu'à l'âge t. La densité de la durée est donnée par :

$$f(t) = \frac{dF(t)}{dt} = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t)}{\Delta t}$$

Elle représente l'intensité d'occurrence d'une durée exactement égale à t. Cette intensité peut être supérieure à l'unité car il ne s'agit pas d'une probabilité mais d'une densité.

$$f(t) = \frac{dF(t)}{dt} = \frac{d}{dt}\big(1 - S(t)\big) = -\frac{dS(t)}{dt}$$

Cependant, nous préférons en pratique utiliser le concept de fonction de survie S(t) qui donne la probabilité que la durée soit supérieure à une valeur donnée t

$$S(t) = \mathbb{P}(T > t) = 1 - F(t)$$

Nous pouvons enfin définir la fonction de hasard

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(t < T \leq t + \Delta t | T > t)$$

La fonction de hasard λ (t) représente une occurrence de mortalité instantanée. C'est la probabilité conditionnelle de sortir (i.e. décéder) à la date t sachant que l'on vécut jusqu'à cette date.

où $\mathbb{P}(B|A) = \mathbb{P}(T > t \mid t < T \leq t + \Delta t) = 1$

d'où $\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}[t < T \leq t + \Delta t | T > t]$

$$= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\mathbb{P}[t < T \leq t + \Delta t]\mathbb{P}(T > t)}{\mathbb{P}(T > t)}$$

$$\frac{1}{\mathbb{P}(T > t)} \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(t < T \leq t + \Delta t)\, f(t)\, S(t)$$

$$= \frac{1}{\mathbb{P}(T > t)} f(t) = \frac{f(t)}{S(t)}$$

$$\frac{-1}{S(t)} \frac{dS(t)}{dt} = \frac{-d \ln S(t)}{dt}$$

C'est ce taux de sortie (ou de mortalité) qui est particulièrement utilisé en démographie et en actuariat.

On en déduit la survie et la densité de la durée :

$$S(t) = \exp\left\{-\int_0^t \lambda(x)dx\right\} \text{ et } f(t) = \lambda(t) \exp\left\{-\int_0^t \lambda(x)dx\right\}$$

Le hasard cumulé ou hasard intégré (cumulative hazard, integrated hazard) est défini par :

$$\Lambda(t) = \int_0^t \lambda(x)dx = -\ln S(t)$$

C'est l'accumulation du hasard au cours du temps. Si $S(\infty) = 0$, alors $\Lambda(\infty) = \infty$. Ceci implique que $\lambda(t)$ ne doit pas décroître trop rapidement sinon le hasard cumulé divergera. Ainsi $\lambda(t) = \exp(-t)$

Enfin, nous avons l'espérance de la durée (i.e. l'espérance de vie à la naissance) est définie par

$$E(T) = \int_0^\infty xf(x)dx = \int_0^\infty S(x)dx = \int_0^\infty \exp\left(\left\{-\int_0^x \lambda(u)du\right\}\right)dx$$

et la durée moyenne restante (ou résiduelle) est l'espérance de la durée qui reste sachant que l'on a déjà atteint t :

$$r(t) = E(T - t|T > t)$$

C'est, par exemple, l'espérance de vie à un âge donné.

En résumé, nous avons :

fonction	symbole	définition	relations
densité	$f(t)$		$f(t) = \frac{dF(t)}{dt}$
répartition	$F(t)$	$\Pr[T \leq t]$	$F(t) = \int_0^t f(x)dx$
survie	$S(t)$	$\Pr[T > t]$	$S(t) = 1 - F(t)$
hasard	$\lambda(t)$	$\lim_{\Delta t \to 0} \frac{\Pr[t<T\leq t+\Delta t \mid T>t]}{\Delta t}$	$\lambda(t) = \frac{f(t)}{S(t)}$
hasard cumulé	$\Lambda(t)$	$\int_0^t \lambda(x)dx$	$\Lambda(t) = -\ln S(t)$

Figure 17 - fonctions utiles à l'analyse de survie

3.6.2 Estimateurs non paramétriques (Kaplan-Meier et Nelson-Aalen)

Lorsqu'aucune hypothèse ne veut être faite sur la distribution des temps de survie, l'estimateur de la fonction de survie le plus utilisé est l'estimateur de Kaplan-Meier. Nous n'en donnons qu'une dérivation heuristique. Les démonstrations relèvent d'une écriture en termes de processus de martingales (cf. Fleming and Harrington (1991), Planchet et Thérond (2011)). Si l'estimateur de Kaplan-Meier est utile pour estimer une fonction de survie, on peut être intéressé par l'estimation d'autres fonctions qui caractérisent la distribution des temps d'évènements. C'est le cas par exemple de l'estimation de la fonction de hasard cumulé, avec l'estimateur de Nelson-Aalen.

En plus de la mise en évidence de la fonction de survie ou du hasard cumulé, il n'est pas rare de s'interroger sur d'éventuels écarts entre les distributions de deux ou plusieurs sous-population (ex : hommes vs femmes, individus soumis à un traitement particulier vs groupe de contrôle, assurés vs non assurés, ...). Cette mise en évidence d'écarts significatifs est souvent utilisée comme une première étape de sélection de variables explicatives avant la mise en œuvre d'estimations de modèles paramétriques ou semi-paramétriques. L'estimateur de Kaplan-Meier généralise la notion de fonction de répartition empirique en tenant compte des données censurées à droite. C'est pourquoi il sert généralement de base à toute étude sur les durées. Il peut en effet guider le choix d'une forme paramétrique particulière. Il doit être calculé pour des populations homogènes.

3.6.2.1 Cas sans censure

Supposons qu'il n'y ait pas de censure. Alors la survie en t peut être simplement estimée par :

$$\hat{S}(t) = 1 - \hat{F}(t)\,, \hat{F}(t) = \frac{n_t}{N}$$

où n_t est nombre de durées inférieures à t et N le nombre total d'observations. On peut remarquer que la fonction de survie estimée peut s'écrire simplement comme un produit de probabilités conditionnelles. Dans le cas simple sans censure et si on n'observe qu'une seule fois chaque valeur de durée, que l'on notera dans l'ordre croissant $t_0, t_1, \ldots, t_N$ avec t_0 = 0, alors :

$$\mathrm{S(t)} = \mathbb{P}(\mathrm{T} > \mathrm{t}) = \prod_{t_i \leq t} \mathbb{P}(\mathrm{T} > t_i \,|\mathrm{T} > t_{i-1}) = \prod_{j<i} 1 - q_i$$

où q_i est la probabilité instantanée de sortir en tj (l'équivalent de la fonction de hasard en temps discret). Cette probabilité q_j vaut alors $\frac{1}{(\mathrm{N-j+1})}$, puisqu'on observe une sortie en j parmi les $N-(j-1)$ personnes qui survivent juste après $t_{j-1} q_1 = \frac{1}{N}, q_2 = \frac{1}{N-1}, \ldots,$

Ces $N-(j-1)$ personnes sont l'ensemble à risque en t_j.

3.6.2.2 Cas avec censure

Si maintenant certaines durées sont censurées à droite, on adapte la notion d'ensemble à risque en t_j. Il sera cette fois défini comme le nombre r_j d'observations ni sorties, ni censurées avant t_j. Alors l'estimateur de $\mathrm{q_i}$ s'écrira $\frac{1}{\mathrm{r_j}}$, et la survie sera estimée par $\prod_{\mathrm{j<i}} 1 - \frac{1}{\mathrm{r_j}}$ de sorties à chaque date j, l'estimateur de Kaplan-Meier pour le hasard à la date j sera :

$$\hat{\lambda}(t_j) = \frac{d_j}{r_j}$$

Et celui de la survie s'écrira

$$\hat{S}(t_j) = \prod_{j|t_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

C'est une fonction décroissante par escalier avec un saut à chaque valeur de durée $(\mathrm{t_0}, \mathrm{t_1}, \ldots, \mathrm{t_n})$.

On peut utiliser cet estimateur non paramétrique de la survie pour calculer une durée moyenne. Comme l'espérance de la durée peut généralement s'écrire :

$$\mathrm{E(T)} = \int_0^\infty \mathrm{S(x)dx}$$

Alors la durée moyenne de vie est :

$$\hat{E}(\mathrm{T}) = \sum_{i=1}^{I} (t_i - t_{i-1})\ \hat{S}(t)$$

Où I est le nombre de durées différentes observées. La durée moyenne ne sera donc la moyenne empirique que s'il n'y a pas de censure.

A partir de l'estimateur de Kaplan-Meier du hasard $\lambda(t_j) = \frac{d_j}{r_j}$, on définit l'estimateur de Nelson-Aalen du hasard cumulé :

$$\hat{\Lambda}(t) = \sum_{j|t_j \leq t} \hat{\lambda}(t_j) = \sum_{j|t_j \leq t} \frac{d_j}{r_j}$$

3.7 Le modèle de Poisson trick

Soit $x_i(t)$ la valeur de vecteur de variables pour l'individu i au moment de la durée t, alors le modèle des risques proportionnels peut être généralisé comme suit :

$$\lambda_i\big(t, x_i(t)\big) = \lambda_0(t)e^{\beta x_i{}'(t)}$$

Dans ce type de modèle, il est difficile d'étudier les effets qui sont fortement colinéaires avec le temps. Le calcul de variables de survie lorsque nous avons des variables variant avec le temps est un peu plus compliqué, car nous devons spécifier un chemin ou une trajectoire pour chaque variable.

Dans ce cas, nous n'utiliserons plus de modèles proportionnels, mais nous pouvons réécrire le problème sous forme de l'équation suivante :

$$\lambda_i(t, x_i) = \lambda_0(t)e^{x_i{}'(t)\beta}$$

Où le paramètre β dépend mainteannt du temps. Ce modèle permet une plus grande généralisation de notre cas, qui se réécrit désormais :

$$\left\{\begin{matrix} \lambda_i(t, x_i) = \lambda_0(t)\ si\ x = 0 \\ \lambda_i(t, x_i) = \lambda_0(t)e^{\beta(t)} si\ x = 1 \end{matrix}\right\}$$

Ce qui nous permet d'avoir deux fonctions de hasard, une pour chaque groupe. Il s'agit alors d'un modèle saturé.

Les extensions précédentes aux variables variant dans le temps et aux effets dépendants du temps peuvent être combinées pour donner la version la plus générale du modèle du taux de risque comme suit :

$$\lambda_i(t, x_i(t)) = \lambda_0(t)e^{x_i(t)'\beta(t)}$$

Dans ce cas, nous remarquons que $x_{i(t)}$ et $\beta(t)$ sont des variables qui dépendant du temps et représentants respectivement les caractéristiques de l'individu i dépendant du temps et les covariables dépendantes du temps.

Dans certains cas, les covariables $x_{i(t)}$ et $\beta(t)$ peuvent être corrélées. Pour mieux préciser le modèle, nous pouvons faire leur ajustement.

Pour se faire, il existe essentiellement trois approches.

3.7.1 Une approche non paramétrique

Celle-ci se concentre sur l'estimation des coefficients de régression β. Elle laisse donc le risque de base $\lambda_0(t)$ inconnu. Elle s'appuie sur le modèle de Cox (1972) qui propose l'estimation d'une fonction de vraisemblance partielle.

3.7.2 Une approche paramétrique

Cette méthode suppose une forme fonctionnelle spécifique pour la fonction risque de base $\lambda_0(t)$. Nous pouvons prendre comme exemples les modèles basés sur les distributions Galla, Weibull…

3.7.3 Une approche flexible ou semi-paramétrique

Celle-ci se fonde sur une stratégie qui fait l'hypothèse que le risque de base $\lambda_0(t)$ est constant par morceaux sur chaque petite subdivision de l'intervalle total de l'exposition au risque. Ceci aboutit à un modèle exponentiel par morceaux.

3.7.3.1 Le modèle exponentiel par morceaux

Nous allons donc considérer dans cette étude le modèle exponentiel par morceaux. Il s'agit d'un modèle de risques proportionnels de la forme :

$$\lambda_i(t,|x_i) = \lambda_0(t)e^{x_i(t)'\beta}$$

Sous les hypothèses relativement faibles concernant le hasard λ_0(t) qui serait constant par morceaux.

Nous pouvons donc voir le hasard de base comme $\lambda_0(t) = \lambda_j$ pour tout $t \in [\tau_{j-1}, \tau_j]$ qui est une subdivision de l'intervalle de temps correspondant à l'exposition de chaque individu de la base. Ceci aboutit donc à une fonction de survie calculée à partir de $\lambda_0(t)$ qui est constante par morceaux (sur les petits intervalles de temps). Nous pouvons choisir ces intervalles de temps de manière judicieuse afin de s'adapter au mieux aux variations du risque étudié. Nous pouvons par exemple tenir compte des durées de fluctuations signifiantes afin que la perte d'information sur notre modélisation n'impacte pas beaucoup les résultats.

3.7.3.2 Un modèle d'aléa proportionnel

En supposant que le risque de base est constant par morceaux $\lambda_0(t)$, nous proposons ainsi le modèle suivant :

$$\lambda_{ij} = \lambda_j e^{x_i'(t)'\beta}$$

Où λ_{ij} est le risque correspondant à l'individu i dans le $jeme$ l'intervalle.

Nous pouvons calculer

$$\log(\lambda_{ij}) = \alpha_j + x_i'(t)'\beta$$

Où α_j représente le logarithme du hasard de base λ_j. Nous aboutissons ainsi à un modèle log-linéaire standard dans lequel les catégories de durées sont traitées sous forme de classes.

3.7.3.3 Le modèle Poisson équivalent

Holford (1980), Laird et Oliver (1981) ont découvert que le modèle à risque proportionnel par morceaux décrit précédemment était équivalent à un modèle de régression de Poisson. Nous rappelons que ce mémoire porte sur l'étude du temps total t_i vécu par le $ieme$ individu de notre portefeuille avant de subir un sinistre (décès ou arrêt maladie) noté $X_i \in [\![0,1]\!]$. A partir de ces informations, nous pouvons calculer des mesures pour chaque intervalle traversé par nos individus.

- La mesure de l'exposition e_{ij} qui désigne le temps vécu par l'individu i dans le $j^{\text{ième}}$ intervalle $[\tau_{j-1}, \tau_j]$.
  Si l'individu n'a pas eu de sinistres alors e_i est égal à la valeur de l'intervalle. En cas de censure, e_i tiendra donc compte de cette information.
- L'indicateur de sinistralité X_{ij} qui vaut 1 si l'individu i est sinistré dans l'intervalle $[\tau_{j-1}, \tau_j]$ et 0 sinon.

Par la suite nous ajusterons le modèle exponentiel par morceaux en supposant que les indicateurs de sinistralité X_i constituent un échantillon de Poisson, i.e. des variables aléatoires de Poisson indépendantes :

$$\mu_{ij} = \lambda_{ij} e_{ij}$$

Or nous avons vu précédemment que le logarithme du taux de risque satisfait l'équation suivante :

$$\log(\mu_{ij}) = \log(t_{ij}) + \alpha_j + x_i'\beta$$

Nous remarquons donc que le risque proportionnel exponentiel par morceaux est équivalent à un modèle log-linéaire de Poisson pour chaque individu i vivant dans l'intervalle j.

Contrairement à ce que l'on pourrait penser, nous remarquons que les variables considérées ne suivent pas une loi de Poisson (qui attribue une probabilité pour les valeurs strictement positives). Ce sont plutôt les fonctions de vraisemblances qui coïncident. Ainsi donc, pour une réalisation de processus de survie exponentiel, nous pouvons trouver une réalisation d'observations de Poisson indépendantes qui auraient les mêmes estimateurs du modèle présenté dans ce mémoire.

3.7.4 Modèle de Machine Learning : LightGBM

3.7.4.1 Introduction

Le développement du Big Data invite beaucoup de domaines à l'utilisation des techniques de machine Learning. C'est ainsi que l'actuariat est de plus en plus enclin à utiliser cette nouvelle technologie dans sa tarification et ses analyses de sinistralité. Parmi les différentes propositions d'algorithmes d'apprentissage automatique, on y trouve les méthodes dites de boosting. Il s'agit de modèles ensemblistes, supervisés et non paramétriques. Mathématiquement, nous pouvons formaliser cela comme suit :

Soit un tableau de n individus avec m caractéristiques

$\mathfrak{D} = \{(x_i, \lambda_i)\}$ $(|\mathfrak{D}| = n, x_i \in \mathbb{R}^m, \lambda_i \in \mathbb{R})$ est l'ensemble de m arbres qui fournit K prédicteurs après entrainement.

La somme de prédicteurs issus de modèles faibles est calculée itérativement. La prédiction $\hat{\lambda}_i$ est calculée comme suit :

$$\hat{\lambda}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \ f_k \in \mathcal{F}$$

Où $\mathcal{F} = \{f(x) = \omega_{q(x)}\}\{q : \mathbb{R}^m \to T, \omega \in \mathbb{R}^T\}$ est l'espace des arbres de régression (également connu sous le nom de CART).

Ici, q représente la structure de chaque arbre qui donne un exemple correspondant à chaque feuille, l'indice de la feuille correspondante. T est le nombre de feuilles de l'arbre. Chaque f_k correspond à une structure d'arbre indépendante q et aux poids ω des feuilles. Contrairement aux arbres de décision, chaque arbre de régression contient un score obtenu à partir de chaque feuille. On utilise $\omega_i$pour représenter le score de la i-ème feuille. Pour un exemple donné, nous utiliserons les règles de décision des arbres (données par q) pour classifier dans les feuilles et calculons la prédiction finale. Cette prédiction est la somme des scores dans les feuilles correspondantes (donné par ω).

3.7.4.2 L'espace des arbres Classification and Regression Tree (CART)

Le socle d'un modèle de boosting est un arbre CART. Cet arbre tente de partitionner l'espace des caractéristiques X en groupes homogènes (regroupés sous une feuille) tout en tenant compte de la variable de prédiction $\hat{\lambda}_i$. Pour se faire, l'arbre partitionne l'espace des caractéristiques de manière récursive. Cet algorithme calcule à chaque étape, les paramètres variable $\mathrm{X}_j(t)$ et seuil $a_j(t)$ qui permettent de scinder au mieux chaque groupe d'individu en sous-groupes homogènes rattachés à des nœuds.

Cette action est répétée autant de fois que nécessaire jusqu'à l'obtention de nœuds les plus purs possibles. C'est alors que la séparation des individus s'arrête car ils ont tous les mêmes caractéristiques ou alors nous n'avons plus assez d'individu ai sein des nœuds pour que le processus puisse continuer sans faire du sur-apprentissage. Il est d'autre part tenu compte du gain de pureté par rapport au temps computationnel, nous arrêterons ainsi la séparation, si ce ratio est très faible.

En effet, l'objectif du gain de pureté est de réduire le mélange des classes au sein du sous ensemble de résultat d'un algorithme de classification. Ceci peut être calculé par l'indice de Gini par exemple.

L'optimisation des variables $\mathrm{X}_j(t)$ et $a_j(t)$ qui permettent scission des nœuds $N_{t>}$ en deux nœuds $N_{t>}$ est pilotée par la fonction $I(.)$ qui mesure l'impureté du nœud en question.

Dans le cas de la régression, il s'agit de variance de la variable réponse chaque nœud. Plus précisément, le couple $(\mathrm{X}_j(t), a_j(t))$ découle de la résolution de l'équation suivante :

$$(\mathrm{X}_j(t), a_j(t)) = arg \max_{\substack{\mathrm{j}\in 1,2,\ldots,\mathrm{m} \\ a_j \in \Omega(X_j)}} \left(I(\mathcal{N}_t) - \left(\frac{|\mathcal{N}_{t,<}|}{|\mathcal{N}_t|}\right) I(\mathcal{N}_{t,<}) + \left(\frac{|\mathcal{N}_{t,>}|}{|\mathcal{N}_t|}\right) I(\mathcal{N}_{t,>}) \right)$$

Dans cette formules, | . | désigne le cardinal d'un nœud et $I(.)$ est la fonction qui calcule la variance intra-échantillon.

Au sein de l'arbre, les nœuds terminaux sont appelés feuilles. Dans le cas de la régression, l'arbre calcule la moyenne des réponses des individus appartenant au même nœud. Ainsi lors de la prédiction, il classe l'individu nécessitant une réponse, dans un des nœuds et lui attribue par la suite la moyenne associée aux individus du nœud en guise de réponse.

La structure des arbres leur permet naturellement de capter les interactions entre les variables, y compris lorsque ces liens ne sont pas linéaires. De plus, leur intelligibilité et leur possibilité de visualisation leur confère l'avantage d'être explicables, ce qui leur vaut leur popularité. Cependant, l'idée d'utiliser un seul arbre manque de robustesse du fait de leur grande sensibilité à l'échantillonnage. Afin de conférer une plus grande stabilité au modèle, nous pouvons nous rabattre sur les méthodes ensemblistes comme le gradient boosting. Il s'agit ici de faire intervenir plusieurs arbres afin de diminuer la variance inter-échantillons qui est calculé ainsi :

$$I(.) = \frac{1}{|\mathcal{N}|}\sum_{i=1}^{|\mathcal{N}|}(y_i - \bar{y})^2$$

3.7.4.3 Gradient Boosting des arbres CART

La méthode du gradient boosting des arbres propose de construire des arbres de manière itérative.

A chaque itération, des poids plus importants sont donnés aux échantillons en erreurs, afin que leurs prédictions soient revues et corrigées par es arbres suivants. L'optimisation est faite pas à pas dans le but de réduire l'erreur par descente de gradient. Les arbres utilisés dans le mécanisme sont dits « modèles faibles », ceci est dû à leur faible profondeur.

Pour mieux comprendre, nous réécrirons l'équations précédente de la manière suivante :

$$\begin{aligned} F_K(\boldsymbol{x}) := \sum_{k=1}^{K}\gamma_k f_k(\boldsymbol{x};\boldsymbol{\theta}_k) &= \sum_{k=1}^{K-1}\gamma_k f_k(\boldsymbol{x};\boldsymbol{\theta}_k) + \gamma_K f_K(\boldsymbol{x};\boldsymbol{\theta}_K) \\ &= F_{K-1}(\boldsymbol{x}) + \gamma_K f_K(\boldsymbol{x};\boldsymbol{\theta}_K) \end{aligned} \qquad (2.3)$$

où $\gamma_k, k = 1, 2, \cdots, K$ sont des paramètres de dilatation qui guident l'importance de chacun

$$\hat{\lambda}_{K(x)} = \sum_{k=1}^{K}\gamma_k f_k(x;\theta_k) = \sum_{k=1}^{K-1}\gamma_k f_k(x;\theta_k) + \gamma_K f_K(x;\theta_K) = \hat{\lambda}_{K-1(x)} + \gamma_K f_K(x;\theta_K) =$$

Avec

- $\gamma_k = \log\left(\frac{1}{\beta_k}\right)$
- β_{mk} est l'erreur relative de l'arbre numéro k
- K ∈ $[\![1, K]\!]$

Le régresser $\hat{\lambda}_{K(x)}$final choisi sera celui qui va maximiser la vraisemblance de notre modèle. Cependant, cette condition est difficile à calculer directement, c'est pour cela que nous approcherons la solution par la méthode récursive de boosting. Nous partirons du postulat que maximiser le maximum de vraisemblance du régresser $\hat{\lambda}_{K(x)}$ revient à minimiser la fonction de perte commise par la modélisation.

Ceci revient donc à calculer

$$(\gamma_k, \theta_k) = \min_{(\gamma,\theta)}\sum_{i=1}^{n} l(y_i - \hat{\lambda}_{m-1(x_i)} + \delta\theta(x_{i,\theta}))$$

Avec $l(\hat{\lambda})$ l'erreur commise lors de l'estimation du taux de hasard λ.

3.7.4.4 Spécificités de l'algorithme LightGBM

LightGBM est un algorithme de machine learning appartenant à la famille des méthodes de boosting par gradient, des arbres de décision. Il a été mis en place par Ke et al en 2017 afin de suggérer une version différente de XGBoost. LightGBM est un algorithme plus efficace car il présente une vitesse d'exécution supérieures à celles du XGBoost (Chen et Guestrin, 2016). En effet, lightGBM présente un temps computationnel allant jusqu'à vingt fois plus vite que XGBoost, tout en gardant des performances de prédiction similaire. La majorité des algorithmes de Boosting par Gradient construisent leurs arbres par divisions des nœuds par étage. Ceci veut dire qu'à chaque itération, les nœuds sont divisés pour que chacun d'eux propose deux sous-branches, et ceci de manière récursive. Il s'agit de l'architecture « level-wise tree growth ».

Contrairement à cette méthode de développement horizontale, LightGBM construit ses arbres plutôt par nœud. Ici l'algorithme se déploie de manière verticale à en séparant les branches de l'arbre à partir d'une feuille. Cette méthode est couramment appelée « leaf-wise tree growth ». LightGBM permet de se concentrer en amont sur les nœuds avec la perte maximale, afin de résoudre le problème de manière plus efficiente.

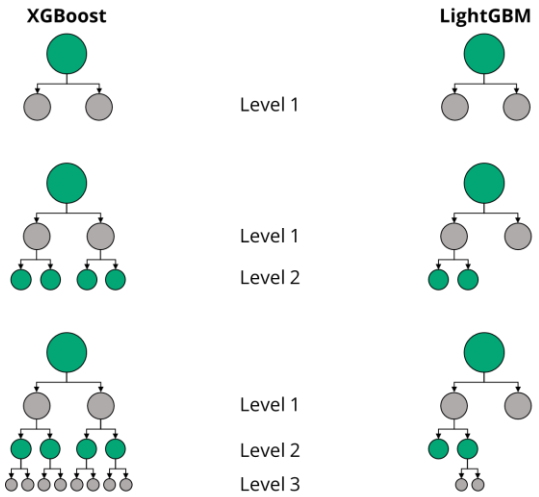


Figure 18 – Différence de fonctionnement entre XGBoost et LightGBM

La prime pure se calcule donc comme ci-dessous

$$P = \sum_{i=1}^{n} \hat{\lambda}_i CRD_i$$

3.7.5 Evaluation du modèle LightGBM

3.7.5.1 Introduction

L'analyse fournit de nombreuses métriques calculées en fonction de la sinistralité historique et les prédictions du modèle utilisé. Nous pouvons donc utiliser eux types d'évaluations :

- **Calibrage** : ces quantités mesurent le degré de correspondance entre les valeurs prédites et les valeurs réelles. Nous pouvons par exemple utiliser des mesures d'étalonnage comme le MSE, la MAE, ou le score de vraisemblance.
- **Discrimination :** Ces quantités mesurent quant à elle la capacité du modèle à faire la distinction entre les niveaux de réponses. En effet, elles évaluent la capacité du modèle à classer correctement les échantillons sur la base des valeurs prédites. Nous pouvons ainsi citer l'indice de Gini en exemple.
- Définition des variables : Avant d'entamer les calculs, nous définirons quelques variables.

Soient :

- y_i la valeur de la sinistralité historique de l'individu i
- $\mu_i = \hat{y}_i = \pi(X_i)$ la prédiction du modèle en guise de réponse pour l'individu i, avec π le prédicteur.
- n est le nombre d'individus dans la base utilisée lors de la modélisation.
- ω_i est le poids de l'individu i au cours de la modélisation. Ce poids est communément appelé : exposition en langage actuariel.

3.7.5.2 Calcul des métriques

- **La ration entre les prédictions et l'historique**
  Cet indicateur évalue l'adéquation entre la sinistralité totale prédite et la sinistralité observée. Il donne une indication sur la calibration globale du modèle. Plus cette valeur est proche de 1 plus l'adéquation globale est bonne.

$$\frac{Total - Pred}{Total - Obs}(y, \mu) = \frac{\sum_{i=1}^{n} e_i \mu_i}{\sum_{i=1}^{n} e_i y_i}$$

- **Indice de Gini**
  Un bon modèle doit refléter l'hétérogénéité de la réponse observée dans ses prédictions. Le coefficient de Gini mesure l'adéquation d'un modèle aux données. Il est obtenu par le rapport des coefficients de Gini des valeurs prédites par celui des valeurs réelles.

- **Déviation moyenne**
  Cette métrique estime la perte d'apprentissage du modèle considéré. Elle évalue donc l'écart entre la valeur prédite et la valeur réelle.

$$\boldsymbol{Dev_{moy}}(y, \mu) = \frac{\mathbf{1}}{\sum_{i=1}^{n} e_i} \sum_{\mathbf{1}}^{n} e_i d(y_i\ ,\ \mu_i)$$

- **Déviance expliquée**

  $\boldsymbol{Dev_{expliquée}}(y, \mu) = 1 - \frac{\mathcal{D}_{Model}}{\mathcal{D}_0}$, avec $\mathcal{D}_{Model}$ la déviance du modèle étudié et $\mathcal{D}_0$ celle du modèle basique (qui prédit la moyenne de la sinistralité). Il s'agit de la généralisation du R^2, donc plus sa valeur est proche de 1 plus le modèle est adéquat.

- **MSE** (Mean Squared Error)

$$MSE(y, \mu) = \frac{\mathbf{1}}{\sum_{i=1}^{n} e_i} \sum_{\mathbf{1}}^{n} e_i (y_i - \mu_i)^2$$

  Il s'agit de la moyenne des carrés des différences entre les prédictions et les sinistres historiques.

- **RMSE** (Root Mean Squared Error)

$$MSE(y, \mu) = \sqrt{\boldsymbol{MSE}(y, \mu)} = \sqrt{\frac{\mathbf{1}}{\sum_{i=1}^{n} e_i} \sum_{\mathbf{1}}^{n} e_i (y_i - \mu_i)^2}$$

  Le RMSE est utilisé pour évaluer la magnitude des erreurs du modèle.

- **MAE** (Mean Absolute Error)

$$MSE(y, \mu) = \frac{\mathbf{1}}{\sum_{i=1}^{n} e_i} \sum_{\mathbf{1}}^{n} e_i |y_i - \mu_i|$$

  C'est la moyenne des valeurs absolues des écarts entre la sinistralité historique et la prédiction.

Il faut remarquer que le MSE, RMSE et la MAE sont sensibles aux valeurs extrêmes. Ces métriques utilisées seules ne donneront pas d'indication locales sur le modèle.

3.7.6 Explication du modèle LightGBM

3.7.6.1 Par le ratio de Shap

Les modèles de machine Learning (ML) sont souvent qualifiés de « boites noires ». En effet, nous évaluons souvent leurs performances sur un ensemble de données, sans savoir exactement pourquoi et comment les décisions sont prises. Or, si obtenir les meilleures performances possibles peut suffire dans certains cas, nous remarquons que cela devient insuffisant. Nous avons désormais besoin d'expliquer les décisions prises par cette intelligence artificielle afin de ne pas laisser une machine aller au-delà du domaine légal.

Si un individu demande des explications, à son conseiller sur le motif de refus de son prêt, celui-ci doit avoir la capacité de les lui fournir.

C'est dans ce cadre que le ratio de Shap intervient. Les avantages que procure sa capacité d'explication est inouïe. Cela permet une meilleure collaboration entre les différents départements qui œuvrent pour la même entreprise, une meilleure maitrise de la modélisation et une meilleure connaissance du produit étudié. Enfin, cela permet une transparence sur les décisions prises afin de rester en adéquation avec la réglementation.

La méthode de Shap est utilisable quel que soit le modèle, même un modèle de type black box, inconnu ou de type Depp Learning.

Définition du ratio de Shap

Les valeurs Shapley ont été introduites par le mathématicien Lloyd Shapley en 1953 dans le domaine de la théorie des jeux coopératifs. Lloyd Shapley était un éminent chercheur en économie et en mathématiques, lauréat du prix Nobel d'économie en 2012 pour ses contributions à la théorie des jeux.

L'idée fondamentale derrière les valeurs Shapley découle de la question de savoir comment attribuer équitablement les gains obtenus dans un jeu coopératif à chaque joueur (ou contributeur) du jeu. Dans un contexte où plusieurs joueurs collaborent pour obtenir un gain global, les valeurs Shapley fournissent une méthode pour attribuer de manière équitable la contribution de chaque joueur à la valeur totale du jeu.

Les valeurs Shapley sont basées sur le concept d'ajouts marginaux. L'idée est que la contribution d'un joueur à un jeu dépend de la valeur qu'il ajoute en rejoignant n'importe quel sous-groupe de joueurs. Les valeurs Shapley quantifient l'apport moyen attendu d'un joueur lorsqu'il joue avec tous les sous-groupes possibles de joueurs.

Les valeurs Shapley sont largement utilisées dans la théorie des jeux pour résoudre divers problèmes, notamment la répartition équitable des coûts, la formation de coalitions et la prise de décision collective. Au fil des années, ces valeurs se sont révélées extrêmement utiles et ont été appliquées à de nombreux domaines, y compris l'économie, la politique, la sociologie, et plus récemment, l'interprétation des modèles de machine learning.

Les valeurs Shapley ont été introduites par le mathématicien Lloyd Shapley en 1953 dans le domaine de la théorie des jeux coopératifs. Lloyd Shapley était un éminent chercheur en économie et en mathématiques, lauréat du prix Nobel d'économie en 2012 pour ses contributions à la théorie des jeux.

L'idée fondamentale derrière les valeurs Shapley découle de la question de savoir comment attribuer équitablement les gains obtenus dans un jeu coopératif à chaque joueur (ou contributeur) du jeu. Dans un contexte où plusieurs joueurs collaborent pour obtenir un gain global, les valeurs Shapley fournissent une méthode pour attribuer de manière équitable la contribution de chaque joueur à la valeur totale du jeu.

Les valeurs Shapley sont basées sur le concept d'ajouts marginaux. L'idée est que la contribution d'un joueur à un jeu dépend de la valeur qu'il ajoute en rejoignant n'importe quel sous-groupe de joueurs. Les valeurs Shapley quantifient l'apport moyen attendu d'un joueur lorsqu'il joue avec tous les sous-groupes possibles de joueurs.

Les valeurs Shapley sont largement utilisées dans la théorie des jeux pour résoudre divers problèmes, notamment la répartition équitable des coûts, la formation de coalitions et la prise de décision collective. Au fil des années, ces valeurs se sont révélées extrêmement utiles et ont été appliquées à de nombreux

domaines, y compris l'économie, la politique, la sociologie, et plus récemment, l'interprétation des modèles de machine learning.

Dans le contexte de l'interprétation des modèles de machine learning, les valeurs Shapley ont été adaptées pour attribuer l'importance des caractéristiques à leurs contributions aux prédictions du modèle, fournissant ainsi une explication cohérente et équitable de la manière dont chaque caractéristique influence les résultats du modèle. Cette adaptation est à la base de la méthode SHAP (SHapley Additive exPlanations) qui est largement utilisée pour l'interprétation des modèles de machine learning.

Les valeurs Shapley sont calculées à l'aide d'une formule mathématique qui repose sur les concepts de la théorie des jeux coopératifs. Dans le contexte de l'interprétation des modèles de machine learning, les valeurs Shapley sont utilisées pour attribuer l'importance des caractéristiques aux prédictions du modèle. Voici la formule générale pour calculer les valeurs Shapley pour une caractéristique donnée :

$$\phi_i = \sum_{S \subseteq N\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}[v(S \cup \{i\})\_v(S)], \text{ où :}$$

- ϕ_i est la valeur de Shapley attribuée à la caractéristique i.
- N est l'ensemble de toutes les caractéristiques
- S est un sous ensemble de N qui ne contient pas la caractéristique i.
- $|N|$ est la taille totale de l'ensemble des caractéristiques.
- $|S|$ est la taille totale du sous-ensemble S.
- $v(S \cup \{i\})$ est la valeur de performance (par exemple la prédiction du modèle) lorsque les caractéristiques dans S son présentes.

La formule exprime la contribution marginale attendue de la caractéristique i. Les facteurs binomiaux et les termes factoriels garantissent que chaque combinaison possible de caractéristiques est prise en compte de manière équitable.

Il convient de noter que la valeur de Shapley peut varier en fonction de la méthode utilisée pour évaluer la performance $v(S \cup \{i\})$.

Dans le contexte de l'interprétation des modèles de machine Learning, des approximations et des méthodes spécifiques peuvent être utilisées pour rendre le calcul plus efficace et réalisable pour les modèles complexes.

Les limites du ratio de Shap

Bien que le SHAP (Shapley Additive explanations) soit une méthode puissante et largement utilisée pour expliquer les prédictions des modèles de machine Learning, il présente également certaines limites et considérations importantes. Voici quelques-unes des limites associées à la méthode SHAP :

- Calcul Computationnel Intensif : Le calcul précis des valeurs SHAP peut être intensif en termes de calcul, en particulier pour les modèles complexes tels que les réseaux de neurones profonds. Le temps de calcul peut augmenter considérablement avec la taille des données et la complexité du modèle.
- Approximations nécessaires : Pour rendre le calcul des valeurs SHAP plus réalisable, des approximations peuvent être utilisées. Ces approximations peuvent réduire la précision des valeurs SHAP dans certains cas, bien que des efforts aient été déployés pour minimiser cet impact.
- Interprétation Complexes : Les valeurs SHAP fournissent des explications individuelles pour chaque prédiction, ce qui peut rendre leur interprétation complexe, en particulier lorsque vous avez un grand nombre de caractéristiques.
- Définition du Baseline : Le choix de la référence (Baseline) à partir de laquelle les contributions sont mesurées peut influencer les valeurs SHAP et donc les interprétations. Il peut être difficile de déterminer la référence appropriée.

- Interactions de Caractéristiques Complexes : SHAP prend en compte les interactions entre les caractéristiques, mais la décomposition des interactions complexes peut être difficile à interpréter et à expliquer de manière concise.
- Dépendance au Modèle Expliqué : Les valeurs SHAP sont spécifiques au modèle expliqué. Si le modèle change ou évolue, les valeurs SHAP peuvent également changer.
- Redondance et Corrélation : SHAP attribue des valeurs à chaque caractéristique indépendamment, ce qui peut ne pas prendre en compte la redondance ou la corrélation entre les caractéristiques.
- Sensibilité aux Données : Les valeurs SHAP peuvent être sensibles aux variations des données d'entraînement, ce qui peut conduire à des interprétations différentes pour des données similaires.
- Taille de l'Ensemble Caractéristiques : Avec un grand nombre de caractéristiques, le nombre total de combinaisons possibles peut augmenter exponentiellement, rendant le calcul des valeurs SHAP plus complexe.
- Malgré ces limites, la méthode SHAP reste l'une des approches les plus avancées et les plus cohérentes pour expliquer les prédictions des modèles de machine Learning. Il est important de considérer ces limites et de les prendre en compte lors de l'interprétation des résultats SHAP.

3.7.6.2 Par Features importance

Il existe différents calculs d'importance de variable (features importance), en particulier pour les modèles de forêt aléatoire (random forest) et xgboost, pour lesquels les librairies Python sklearn et xgboost (respectivement) proposent de base plusieurs calculs.

La mise en pratique de cette méthode est présentée dans la section dédiée à la modélisation de la fréquence LightGBM (4.5.5.1.2).

3.8 Construction de la prime pure

L'assureur dispose de deux bases historiques :

- Base de sinistres qui contient le nombre de sinistres et leurs montants individuels.
- Base de souscription qui liste les facteurs de risques à savoir les caractéristiques des individus assurés.

Tout au long de notre explication, nous noterons ;

- $X = (X_1, X_2, \ldots, X_m)$: le vecteur aléatoire contenant les facteurs de risque (ou variables explicatives du modèle). Il s'agit des observations disponibles caractérisant l'assuré numéro i, qui sera désigné par la variable $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$
- n : la variable déterministe entière qui indique pour chaque assuré le nombre d'année de souscription du contrat.
- CRD_{n_j}: variable déterministe strictement positive qui indique le montant du capital restant dû à chaque année j.
- $T = \begin{cases} 0 \text{ si l'individu numéro } i \text{ n'a pas subi de sinistre} \\ 1 \text{ si l'individu numéro } i \text{ a subi un sinistre} \end{cases}$ la variable booléenne indiquant la sinistralité.
- e_i la variable calculant l'exposition de l'individu numéro i au cours d'une année j, nous avons alors $e_i = \frac{\text{Nombre de jours où le contrat est couvert}}{365.25}$.
- $\lambda_i = \frac{T_i}{e_i}$ la variable indiquant le taux de hasard de l'individu i durant l'année j. Il s'agit de la variable réponse du modèle.

L'objectif de l'assureur est d'estimer λ_i à partir des caractéristiques individuelles X, donc des variables connues a priori. L'assureur cherche particulièrement à construire un prédicteur Φ définie par

$$\Phi(x) := E[\hat{\lambda}_i | x].$$

Ce prédicteur (ou tarifeur) est la fonction qui, à partir des caractéristiques de l'assuré, lui associe son taux de risque instantané.

Le calcul de ce taux de hasard nous servira finalement à calculer l'engagement de l'assureur vis-à-vis de chacun de ses adhérents, à savoir le risque pur encouru par celui-ci via cette formule :

$$R(x_i)_{contrepartie(i)} = \sum_{j=1}^{n} CRD(j)\lambda_{ij}(x_i)$$

3.9 Les limites de la tarification individuelle

Afin de mesurer le risque décès ou le risque arrêt maladie inhérent à un dossier, il est dans l'intérêt de l'assureur de détecter et de connaitre la manière dont les différents risques modélisés interagissent entre eux.

En effet, il est évident que le risque global est différent lorsque la survenance d'un risque rend impossible la survenance d'un autre, ou quand la survenance du premier risque entraîne mécaniquement la survenance du second.

C'est dans ce cadre que nous entreprenons l'étude sur la dépendance entre les risques au sein d'un couple (risque décès ou risque arrêt maladie). Nous nous intéresserons à l'effet qu'aurait une structure de dépendance de ses risques sur la prime pure.

3.10 Mesure de la dépendance / les Copules

3.10.1 Définition

Le mot copule vient du latin « Copula » et signifie lien. La copule est un outil mathématique utilisé pour apprécier la dépendance entre deux variables aléatoires.

La copule est une écriture « uniformisée », c'est à dire indépendante des marges. Elle est d'une importance cruciale en assurance, car ce qui nous intéresse dans l'étude des risques, ce sont principalement les dépendances au sein de nos assurés.

Mathématiquement, la copule est une distribution multivariée dont les marginales sont uniformes.

Nous devons la notion de copule au mathématicien Abé Sklar, en 1969. Ce dernier s'est servi des travaux du mathématicien français Maurice Fréchet et de ceux du mathématicien autrichien Karl Menger (sur les espaces métriques aléatoires qui sont une généralisation de l'espace métrique usuel introduit par Maurice Fréchet).

La copule propose une nouvelle approche quant à la manière de considérer les comportements des risques d'individus liés au sein d'un ménage, et les mesures de corrélation qui en émanent. Devenu aujourd'hui un outil standard d'étude des dépendances, la copule s'applique déjà à l'assurance de de marché et permet de remédier aux problèmes de normalité des méthodes statistiques traditionnelles.

C'est suite à ce succès que nous tenterons de l'appliquer à l'assurance des emprunteurs afin d'améliorer les outils fondés sur l'indépendance de ce type de risque au sein d'un couple emprunteur.

3.10.2 Contexte mathématique

On s'intéresse à un vecteur aléatoire $X = (X_1, \dots, X_n)$, avec n un entier naturel.

L'objectif est de chercher à connaître la loi du vecteur $(X_1, \dots, X_n)$, dite **loi jointe**.

Une copule est donc la fonction de répartition d'un vecteur aléatoire, dont les marginales sont des variables aléatoires uniformes sur [0,1].

On a donc pour une copule C :

$$C(t_1, \dots, t_n) = \mathbb{P}(U_1 \leq t_1, \dots U_n \leq t_n)\,(t_1, \dots, t_n) \in [0,1]$$

Si X est un vecteur aléatoire de $\mathbb{R}^n$, alors nous avons

$$C(t_1, \dots, t_n) = C(F(X_1) \dots F(X_n))$$

3.10.2.1 Théorème de Sklar

Soit X un vecteur aléatoire, de loi F et de lois marginales $F_1, \dots, F_n$. Alors, il existe au moins une copule C, telle que :

$$F(t_1, \dots, t_n) = C(F(t_1) \dots F(t_n))$$

De plus, C est unique dès que les lois marginales F_i sont absolument continues.

Dans ce cas, C est la copule associée au vecteur X.

3.10.2.2 Définition de la densité d'une copule

Si elle existe, la densité d'une copule C se définit par : $\mathrm{C}(t_1, \dots, t_n) = \frac{\partial C^n}{\partial t_1, \dots, \partial t_n}(t_1, \dots, t_n)$

On en déduit la densité f de la variable aléatoire X :

$$f(t_1, \dots, t_n) = c\,(F_1\,(t_1), \dots, F_n\,(t_n)) \prod f_1\,(t_1) \dots f_n\,(t_n)$$

3.10.2.3 Propriétés des copules

Une fonction $C : [0,1]2 \rightarrow [0,1]$ est une copule si et seulement si :

$\rightarrow \forall u \in [0,1],\ C(u,0) = C(0,u) = 0$

$\rightarrow \forall u \in [0,1],\ C(u,1) = c(1,u) = u$

$\rightarrow C$ est une fonction 2-croissante, c'est-à-dire :

$\forall\ 0 \leq u_1 \leq v_1\ et\ 0 \leq u_2 \leq v_2,\ C(v_1,v_2) - C(v_1,u_2) + C(u_1,u_2) \geq 0$

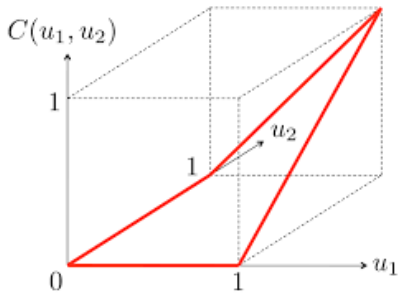


Figure 19 : Copule

On définit deux copules particulières

- La copule antimonotone :

$$C^-(u_1, u_2) = \max(u_1 + u_2 - 1,\ 0) = (u_1 + u_2 - 1)_+$$

- La copule comonotone (généralisable en dimension n) :

$$C^+(u_1, u_2) = \min(u_1, u_2)$$

On peut montrer que toute copule C est vérifié : $C^- \leq C \leq C^+$ (ce qui est un encadrement assez précis)

3.10.2.4 Quelques exemples de Copules

Copule de Gauss

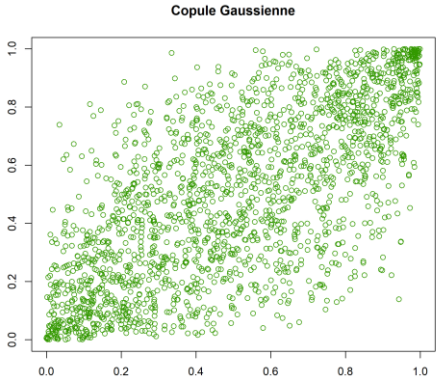


Figure 20 : Graphe de la copule de Gauss

Soit Σ une matrice symétrique définie positive.

La copule gaussienne est

$$C(t1, \ldots, tn) = \Phi\Sigma(\Phi^{-1}(t1), \ldots, \Phi^{-1}(tn))$$

Avec Φ et $\Phi\Sigma$ les fonctions de répartition des lois normales $(0,1)$ et $(0, \Sigma)$ (en dimension n).

Remarques :

- Paramètres = corrélations (beaucoup de paramètres…)
- Cas limites corrélations = -1, 0, 1
- Hors des cas -1 et 1, pas de dépendance de queue.

Copule de Student

Soit $r \in [-1, +1]$, et $\nu \geq 2$, alors la copule de Student de paramètres (r, ν) est définie par sa loi C comme :

$$C(t_1, t_2) = \int_{-\infty}^{T_\nu^{-1}(u_1)} \int_{-\infty}^{T_\nu^{-1}(u_2)} \frac{1}{\pi\nu\sqrt{1-r^2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{{t_1}^2 - 2rt_1t_2y + {t_2}^2}{\nu(1-r^{2)}}\right)^{\left(-\frac{\nu}{2}+1\right)} dt_1 dt_2$$

Où $T_\nu(T)$ est la fonction de répartition de la loi de Student univariée et ν le nombre de degrés de libertés.

Copules de Joe

La copule de Joe est efficiente lorsque nous souhaitons modéliser une dépendance de queue à droite. Cela stipule que les deux variables considérées ont une sorte corrélation positive. Le paramètre θ de la copule de Joe est définie sur $[1, +\infty[$.

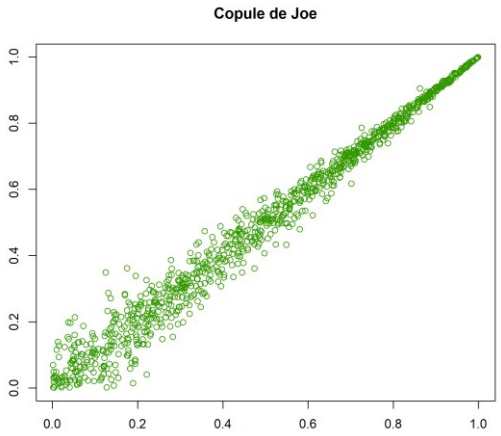


Figure 21 - Copule de Joe

$$C(t_1, t_2) = 1 - \left[(1-t_1)^\theta (1-t_2)^\theta\right]^{\frac{1}{\theta}}$$

Copule de Franck

La copule de Franck prend en compte un large panel de dépendance. Elle permet en effet d'approcher les bornes supérieure et inférieure de Fréchet-Hoeffding [127]. Par conséquent, cette copule est très utile lorsque nous avons des valeurs positives et négatives. Le paramètre de la copule θ est défini sur $]-\infty, +\infty[$.

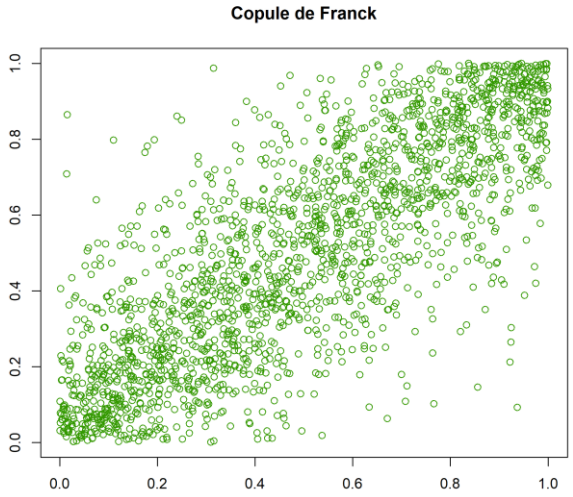


Figure 22 - Copule de Franck

$$C(t_1, t_2) = -\theta^{-1}\log\left\{1 + \frac{(e^{-\theta t_1}-1)(e^{-\theta t_2}-1)}{(e^{-\theta}-1)}\right\}$$

Copule de Joe-Franck (BB8 Copula)

La copule de Joe-Franck (BB8) Copula est un mixte entre la Copule de Joe et celle de Franck.

Elle combine ainsi leur propriété de dépendance à droite (Joe) et la forte dépendance de Franck (positive ou négative). La copule de Joe-Franck est définie par deux paramètres (θ, δ) qui sont respectivement caractérisés par :

- $\theta \leq 1$
- $0 < \delta < 1$

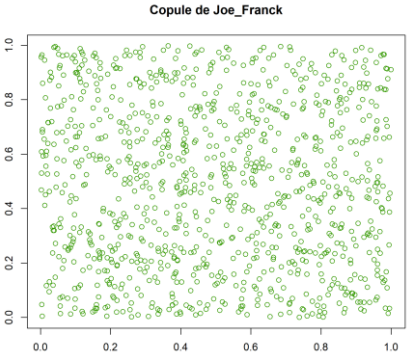


Figure 23 - Copule de Joe-Franck

$$C(t_1, t_2) = \frac{1}{\theta}\left(1 - \left[\left(1 - \frac{1}{1-(1-\delta)^{\theta}}\right)(1-(1-\delta t_1)^{\theta}\ ((1-1\delta t_2)^{\theta})\right]\right)^{\frac{1}{\delta}}$$

3.11 Risques compétitifs

3.11.1 Définition

Contrairement à la censure, le risque compétitif émane d'un évènement qui élude l'arrivé un autre. Ceci peut être le cas lorsqu'en biologie, nous étudions le décès lié à une maladie précise. Dans ce cas, le décès causé par d'autre maladies sans jamais n'avoir déclaré la maladie d'intérêt représente un risque compétitif. En effet, celui-ci nous a potentiellement privé d'observer cette maladie chez l'individu décédé suite à une autre cause.

Dans notre cas, lors de la souscription à un contrat d'assurance emprunteur, les assurés, sous réserve d'acceptation de l'organisme emprunteur, peuvent choisir leur taux de couverture. Lorsque l'emprunt est contracté à deux personnes, non seulement le taux de couverture répond à la même règle, mais chacun est aussi libre de choisir son taux de couverture individuel. Ainsi, lorsque le cout de couverture de l'emprunt dépasse 100%, nous observons l'apparitions de risques compétitifs partiels ou totaux.

Lors d'un emprunt à deux têtes, l'engagements annuel de l'assureur est noté :

$$R_{contrepartie(i)} = (quot_1 \mathbb{P}(DC_1) + quot_2\ \mathbb{P}(DC_2))\ \times CRD_i$$

Avec la variable $R_{contrepartie}$ qui fait référence au risque de défaut de remboursement de l'emprunt, et

La variable $quot_j$, la part d'assurance affectée à chaque co-emprunteur.

Cependant, lorsque $quot_1 + quot_2 > 100\%$, le risque de contrepartie va être dépendant de la survenance d'un événement garantie sur l'un des deux assurés, il y a donc risques compétitifs.

3.11.2 Risque compétitifs partiels

Lorsque $100\% < quot + quot < 200\%$, nous avons à faire à un risque compétitif partiel. En effet, ceci implique que la forme du contrat est un mixte entre une temporaire décès au premier décès, si c'est l'individu qui est couvert à 100% qui décède, et temporaire décès au second décès dont le taux de couverture est inférieur à 100% qui trépasse. Risque compétitifs totaux

3.11.3 Risques compétitifs totaux

Lorsque $quot_1 + quot_2 = 200\%$, nous observons alors, un risque compétitif total. En effet, ceci implique que la forme du contrat est celle d'une temporaire décès au premier décès. En effet, le décès de l'un deux conjoints entrainent la clôture du contrat et le remboursement du capital restant dû. Dans ce cas, le décès du second conjoint à une date ultérieure n'aura aucun impact sur le montant de la prime. Ce raisonnement s'applique autant à l'arrêt maladie qu'au décès. Lorsque nous considérons un couple, nous nous rendons compte que le système des quotités change la manière de considérer le risque.

Les quotités		Etat du contrat en cas de sinistre décès d'un des deux individus			
Individu 1	Individu 2	Ratio de couverture	Ratio d'indemnisation maximal	Etat du contrat	Remarque
50%	50%	100%	100%	Continuité	
80%	20%	100%	100%	Continuité	
100%	100%	200%	100%	Clôture du contrat	
80%	50%	130%	100%	Continuité	Sur couverture
20%	30%	50%	50%	Continuité	
30%	50%	80%	80%	Continuité	
100%	50%	150%	100%	Dépend de l'individu sinistré	

Figure 24 - Exemple de répartition des quantités au sein d'un couple

Supposons un couple d'assuré assurés respectivement à 60% et 100% du capital emprunté, qui est de 100.000 euros avec un taux d'intérêt nulle (pour simplifier les calculs) sur une durée de vingt ans.

Supposons d'autre part que la première tête décède au bout de dix ans. Le risque compétitif interviendra.

Dans ce cas, la projection faite lors de la souscription se base sur l'hypothèse suivante :

« *Le conjoint survivant continuera à payer 100% de la prime dont l'assiette de couverture est le capital restant dû avant le décès* »

Cependant, le capital restant dû après le décès, serait le capital restant dû avant le décès amoindri du remboursement de la partie couverte du défunt.

Par conséquent le conjoint survivant continuera de payer une prime indexée sur 50.000 euros (capital restant dû au bout de dix ans), alors qu'en réalité ceci est amoindri par le remboursement de 30000 euros par l'assurance (50.000 X 60%) suite au décès de la première tête.

Le risque du conjoint survivant sera de ce fait 20.000 euros.

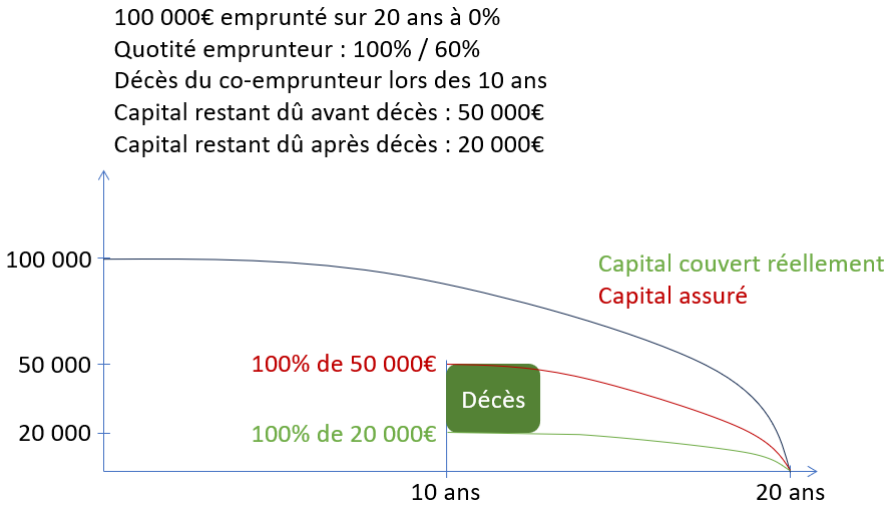


Figure 25 - Exemple de risques compétitifs

3.11.4 Mathématisation du problème

Le principe de risques compétitifs peut être vu comme un processus $(J_t)_{t>0}$ notant pour chaque date t l'état dans lequel se trouve un individu, $J_t \in \{0,1,2\}$. Un individu est initialement à l'état 0 et y demeure tant qu'aucun événement (d'intérêt ou compétitif) n'aient eu lieu.

L'individu peut :

- soit rester en 0 avec une probabilité $\mathbb{P}_{00}$.
- soit migrer à l'état 1 si l'événement d'intérêt se produit $\mathbb{P}_{01}$.
- soit migrer à l'état 2 si l'événement compétitif se produit $\mathbb{P}_{12}$.

Les états 1 et 2 sont absorbants ou transients en fonction des conditions du contrat.

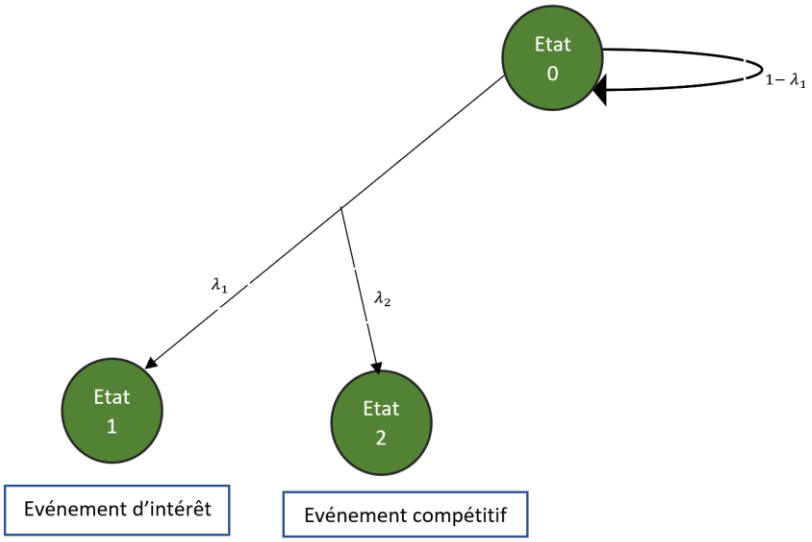


Figure 26 - Le processus J

Considérons maintenant, le temps d'arrêt :

$\boldsymbol{\tau = inf(t > 0 | J_t \neq 0)}$; $\boldsymbol{\tau}$ est le temps d'occurrence d'un évènement.

L'observation du processus $(J_t)_{t>0}$ peut être soumise à une censure à droite mais pas à une troncature à gauche, car l'échantillon observé prend en compte tous les individus depuis le début du contrat. Dans le cas d'une censure à droite $\boldsymbol{C}$, nous observons alors, la variable $\overline{\boldsymbol{T}} = \min(\boldsymbol{T}, \boldsymbol{C}), (J_t)_{t>0})$. Ainsi lorsque l'évènement est censuré, l'observation est arrêtée alors que l'individu est encore dans l'état initial 0 (ce qui sous-entend que l'évènement d'intérêt n'est pas encore eu lieu). Dans le cas contraire, le temps d'occurrence d'un évènement et le type d'évènement ont été observés.

Rappelons ici que la fonction de hasard représente la probabilité instantanée, que le contrat subisse un sinistre suite à la cause k, sachant qu'il était toujours en vigueur à la date t, cela se traduit par la formule suivante :

$$\lambda(t) = \frac{f(t)}{S(t)} = = \lim_{\Delta \to 0} \frac{\mathbb{P}[t \leq T \leq t+\Delta | T > t]}{\Delta}$$

Avec :

- Δ *aussi petit que l'on veut.*
- $f(t)$ *la densité de la variable aléatoire*
- $S(t)$ *la fonction de survie* ; $S(t) = \mathbb{P}[T > t]$

Nous pouvons maintenant définir la fonction de hasard spécifique telle que :

$$\lambda_k(t) = \frac{f(t)}{S(t)} = = \lim_{\Delta \to 0} \frac{\mathbb{P}[t \leq T \leq t+\Delta, J=k, | T > t]}{\Delta}$$

Le hasard spécifique est donc le taux de hasard causé par chacun des risques compétitifs de manière individuelle.

Dans cette équation, $\boldsymbol{J}$ est la variable aléatoire référente aux causes possibles de la sinistralité d'un contrat (le décès d'un des conjoints couverts à 100% par exemple). La fonction de hasard cumulée qui lui est associée est définie par :

$$\Lambda_k(t) = \int_0^t \lambda_k(s) ds$$

Elle traduit quant à elle, le hasard sur la durée totale de l'exposition émanant chacun des risques rivaux.

Le processus stochastique J, est donc entièrement déterminé par l'ensemble des fonctions de hasard.

Enfin, nous pouvons définit le taux de hasard toutes causes, $\lambda(t)$ qui correspond au taux de hasard induit par tous les risques compétitifs si nous ne faisons pas la distinction causale. Il est noté :

$$\lambda(t) = \sum_{k=0}^{n} \lambda_k(t)$$

Et la fonction de survie du processus J quant à elle, se déduit des fonctions précédentes par :

$$s(t) := \mathbb{p}[T > t] = exp^{-\int_0^t \sum_{k=1}^{n} \lambda_{k(s)} ds}$$

Avec :

- n le nombre de risques compétitifs
- t la durée du contrat

Dans notre étude, $s(t)$ représente la probabilité de survie d'un individu dans le risque décès.

C'est à partir de la théorie des risques compétitifs basée sur l'analyse de survie que l'on a développé des modèles afin d'estimer la sinistralité en décès et en incapacité de travail de notre portefeuille.

3.11.5 Modélisation

En utilisant l'algorithme de machine learning lightGBM, nous modéliserons le taux de hasard de chacun des individus constituant notre couple.

Cette approche consiste à modéliser l'indemnisation d'un contrat liée à une cause pacifique. Nous pouvons prendre l'exemple du décès d'un des deux conjoints seulement.

Cette technique nous permet de connaitre l'effet des variables au sein du couple sur le processus J. Afin de prendre en compte les risques compétitifs au sein d'un couple, et pouvoir déterminer une proportion d'individus sur un portefeuille subissant le risque d'intérêt. Il est donc nécessaire de modéliser toutes les fonctions de hasard spécifiques afin d'obtenir :

λ (t) et F(t) :

$$\begin{cases} \lambda^1(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(t < T_1 \leq t + \Delta t | T_1 > t) \\ \lambda^2(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(t < T_2 \leq t + \Delta t | T_2 > t) \end{cases}$$

Nous calculerons par la suite les taux de hasard cumulés pour chacun des partenaires :

$$\begin{cases} \Lambda_1(t) = \int_0^t \lambda_1(s) ds \\ \Lambda_2(t) = \int_0^t \lambda_2(s) ds \end{cases}$$

Nous déduirons enfin la fonction de survie du processus :

$$S(t) = \mathbb{P}[T > t] = \exp\left(-\int_0^t \sum_{k=1}^{n} \lambda_k(s) ds\right)$$

Le risque absolu représente la probabilité d'avoir subi l'évènement k entre les temps s et t et peut être estimé de manière non paramétrique par la formule d'Aalen-Johansen (Beyersmann et Scheike, 2014).

3.11.6 Les métriques : Le critère d'information d'Akaike

Le critère d'information d'Akaike, (AIC en anglais pour Akaike Information Criterion) mesure la qualité d'un modèle statistique. Ce critère proposé par Hirotugu Akaike en 1973, est calculé comme suit :

$$AIC = 2k - 2ln(L) \text{ où}$$

- k est le nombre de paramètres à estimer du modèle
- L est le maximum de la fonction de vraisemblance du modèle.

Lors de la modélisation d'un problème statistique, il est souvent préférable d'augmenter la vraisemblance du modèle en ajoutant un paramètre. Le critère d'information d'Akaike permet de pénaliser les modèles en fonction du nombre de paramètres afin de satisfaire le critère de parcimonie.

En considérant un ensemble de modèles candidats, le modèle choisi est celui qui aura la plus faible valeur d'AIC. En effet, l'AIC repose sur un compromis entre la qualité de l'ajustement et la complexité du modèle, en pénalisant les modèles ayant un grand nombre de paramètres, ce qui limite les effets de surajustement (plus on a de paramètre meilleur est la qualité du modèle).

En se basant sur la théorie de l'information, l'AIC propose une estimation de la perte d'information en fonction du choix du modèle. Il ne s'agit donc pas d'une réponse absolue quant à la qualité du modèle. Même si la valeur minimale est choisie, ceci ne garantit pas une bonne adéquation quant au modèle concerné.

3.11.7 L'adéquation du modèle

Afin d'estimer l'adéquation du modèles les métriques calculées sont :

3.11.7.1 La logloss

En statistique, la régression de Poisson est une forme d'analyse de régression utilisée pour modéliser les données de comptage et les tableaux de contingence. La régression de Poisson suppose que la variable de réponse Y a une distribution de Poisson, et que le logarithme de sa valeur probable peut être modélisé par une combinaison linéaire de paramètres calculés par suite à la régression. Un modèle de régression de Poisson est parfois appelé modèle log-linéaire, notamment lorsqu'il est utilisé pour modéliser des tableaux de contingence. Les modèles de régression de Poisson sont des modèles linéaires généralisés avec le logarithme comme fonction de lien (canonique), et la fonction de distribution de Poisson.

Le modèle de régression s'écrit comme suit :

Soit X $\in \mathbb{R}^n$, un vecteur de variables **i.i.d,** nous pouvons alors définir le modèle de régression sous la forme

$$\log\left(\mathbb{E}(Y|x)\right) = a'x + b$$

Où $a \in \mathbb{R}^n$ et $b \in \mathbb{R}$

Nous pouvons également le voir sous la forme concaténée $\log\left(\mathbb{E}(Y|x)\right) = \theta'x$

Nous déduisons donc que $\mathbb{E}(Y|x) = e^{\theta'x}$

Nous pouvons estimer les paramètres par maximum de vraisemblance, en effet :

Soit θ, un ensemble de paramètres et X un vecteur d'observations, alors nous pouvons calculer l'espérance de la distribution de Poisson théorique. En Effet, $\mathbb{E}(Y|x) = e^{\theta'x}$ d'autre part, la densité de probabilité donnée pour la fonction de Poisson est $\mathbb{P}(y|x;\theta) = \frac{e^{y(\theta'x)}e^{-\theta x}}{y_i!}$

Ainsi la vraisemblance et son logarithme sont comme déduits comme suit :

$$L(y_1, \dots, y_n|; \theta) = \prod_{i=1}^{n} \frac{e^{y_i(\theta' x_i)} e^{-\theta x_i}}{y_i!}$$

$$\ell\left(\theta|x_{1,\dots,}, x_{1,\dots,}; y_1, \dots, y_n\right) = \sum_{i=1}^{n} y_i(\theta' x_i) - \theta x_i - \log(y_i!)$$

Le but de cette partie est de trouver le paramètre θ qui maximise le logarithme de vraisemblance. Cela équivaut à minimiser la fonction $\sum_{i=1}^{n} y_i(\theta' x_i) - \theta x_i$.

Or, la résolution de cette équation ne peut pas se faire de manière analytique, nous pouvons toutefois minimiser la fonction - ℓ par descente de gradient étant donné qu'elle est convexe.

3.11.7.2 La déviance

La déviance comme son nom l'indique représente la différence entre la vraisemblance estimée comme étant la meilleure et le modèle saturé. Le modèle saturé est celui qui s'ajuste aux données de manière exacte.

La déviance est donc calculée comme suit $D_{\widehat{\mathcal{B}}} = -2(\ln(\mathcal{L}_{\widehat{\beta}}) - \ln(\mathcal{L}_s) = 2\left(\frac{\max(y_i,0)^{2-p}}{(1-p)(2-p)} - \frac{y_i \hat{\mu}_i^{1-p}}{1-p} + \frac{\hat{\mu}_i^{2-p}}{2-p}\right)$

Il s'agit donc d'un indicateur qui mesure l'adéquation du modèle. Plus sa valeur est faible, meilleure est l'ajustement.

3.11.7.3 La déviance expliquée

La déviance expliquée est ce qui permet de connaitre la part de déviance maitrisée par le modèle. Il est une généralisation du coefficient de détermination R^2utilisé dans les régressions linéaires.

Nous pouvons le calcule comme suit $\mathcal{D}_e = 1 - \frac{D_{\widehat{\mathcal{B}}}}{D_{\mathcal{B}_0}}$ avec $\mathcal{B}_0$ la déviance du modèle sans variables explicatives. Si le modèle choisi admet une adéquation parfaite alors cet indicateur vaudra 1. Il vaudra donc 0 dans le cas contraire.

4 MODELISATION DE LA FREQUENCE (LIGHTGBM, COPULE, COUPLE)

4.1 Introduction et matériel disponible

Nous avons accès à deux bases de données importantes qui concernent des bénéficiaires de l'assurance emprunteur et à leurs sinistres. Dans ce mémoire, nous analyserons seulement les risques de décès et le risque d'entrée en incapacité de travail.

Nous modéliserons ainsi le taux de décès instantané/le taux annuel d'entrée en incapacité qui correspond mathématiquement au taux de Hasard.

Comme la méthode de modélisation est la même pour ces deux risques, nous les désignerons par la suite en utilisant le terme **taux de Hasard** sans précisions supplémentaires.

L'ensemble des données ont été rassemblées sur une période d'observation de dix ans et sont disponibles dans les bases suivantes :

- **Une base contenant les informations relatives aux adhésions.**
  Cette base collecte les informations nécessaires à l'établissement d'un tarif, comme l'âge, le montant du prêt, la quotité couverte, l'ancienneté de l'assuré au sein de la boite, le poids, la taille, des information médicales …
- **Une base contenant les informations relatives aux sinistres**.
  Cette base consigne les renseignements relatifs au assurés sinistrés. Elle contient donc autant de lignes que de sinistres, et y enregistre leurs montants.

4.2 Traitement des données

4.2.1 Traitement des valeurs manquantes

Comme pour toutes les bases de données, on y observe des erreurs opérationnelles. Certaines valeurs sont manquantes comme :

4.2.1.1 La catégorie socio-professionnelle

La catégorie socio-professionnelle est une variable très importante dans la tarification du risque emprunteur. En effet, la santé des individus peut être sévèrement influencée par le métier exercé. Comme indiqué par cette étude, cette variable a une part de variance explicative non négligeable vis-à-vis de l'espérance de vie d'une personne.

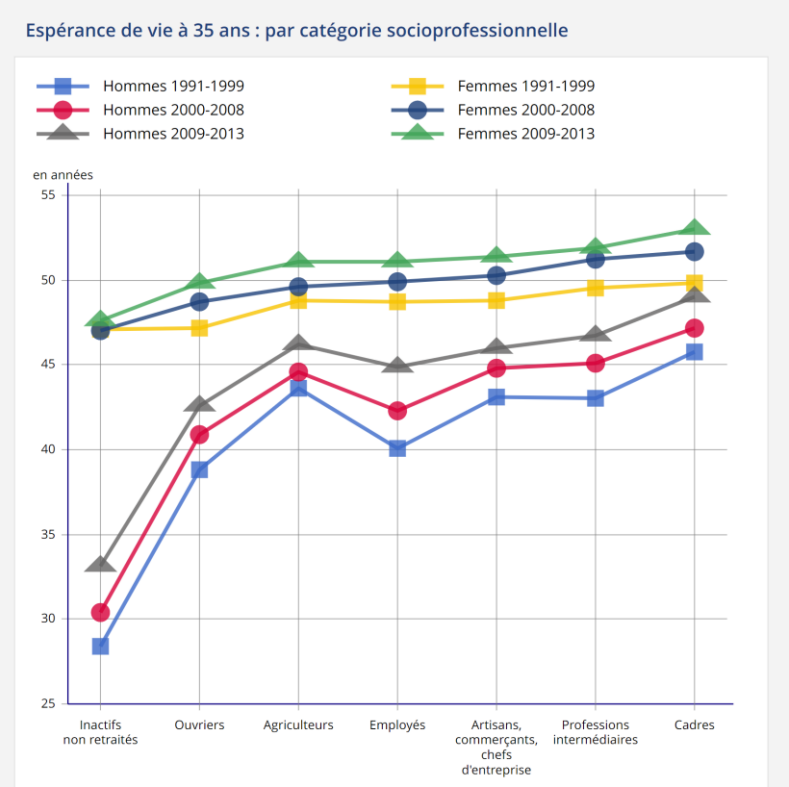


Figure 27 - Espérance de vie

- Lecture : en 2009-2013, l'espérance de vie à 35 ans des hommes cadres est de 49,0 ans, soit 6,4 ans de plus que celle des hommes ouvriers.
- Champ : France métropolitaine.
- Source : Insee, échantillon démographique permanent.

Nous remarquons par cette étude de l'INSEE que la catégorie socio-professionnelle des individus constitue une information très significative dans l'estimation du risque de mortalité. De la même manière, le risque incapacité de travail est très corrélé au métier de la personne étudiée.

Ce graphique nous donne en outre une information cruciale sur la dépendance de l'espérance de vie par rapport à l'année d'observation. Nous pouvons donc conclure que la probabilité de décès d'un individu dépendant non seulement de sa catégorie socio-professionnelle, de son âge mais aussi de la date des observations.

4.2.1.2 Le département

D'après les chiffres de l'INSEE, nous pouvons voir sur ce graphique que le taux de mortalité n'est pas le même en fonction de la résidence de l'assuré. En effet, ce graphique suggère qu'il existe une tendance de la sinistralité décès en fonction du département de résidence de l'assuré.

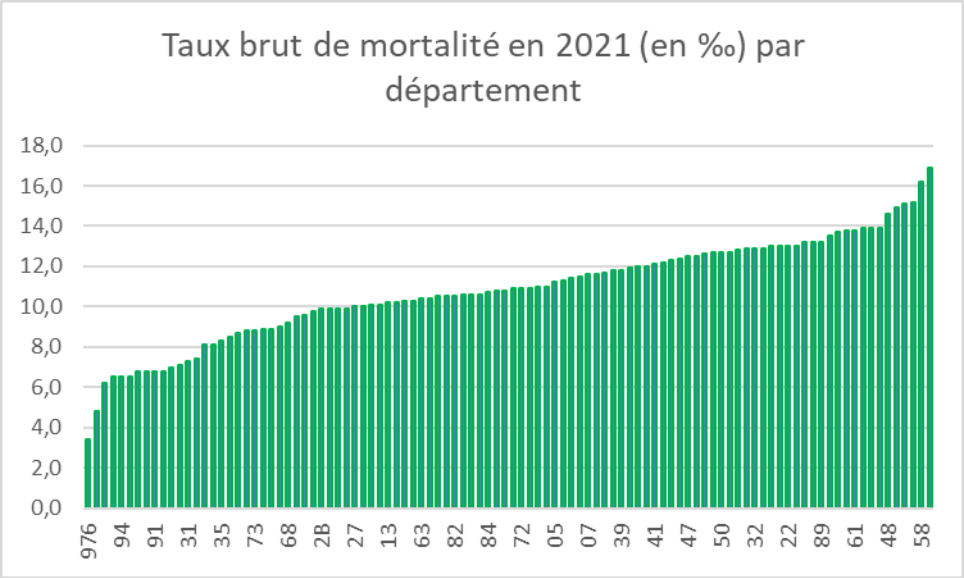


Figure 28 - Taux brut de mortalité en 2021 par département

Comme nous pouvons le constater sur ce graphique dont les données proviennent de l'INSEE, le taux de mortalité de la population française est très variable en fonction du département. Nous pouvons ainsi remarquer que la probabilité de décès en vivant dans la Creuse est 5 fois plus importante que la probabilité de décès à Mayotte. Des recherches complémentaires ont montré que la différence de l'âge moyen par département explique en partie cette variation dans le taux de décès mais n'explique pas tout. En effet, C'est la conclusion d'une étude publiée en 2019 par l'Association des Maires ruraux de France (AMRF) et la Macif. Il en résulte qu'un homme vivrait 2,2 ans de moins en zone rurale que dans les grandes villes.

De ce fait, lorsque cette variable est manquante, elle est donc corrigée à partir du code postal afin de ne pas biaiser la tendance.

4.2.2 La mensualité

Cette variable est incomplète et présente 30% de valeurs inconnu. Un recalcul a donc été effectué sous l'hypothèse que les prêts soient remboursés mensuellement avec une mensualité constante.

4.2.3 La variable cible

Suite à l'ajout de ces compléments d'informations, nous avons par la suite subdivisé la base des conjoints en années afin de calculer le taux de hasard de chaque tête sur une période d'un an. On y trouvera ainsi autant de lignes que d'années durant lesquelles l'assuré est observable dans la base. Cette architecture présente des bienfaits car elle permet de lier les informations de chaque assuré à celles de de son conjoint.

ID_DOSS	ID_IND	Age_Step	CD_Sexe	CSP	Form_Med
1	1	36	F	1	1
1	2	42	H	2	1
2	3	56	H	3	0
2	4	31	F	4	0
3	5	23	F	4	3
3	6	18	F	2	3

ID_DOSS	ID_IND	Age_Step	CD_Sexe	CSP	Form_Med	period
1	1	36	F	1	1	1
1	1	37	F	1	1	2
1	1	38	F	1	1	3
1	2	42	H	2	1	1
1	2	43	H	2	1	2
1	2	44	H	2	1	3
2	3	56	H	3	0	1
2	3	57	H	3	0	2
2	3	58	H	3	0	3
2	3	59	H	3	0	4
2	4	31	F	4	0	1
2	4	32	F	4	0	2
2	4	33	F	4	0	3
2	4	34	F	4	0	4
3	5	23	F	4	3	1
3	6	18	F	2	3	1

Figure 29 - Calcul du taux de hasard

4.2.4 Ajout de variables calculées

Pour calculer le taux de décès instantané, nous avons besoin de connaitre l'historique de la sinistralité pour chaque temps d'exposition Δ_t. C'est pour cela que la base est transformée. Nous devons donc voir l'état de chaque individu de manière annuelle. Cela se traduit dans la base par la répétition de chaque individu par autant d'année qu'il reste dans le portefeuille.

ID_DOSS	ID_IND	Age_Step	CD_Sexe	CSP	Form_Med	period
1	1	36	F	1	1	1
1	1	37	F	1	1	2
1	1	38	F	1	1	3
1	2	42	H	2	1	1
1	2	43	H	2	1	2
1	2	44	H	2	1	3
2	3	56	H	3	0	1
2	3	57	H	3	0	2
2	3	58	H	3	0	3
2	3	59	H	3	0	4
2	4	31	F	4	0	1
2	4	32	F	4	0	2
2	4	33	F	4	0	3
2	4	34	F	4	0	4
3	5	23	F	4	3	1
3	6	18	F	2	3	1

Figure 30 - Transformation base par décomposition par période

Pour se faire, des lignes mais aussi des variables sont ajoutées, à savoir les variables :

- Exposition (c'est le temps durant lequel l'individu est exposé par année)
- Décès (booléen qui indique si le décès a eu lieu ou non)
- Age_step : Il s'agit d'une variable qui calcule l'âge de l'individu à chaque étape de l'observation.
- Seniority : cette variable nous permet de connaitre et d'utiliser l'ancienneté de chaque individu au sein du portefeuille.
- CSR : elle indique le capital sous risque.
- Mensualité : qui indique le coût mensuel du remboursement du crédit.
- Target : cette variable correspond au taux de hasard qui est défini pat : $Target = \frac{décès}{exposition}$

C'est cette dernière variable qui sera estimée dans nos modèles.

4.2.5 Base finale de la modélisation individuelle

Suite au nettoyage des données, aux calculs que nous avons effectués et à la sélection des variables explicatives, nous obtenons une base classique qui constitue l'entrée de notre modèle. Nous avons d'autre part complété les modalités manquantes et sélectionné les variables nécessaires à la tarification.

Les variables explicatives sont donc :

Age_step : L'âge de l'individu à chaque période d'observation. L'utilisation de cette variable nous apportera plus de précision que l'âge à la souscription. En effet, connaitre l'âge du décès est plus pertinent que de connaitre l'âge à la souscription, vu que cette information est modulée par la durée du prêt. L'âge step nous donnera quant à lui l'ampleur du risque de manière directe.

K_ass : La tranche du capital enrichit des autres prêts et des montants d'assurance déjà payé par tête.

CD_sexe : Le sexe de l'individu considéré.

Mensualite : La mensualité du prêt bancaire.

Top_fumeur : Un booléen qui indique si l'individu et fumeur ou non.

Dur_pret : La durée Durant laquelle court le prêt.

CSP : C'est la catégorie socio-professionnelle.

Departement: Le département dans lequel vit l''assuré.

Form_med: Le type de questionnaire médical considéré.

4.2.6 Quelques chiffres sur la base de données

Après retraitements des données, la base de données qui servira pour notre étude contient 100.000 couples couverts par ce produit d'assurance emprunteur. Parmi ces individus, 1300 cas de décès sont observés sur une période de 12 ans. Cela suggère un taux moyen de décès de 0.6% pour les femmes et 0.7% pour les hommes. La dispersion par sexe suggère 51% d'hommes et 49% de femmes. La base est donc équilibrée selon ce segment, cependant elle ne l'est moins quant à la dichotomie en fonction de la survie ou du décès. Cette information est cruciale car il n'est pas chose aisée de mettre en place une adéquation sur la prédiction d'une valeur cible déséquilibrée.

4.3 Création des échantillons d'entrainement et de test

Lors de la modélisation individuelle, la création des échantillons d'entrainement et de test n'a pas été faite de manière conventionnelle. En effet, la structure compliquée a conduit à un découpage sans utiliser d'aléas.

En effet, les individus se répètent autant de fois que le nombre d'années de leur présence dans le contrat. Et un découpage aléatoire aurait conduit à la présence de la même personne tantôt dans l'échantillon d'entrainement, tantôt dans l'échantillon test. Il a donc été choisi de faire un découpage forcé en alternant au sein des personnes dans la base de données.

Dans ce mémoire, ce découpage a été fait en fonction du numéro de dossier d'adhésion. Cela a donc permis de garder chaque couple au sein du même bloque (d'entrainement ou de test) tout en application une sélection aléatoire et en veillant à l'équilibre de la proportion de sinistre entre les deux échantillon (d'entrainement et de test).

La même méthode a été utilisée dans la cross-validation utilisée par le modèle LightGBM.

5 MISE EN PLACE DES TABLES D'INCIDENCE

L'étude brute des taux de mortalité étant faite, nous allons modéliser par la suite les taux de hasard en utilisant le modèle de Nelson Allen sur la population couple. Par la suite, nous utiliserons une méthode plus segmentée avec le modèle LightGBM appliqué à la même population avec prise en compte des variables du conjoint.

Pour finir nous proposerons une approche qui est fondée sur les risques compétitifs avec la prise afin d'estimer le risque réel encouru par l'assureur.

5.1.1 Modélisation des probabilités de décès par Nelson Allen

5.1.1.1 Décès

Nous commencerons par modéliser le taux de décès instantané en utilisant le modèle de Nelson Allen.

Pour cela nous utiliserons seulement la variable âge.

	Model_name	Poisson_loss	Explained_deviance	mean_poisson_deviance
0	NA_estimate_real_smooth	0.001985	0.126253	0.007360
1	NA_estimate_real	0.003265	-1.212211	0.018633

Figure 31 – Table des métriques

La table des métriques indique des chiffres acceptables en termes de fonction perte et de déviance.

Dans le premier modèle (NA_estimate_smooth), la modélisation est lissée par Whittaker et semble présenter de meilleures métriques. En effet, le lissage permet de diminuer la perte ainsi que la déviance mais il permet également d'améliorer l'adéquation (qui est de 0.12) qui nous permet de choisir Nelson-Allen face au modèle de base qui calculerait la fréquence moyenne par âge.

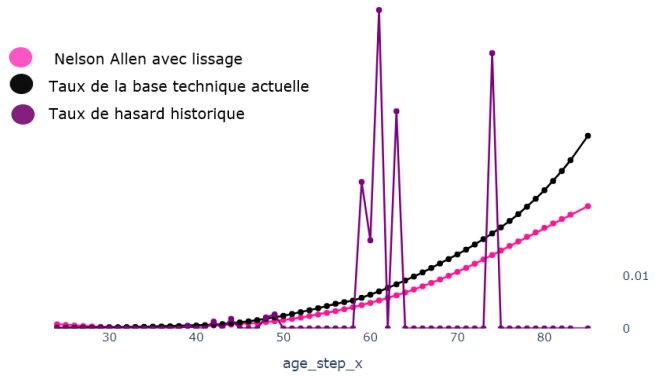


Figure 32 – Comparaison des taux de hasard

Nous obtenons le graphique suivant : en effet de 18 à 50 ans, le modèle de Nelson Allen semble en adéquation avec la sinistralité historique. Cette adéquation est également valable pour la sinistralité

contenue dans la base technique. Cependant, après 50 ans, la sinistralité historique semble très erratique, et nous observons que les taux d'incidence de la table technique s'éloignent un peu trop de la sinistralité historique. Le modèle de Nelson Allen semble un peu moins éloigné ce cette dernière. Cette modélisation est due aux sinistres extrêmes qui tirent la modélisation un peu vers le haut.

Nous aurions pu faire une modélisation sans les valeurs extrêmes mais la sinistralité étant parcimonieuse, nous perdrions les caractéristiques de ces personnes. D'autre part, nous avons besoin de connaitre le taux de décès instantané de chaque personne et enlever les grandes valeurs reviendrait à exécuter l'apprentissage de la modélisation uniquement sur les personnes qui décèderaient en fin d'année.

5.1.1.2 Arrêt de travail

	Model_name	Poisson_loss	Explained_deviance	mean_poisson_deviance
0	NA_estimate_real_smooth	0.007184	-0.004797	0.027146
1	NA_estimate_real	0.007207	-0.073149	0.028993

Figure 33 – Métriques arrêt de travail

La table des métriques indique des chiffres acceptables en termes de fonction perte et de déviance.

De plus, le premier modèle (NA_estimate_smooth) qui inclut le lissage par Whittaker semble présenter de meilleures métriques. En effet, il permet de diminuer la perte ainsi que la déviance.

Cependant, la métrique qui calcule l'explained variance qui nous permets d'estimer l'adéquation du modèle tête par tête est négative. Cela veut dire qu'il est moins bien que le modèle de base en termes d'adéquation. Ceci implique le modèle qui permet d'obtenir la moyenne par âge est meilleur.

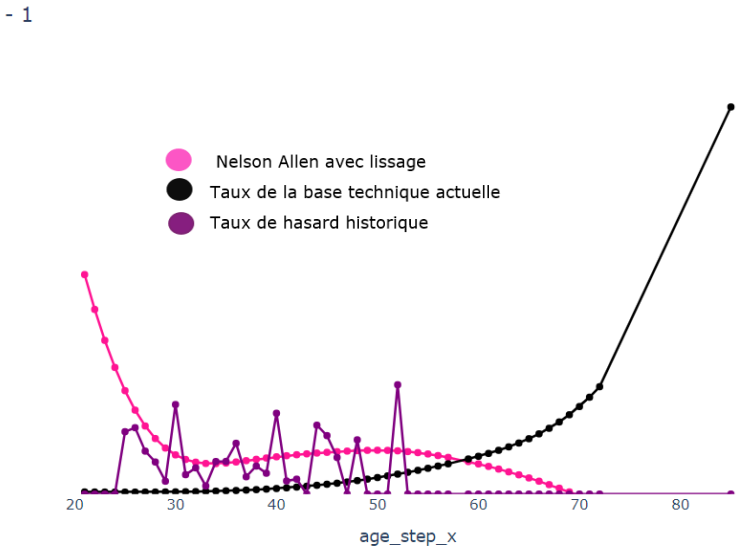


Figure 34 – Comparaison des taux de hasard

Pour la sinistralité arrêts de travail, le modèle de Nelson Allen semble très différent de la loi utilisée dans la base technique. La modélisation par Nelson Allen, semble plus en adéquation avec la sinistralité historique. De plus, nous observons qu'à partir de 60 ans, la sinistralité en arrêt de travail semble diminuer avec Nelson Allen ainsi qu'avec la sinistralité historique, ce qui est entièrement réaliste avec l'âge de départ à la retraite. En effet, peu d'effectifs restent dans la base après cet âge. Contrairement à cela, la base technique semble y indiquer une sinistralité plus importante.

5.2 Modélisation LightGBM avec le lien couple

5.2.1 Collecte d'informations par dossier

Les bases de données utilisées ont un format habituel, c'est-à-dire que chaque ligne correspond à un individu que nous avons segmenté en observations annuelles comme expliqué dans le chapitre **4.2.4**. Dans ce mémoire, il s'agit d'estimer la sinistralité relative à un couple. C'est une vision un peu différente par rapport à celle dans laquelle nous traitons les informations liées à un individu. Ici au contraire, il s'agirait plutôt de modéliser le risque par tête tout en prenant en compte le lien au sein d'un couple.

Pour ajouter les informations relatives au couple, nous avons mis en place une nouvelle base de données sur deux têtes qui collecte les informations nécessaires à notre analyse et les présente sur chaque ligne.

A partir de la première base dans laquelle sont recensées les observations, nous avons lié les personnes grâce à une clef unique par dossier.

Cette nouvelle base permettra donc de modéliser le taux de hasard par tête tout en prenant en compte les variables explicatives du conjoint.

ID_DOSS	ID_IND	Age_Step	CD_Sexe	CSP	Form_Me	period
1	1	36	F	1	1	1
1	1	37	F	1	1	2
1	1	38	F	1	1	3
1	2	42	H	2	1	1
1	2	43	H	2	1	2
1	2	44	H	2	1	3
2	3	56	H	3	0	1
2	3	57	H	3	0	2
2	3	58	H	3	0	3
2	3	59	H	3	0	4
2	4	31	F	4	0	1
2	4	32	F	4	0	2
2	4	33	F	4	0	3
2	4	34	F	4	0	4
3	5	23	F	4	3	1
3	6	18	F	2	3	1

ID_DOS	ID_IND	Age_Step	CD_Sexe	CSP_	Form_Med	CD_Fum	ID_IND_	Age_Step	CD_Sexe	CSP_1	Form_Med	CD_Fum	Period
1	1	36	F	1	1	N	2	42	H	1	1	O	1
1	1	37	F	1	1	N	2	43	H	1	1	O	2
1	1	38	F	1	1	N	2	44	H	1	1	O	3
1	2	42	H	2	1	O	1	36	F	3	1	N	1
1	2	43	H	2	1	O	1	37	F	3	1	N	2
1	2	44	H	2	1	O	1	38	F	3	1	N	3
2	3	56	H	3	0	O	4	31	F	3	0	O	1
2	3	57	H	3	0	O	4	32	F	3	0	O	2
2	3	58	H	3	0	O	4	33	F	3	0	O	3
2	3	59	H	3	0	O	4	34	F	3	0	O	4
2	4	31	F	4	0	O	3	56	H	3	0	O	1
2	4	32	F	4	0	O	3	57	H	3	0	O	2
2	4	33	F	4	0	O	3	58	H	3	0	O	3
2	4	34	F	4	0	O	3	59	H	3	0	O	4
3	5	23	F	4	3	N	6	18	F	4	3	N	1
3	6	18	F	2	3	N	5	23	F	1	3	N	1

Figure 35 – Ajout de données pour préparation de la modélisation

5.2.2 Paramètres lightGBM

Lors de l'apprentissage, nous avons utilisé une validation croisée afin d'obtenir une modélisation plus robuste. L'outil utilisé est RandomizedSearchCV de scikit learn. Il s'agit d'une méthode qui permet une recherche aléatoire d'hyper paramètre dans le cadre de l'optimisation des modèles d'apprentissage automatique. Elle est semblable à GeidSearch mais effectue une recherche des hyper paramètres de manière aléatoire (elle ne parcourt donc pas toute la grille des possibles) ce qui permet un gain de temps considérable. Le choix s'est porté là-dessus à cause des limitations computationnelles.

Après l'entrainement et l'optimisation, les résultats des paramètres obtenus donnent satisfaction.

Entrainement sur 75% de la base	
objective	poisson
eval_metric	poisson
learning_rate	0.01
max_depth	4
subsample	0.9
subsample_freq	1
boosting_type	gbdt
colsample_bytree	0.5
colsample_bylevel	0.6
reg_alpha	0.01
reg_lambda	5
gamma	1
min_child_samples	50
feature_pre_filter	False
n_estimators	612

Entrainement sur 95% de la base	
objective	poisson
eval_metric	poisson
learning_rate	0.02
max_depth	5
subsample	0.5
subsample_freq	1
boosting_type	Gbdt
colsample_bytree	0.6
colsample_bylevel	1
reg_alpha	0.01
reg_lambda	0
gamma	0
min_child_samples	100
feature_pre_filter	False
n_estimators	232

Figure 36 - Paramètre de la modélisation $M_{c(x,y)}$ par LightGBM du risque décès

En effet, avec une profondeur d'arbre max_depth = 4, et un paramètre de régularisation reg_lambda = 5 (donc grand), nous pouvons dire que le modèle est robuste et régularisé.

Ce sont ces paramètres qui seront retenus pour la modélisation du risque décès avec ajout des variables du conjoint, soit le modèle $M_{c(x,y)}$.

Concernant le même modèle pour le risque arrêt de travail, nous en tirons les mêmes conclusions.

Entrainement sur 75% de la base	
objective	poisson
eval_metric	poisson
learning_rate	0.1
max_depth	4
subsample	0.3
subsample_freq	1
boosting_type	gbdt
colsample_bytree	0.5
colsample_bylevel	0.6
reg_alpha	1
reg_lambda	50
gamma	1
min_child_samples	100
feature_pre_filter	False

Figure 37 - Paramètre de la modélisation $M_{c(x,y)}$ par LightGBM du risque incapacité de travail

5.2.3 Résultat de l'adéquation pour le décès

Après l'entrainement du modèle LightGBM sur la base couple, nous obtenons les métriques suivantes :

	Model_name	Poisson_loss	Explained_deviance	mean_poisson_deviance
0	Predictions_LGBM	0.00187	0.18889	0.00346

Figure 38 - Schéma modélisation couples prédit sur les couples

Explained deviance

	Model_name	Poisson_loss	Explained_variance	mean_poisson_deviance
0	predictions_LGBM	0.003431	0.086875	0.006382

Figure 39 - Schéma modélisation individuelle prédit sur les couples

Nous remarquons que la modélisation en couple, nous permet d'avoir un gain considérable.

En effet, cette nouvelle modélisation diminue la fonction perte de 45% et améliore l'adéquation de 50%.

Nous pouvons également constater que le graphique lift-chart appliqué à l'échantillon test qui permet de rendre compte de la moyenne des erreurs, nous indique que le modèle LightGBM en couple est meilleur que le modèle Nelson Allen.

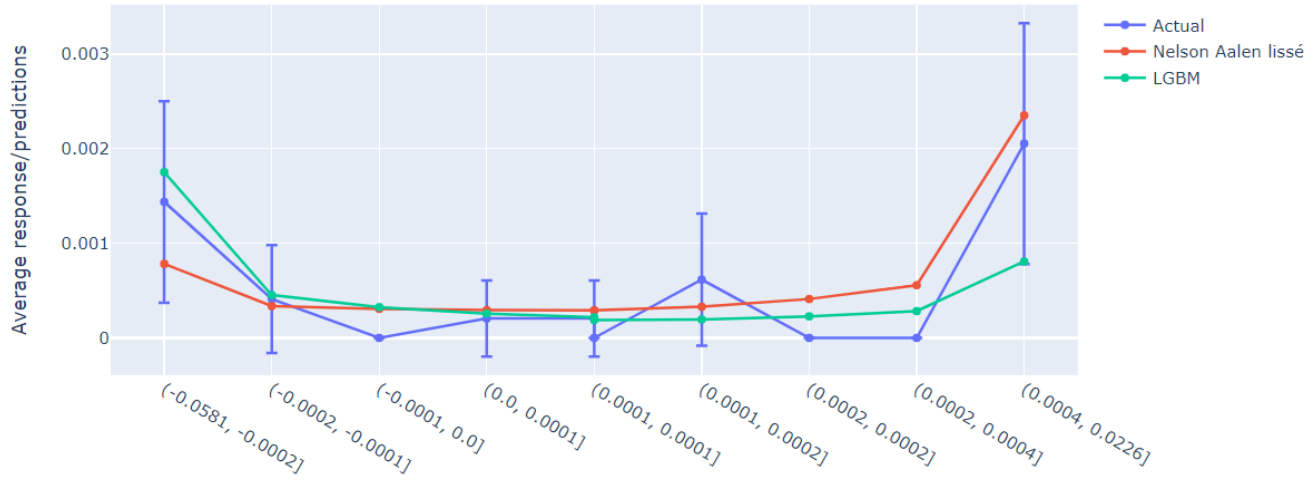


Figure 40 – Lift chart appliqué à l'échantillon de test

Nous remarquons que le modèle LightGBM décrit bien les valeurs empiriques. D'autre part, nous pouvons écarter tout soupçon de sur-apprentissage car ce graphique a été appliqué à l'échantillon test. Nous observons cependant une sous-estimation du modèle LightGBM sur la dernière classe d'erreur observée. Cependant l'exposition y est très faible. D'autre part, ceci peut être compensé par la surestimation du modèle dans la première classe et la neuvième.

Ce raisonnement est conforté par les valeurs des métriques utilisées à savoir :

	Model_name	Poisson_	Explained_deviance	mean_poisson_deviance
0	NA_estimate_real_smooth	0.001985	0.126253	0.007360
1	NA_estimate_real	0.003265	-1.212211	0.018633
2	Predictions_LGBM	0.001870	0.188890	0.003460

Figure 41 - Valeurs des métriques

Le premier modèle est l'application de Nelson Allen avec un lissage par la méthode de **whittaker-henderson**, le second est le modèle de Nelson Allen sans lissage, tandis que le dernier est le modèle LightGBM.

Nous remarquons que la Poisson loss (qui est la fonction perte) est plus petite pour le modèle LightGBM que pour les deux autres, quoique celle de LightGBM et celle de Nelson-Allen avec lissage soient très proches.

La métrique « explained-variance » fait référence au taux d'adéquation du modèle. Bien que les trois soient très faibles dans l'absolue, nous remarquons que LightGBM est meilleur.

Enfin, la mean-Poisson-deviance fait référence à la variance du modèle. Cette métrique témoigne également du fait que LightGBM reste plus adéquat.

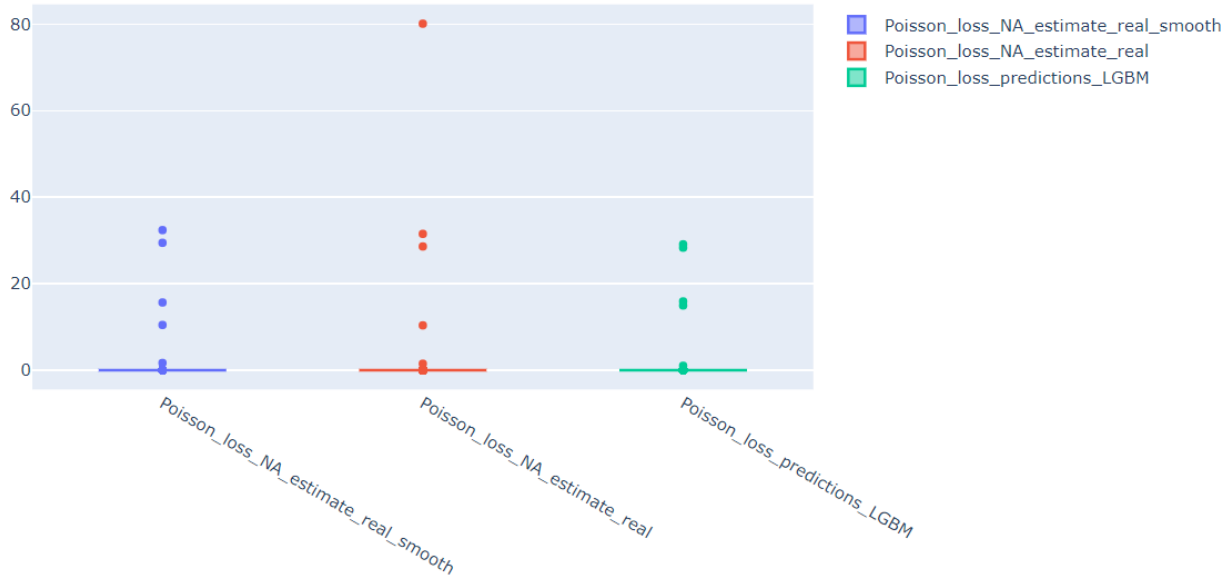


Figure 42 - Mean Poisson Deviance

Ce graphique donne la dispersion des erreurs en fonction des trois modèles. Ces résultats montrent indéniablement que les erreurs commises par la modélisation LightGBM sont beaucoup plus petites dans l'absolue, et la plupart d'entre elles sont basses.

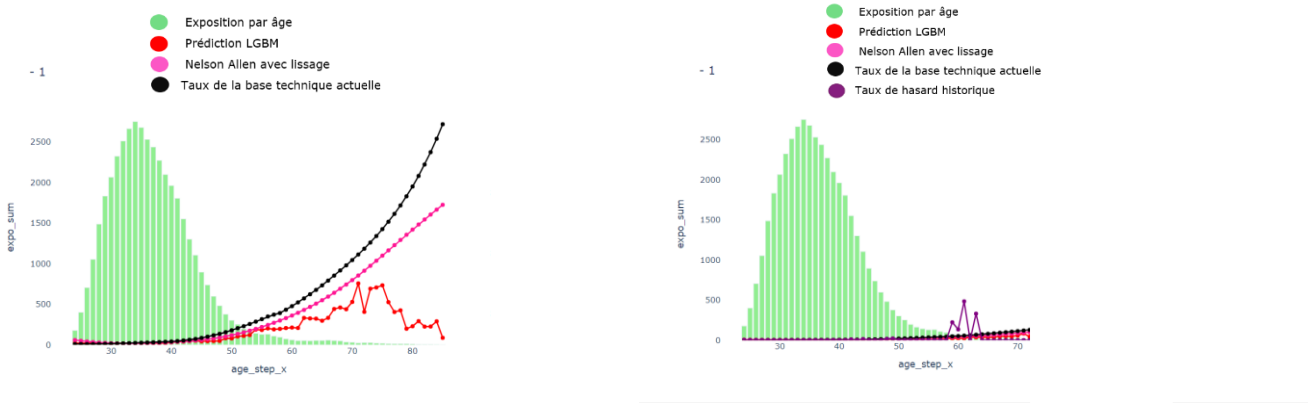


Figure 43 – Graphiques de dispersion des erreurs en fonction des trois modèles

Si nous comparons les prédictions du taux de décès instantané par rapport au taux de la base technique, nous remarquons que le modèle LightGBM est beaucoup plus optimiste. Ceci est une bonne chose car le graphique de droite montre que le modèle LightGBM est beaucoup plus proche de la sinistralité historique que le modèle utilisé actuellement.

Bien que notre modèle de prédilection semble erratique vers les grands âges, il reste plus réaliste. Si nous regardons l'exposition, nous déduisons que le modèle LightGBM est erratique seulement pour les classes où nous avons peu d'exposition, donc peu de monde (après 60 ans).

Pour cette partie, nous pouvons soit prendre la sinistralité historique, soit une combinaison entre le modèle de Whittaker et le modèle LightGBM, soit le modèle de Whittaker seul, mais cette dernière proposition peut aboutir à du sur-apprentissage.

5.2.4 Résultat de l'adéquation pour l'arrêt de travail

Les mêmes résultats sont obtenus pour la table d'incidence en arrêt de travail.

Le résultat du modèle effectué par tête sur la population entière fournit une meilleure adéquation mais propose une fonction perte de 0.021 et une variance de 0.036.

	Model_name	Poisson_loss	Deviance	mean_poisson_deviance
0	predictions_dummy	0.022234	-0.000143	0.038111
1	predictions_LGBM	0.021412	0.042970	0.036468

Figure 44 - Résultats modèle par tête sur population entière

Tandis que le modèle obtenu à partir de la population des couples fournit une adéquation plus faible qui est de 0.029 celui-ci diminue fortement la fonction coût qui passe de 0.021 à 0.014. Ce modèle diminue également la variance de la modélisation qui passe de 0.036 à 0.026.

	Model_name	Poisson_loss	Deviance	mean_poisson_deviance
0	predictions_LGBM	0.014323	0.029071	0.026233

Figure 45 - Résultats modèle par couple

Par conséquent, le modèle entrainé sur la population des couples est globalement meilleur car le gain en termes de fonction perte est de 33% mais cela se fait au détriment de la qualité de l'adéquation qui perd 32%.

Enfin, le modèle couple offre un gain de 28%, donc non négligeable en termes de variance du modèle. Il est donc plus stable que le modèle qui suppose l'indépendance des individus au sein du couple.

Avec le graphique double-lift-chart nous obtenons les résultats suivants :

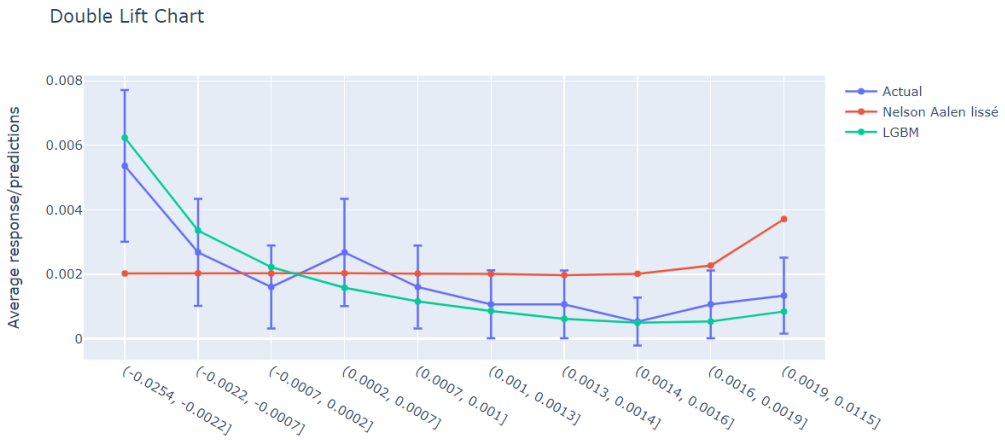


Figure 46 - Double Lift Chart

Le modèle LightGBM se démarque largement du modèle Nelson-Allen en offrant une meilleure adéquation par rapport à **l'Actual** qui est la sinistralité historique.

D'autre part, le calcul des métriques permet de confirmer ce graphique :

	Model_name	Poisson_loss	Déviance	mean_poisson_deviance
0	NA_estimate_real_smooth	0.007184	-0.004797	0.027146
1	NA_estimate_real	0.007207	-0.073149	0.028993
2	predictions_LGBM	0.007066	0.022481	0.026410

Figure 47 - Métriques du modèle

Nous avons une adéquation de 2% avec le modèle LightGBM contre une adéquation négative pour les modèles Nelson-Allen. Ceci implique que la modélisation avec Nelson-Allen est pire que le modèle de base qui consiste à prendre la moyenne des entrées en Arrêt de travail sur toute la base.

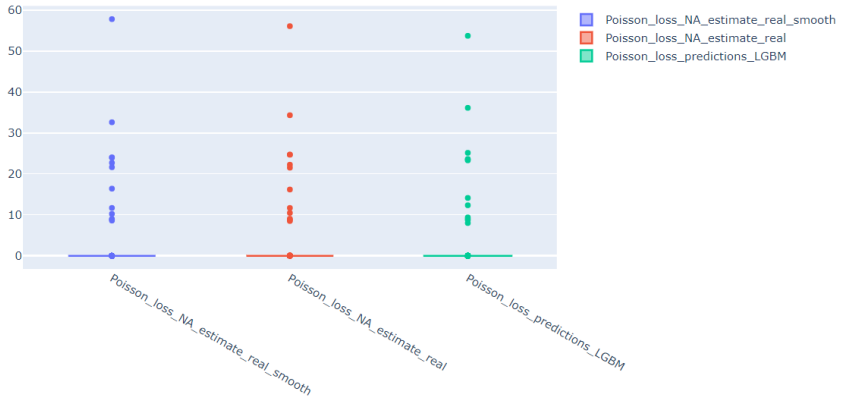


Figure 48 – Mean Poisson Deviance

Lorsque nous traçons le graphique des erreurs, nous remarquons que la plupart des erreurs commises par le modèle LightGBM présentent des valeurs plus basses que les deux autres modèles.

5.2.5 Autres métriques

5.2.5.1 Pour le taux de décès instantané

Dans cette partie, nous utilisons quelques métriques afin de connaitre la qualité des prédictions issues des trois modèles. Nous analyserons les prédictions de la modélisation effectuée sur la population en couple sans ajout des variables du conjoint et celles avec ajout des variables du conjoint. Enfin, nous étudierons les prédictions sur la même population issue de la modélisation individuelle.

Nous utiliserons pour cela trois modèles :

- M_i : le modèle entrainé sur toute la population du portefeuille
- $M_{c(x)}$: le modèle entrainé sur la population couple sans ajout des variables du conjoint de chaque tête.
- $M_{c(x,y)}$: le modèle entrainé sur la population couple avec ajout des variables du conjoint de chaque tête.

Nous étudierons également le comportement local et global des modèles afin de s'assurer de leur cohérence. Nous utiliserons pour cela les valeurs de Shapley ainsi que l'algorithme Feature importance et Feature dependance.

Comparaison des modèles M_i et $M_{c(x)}$

Nous constatons que l'adéquation du modèle M_i est de 9% tandis que celle de $M_{c(x)}$ est de 10%. Ceci signifie que la limitation de la modélisation à la population couple offre un gain de 11% en termes d'adéquation du modèle.

Lgbm-Model

metrics	On Task data (100%)
filter data...	
Deviance Explained	0.09074
RMSE	0.06514
MAE	0.00076
TotaPred/TotalObs	0.93291
Gini Index (Normalized)	0.56974
Average Deviance	0.00636
AveragePred	0.00037
MSE	0.00424

Figure 49 – Métriques de la modélisation totale tête par tête M_i

Lgbm-Model

metrics	On Task data (100%)
filter data...	
Deviance Explained	0.10761
RMSE	0.05279
MAE	0.00059
TotaPred/TotalObs	0.90080
Gini Index (Normalized)	0.64125
Average Deviance	0.00494
MSE	0.00279
AveragePred	0.00028

Figure 50 - Entrainement du modèle sur la population en coupe sans ajout des variables du conjoint $M_{c(x)}$

Les autres métriques montrent globalement que le modèle $M_{c(x)}$ est légèrement meilleur.

En regardant le tableau suivant, nous remarquons que les pertes avec la modélisation $M_{c(x)}$ sont moindres tandis que les métriques d'adéquation ou de robustesse comme la variance expliquée ou l'indice de Gini sont meilleurs.

Modèle	M_i	$M_{c(x)}$	Gain
Déviance expliquée	9%	10%	10%
RMSE	0%	5%	14%
MAE	0.0008	0,0006	25%
Tot_pred/Tot_Obs	93%	90%	-3%
MSE	0.0042	0,0029	-31%

Figure 51 -Comparaison des métriques d'adéquations et de robustesse

Il y a tout de même une légère baisse du taux d'adéquation global de 3%.

Le choix de la modélisation en couple ne semble pas améliorer les prédictions de manière nette mais ceci s'explique par une réduction drastique de la base d'entrainement Nous observons cependant un signal du fait que 80% des métriques sont améliorées.

Comparaison des modèles $M_{c(x)}$ et $M_{c(x,y)}$

Lgbm-Model

metrics	On Task data (100%)
filter data...	
Deviance Explained	0.18889
RMSE	0.04366
MAE	0.00095
TotaPred/TotalObs	0.94835
Gini Index (Normalized)	0.84977
AveragePred	0.00047
Average Deviance	0.00683
MSE	0.00191

Figure 52 - Métriques en couple avec ajout des variables du conjoint

Pour le modèle $M_{c(x,y)}$, nous obtenons un taux d'adéquation de 18% tandis que celui-ci est de 10% pour $M_{c(x)}$. Cela signifie que l'ajout des variables explicatives du conjoint dans la modélisation du taux de hasard pour le risque décès est très significatif. Cela démontre que modèle LightGBM permet de capter le lien entre la tête modélisée et son conjoint vis-à-vis du risque de décès encouru par l'assuré.

Si nous faisons un récapitulatif de la comparaison entre les deux modélisations, nous voyons que toutes les métriques s'accordent sur le fait que la modélisation en couple avec l'ajout des informations du conjoint donne une modélisation plus fidèle du risque encouru par l'assuré.

Modèle	$M_{c(x,y)}$	$M_{c(x)}$	Gain
Déviance expliquée	18%	10%	80%
RMSE	4,3%	5%	14%
MAE	0,000095	0,0006	25%
Tot_pred/Tot_Obs	95%	90%	6%
MSE	0,00191	0,0029	34%

Figure 53- Comparaisons des métriques entre $M_{c(x,y)}$ et $M_{c(x)}$

Comme expliqué dans la section 3.7.5, le MSE et le RMSE donnent l'erreur commise par la modélisation. Le modèle $M_{c(x,y)}$ permet de les diminuer afin d'avoir un gain de 34% et 14% respectivement. La MAE est l'erreur absolue commise par la modélisation. L'ajout des variables du conjoint permet de la diminuer de 25%.

L'amélioration de l'adéquation ainsi que la diminution de la perte sont clairement liées à la modélisation des têtes en prenant en compte les variables de leur conjoint.

Dans la suite de l'étude, nous comparerons les modèles $M_{c(x,y)}$ et M_i afin d'expliquer la différence entre le modèle actuel qui consiste à effectuer une modélisation tête par tête et celui qui consiste à limiter la modélisation à la population couple avec ajout des variables du conjoint.

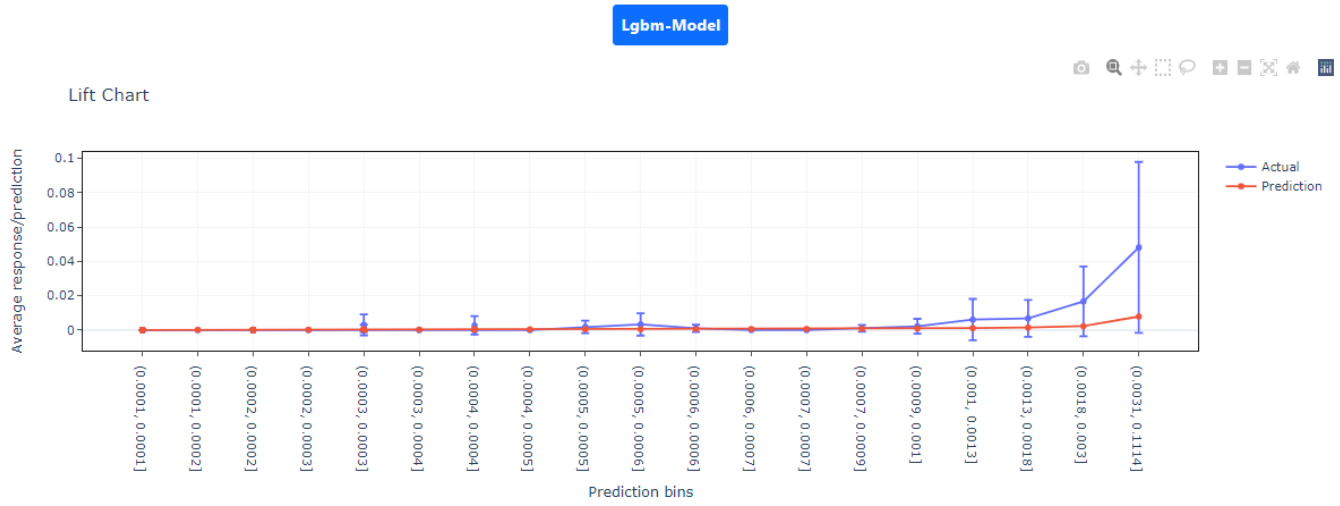


Figure 54 - Lift chart couple avec ajout des variables du conjoint

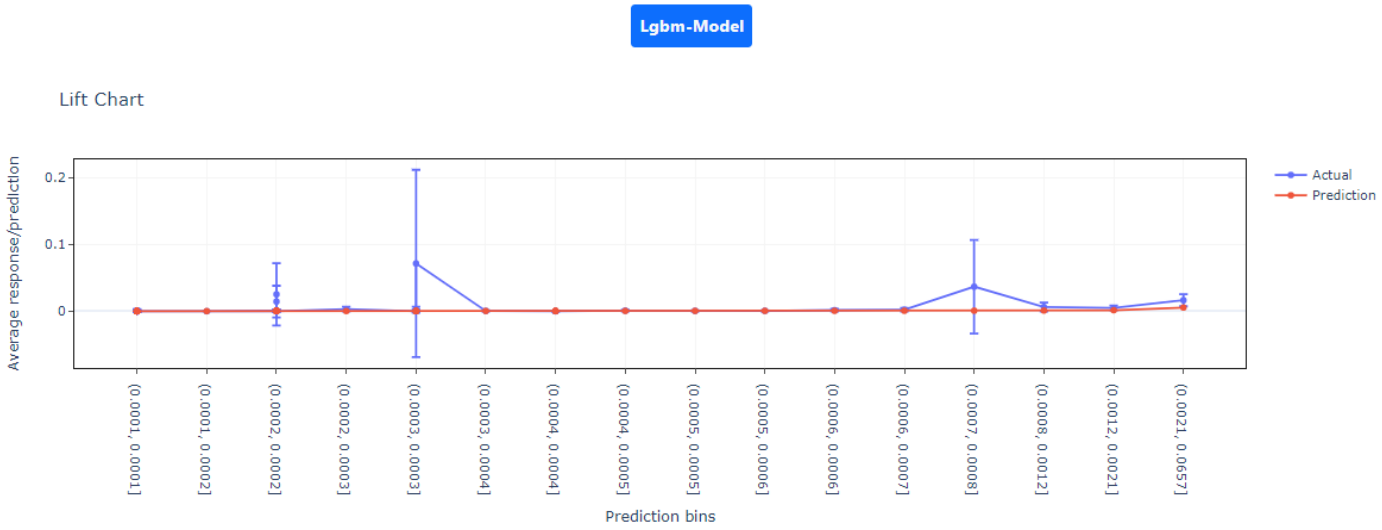


Figure 55 - Lift chart population totale

Si nous nous intéressons au graphique de Lift-chart du modèle $M_{c(x,y)}$, nous remarquons que les erreurs commises par le modèle $M_{c(x,y)}$ sont moins importantes que celles commises par M_i sur la population couple. Ceci explique la raison pour laquelle nous avons une adéquation globale plus pertinente avec $M_{c(x,y)}$.

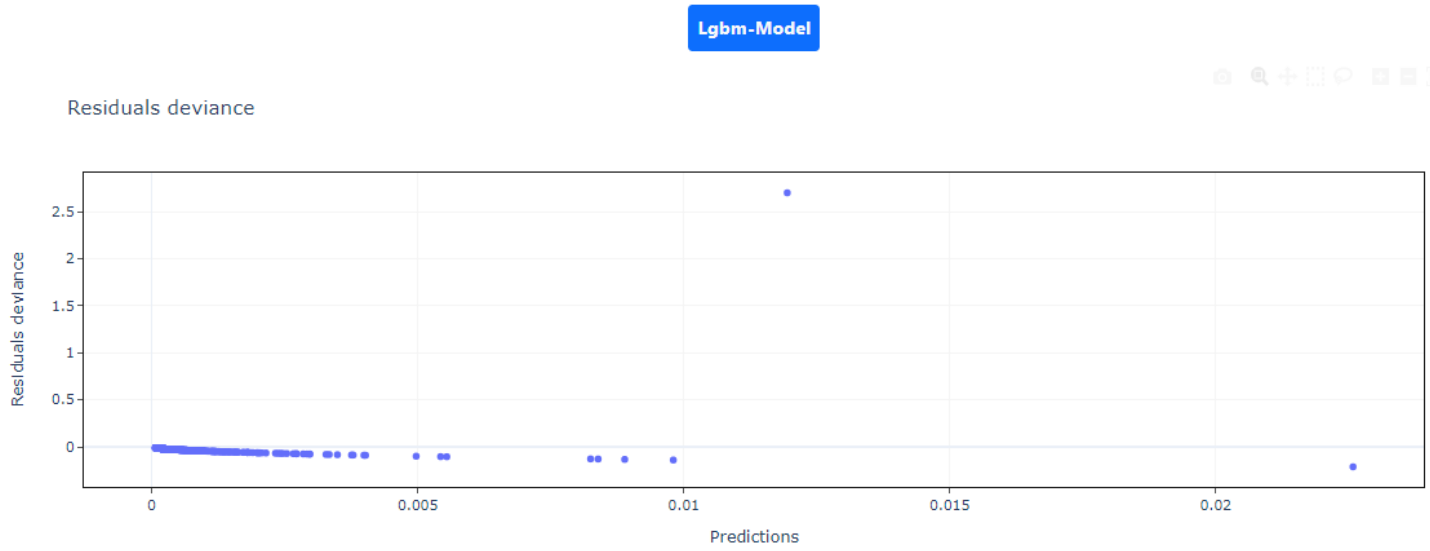


Figure 56 - Déviance résiduelle du modèle couple

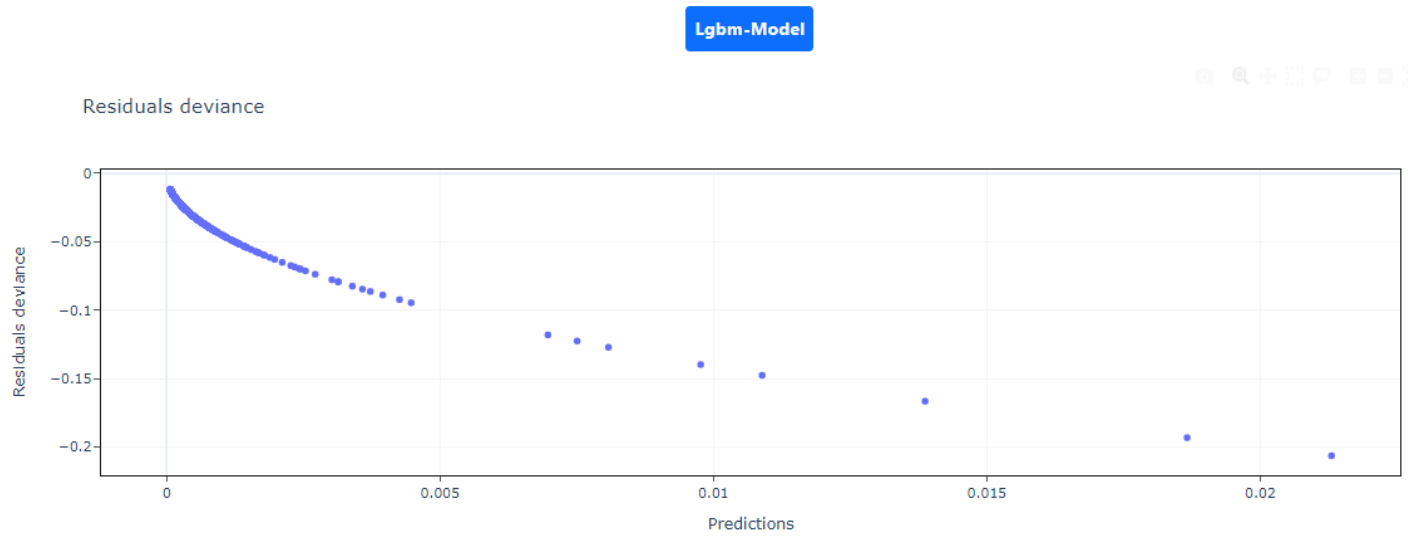


Figure 57 - Déviance résiduelle du modèle individuel

Le graphique donnant les résidus de déviance de $M_{c(x,y)}$ montre que ceux-ci sont plus proches de 0 que ceux de la modélisation individuelle. Ceci peut supposer que la population globale a un comportement diffèrent par rapport à celle des couples.

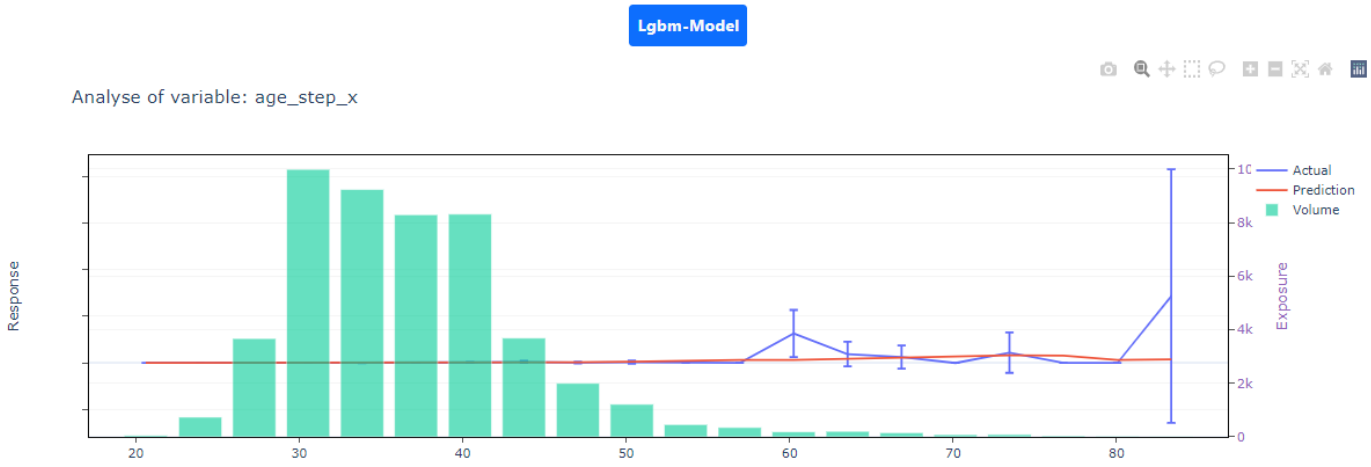


Figure 58 - Analyse de la variable age_step en fonction de l'exposition du modèle couple

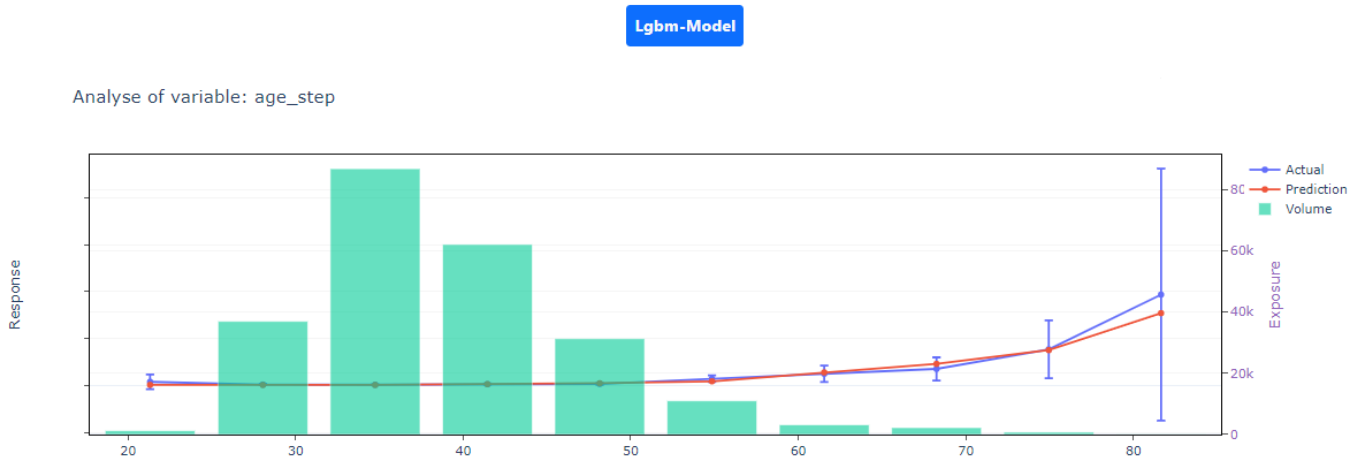


Figure 59 - Analyse de la variable age_step en fonction de l'exposition du modèle individuel

Ces deux graphiques comparent le comportement de la sinistralité prédite et la sinistralité historique en fonction de l'âge de la tête modélisée.

Malgré la plus forte exposition de la population totale, le modèle entrainé sur la population en couple semble présenter une très bonne adéquation avec la sinistralité observée de 20 à 65 ans. A partir de 65 ans, nous observons de petits écarts que le modèle résout avec une mutualisation des risques de 65 ans à 80 ans. Ces écarts restent quand même dans l'intervalle de confiance qui est représenté par la barre bleue verticale, ce qui est acceptable. L'intervalle de confiance est de 99%.

Le modèle individuel quant à lui présente des écarts à partir de 72 ans, mais propose une correction moins efficace que le modèle en couple. Cela conduit à une adéquation globale de 90% contre 95% pour le modèle en couple.

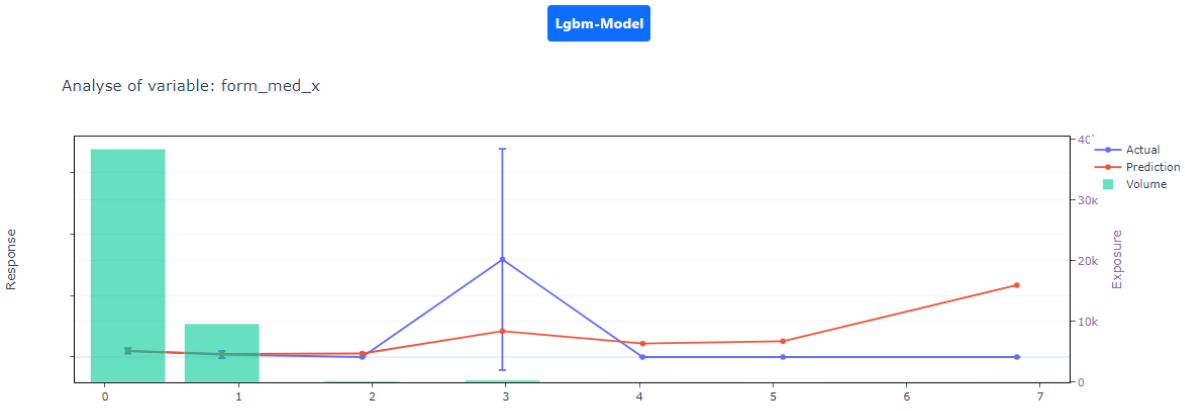


Figure 60 - Analyse de la variable form_med en fonction de l'exposition du modèle couple

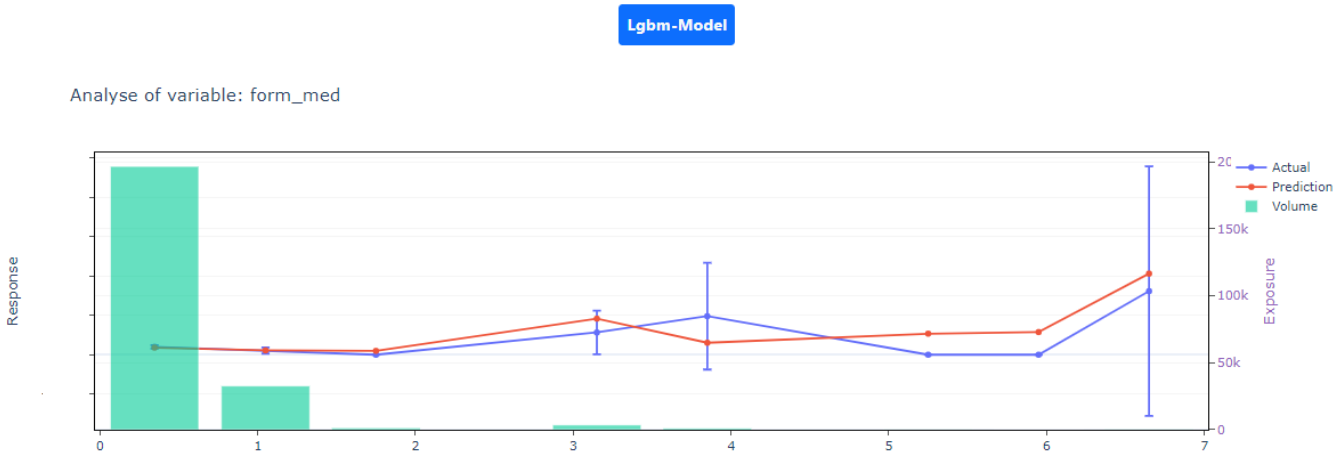


Figure 61 - Analyse de la variable form_med en fonction de l'exposition du modèle individuel

Lorsque nous analysons les prédictions en fonction du type de formalités médicales subites par l'assuré, nous remarquons que les prédictions des deux modèles sont fidèles à la réalité pour les formalités 0,1 et 2.

A partir de la formalité 3, la fracture entre sinistralité historique et prédite est plus importante pour le modèle entrainé sur la population en couple que celui entrainé sur la population totale prédit sur les couples noté M_i.

Ceci peut se justifier par le fait de l'extrême faible exposition de la population en couple dans ces catégories.

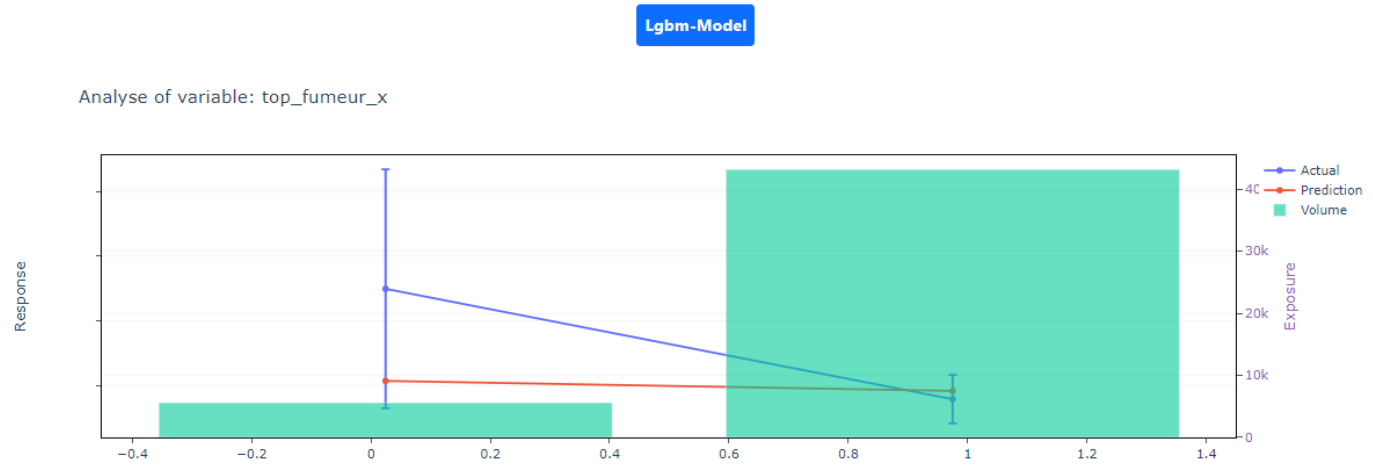


Figure 62 - Analyse de la variable top_fumeur en fonction de l'exposition du modèle couple

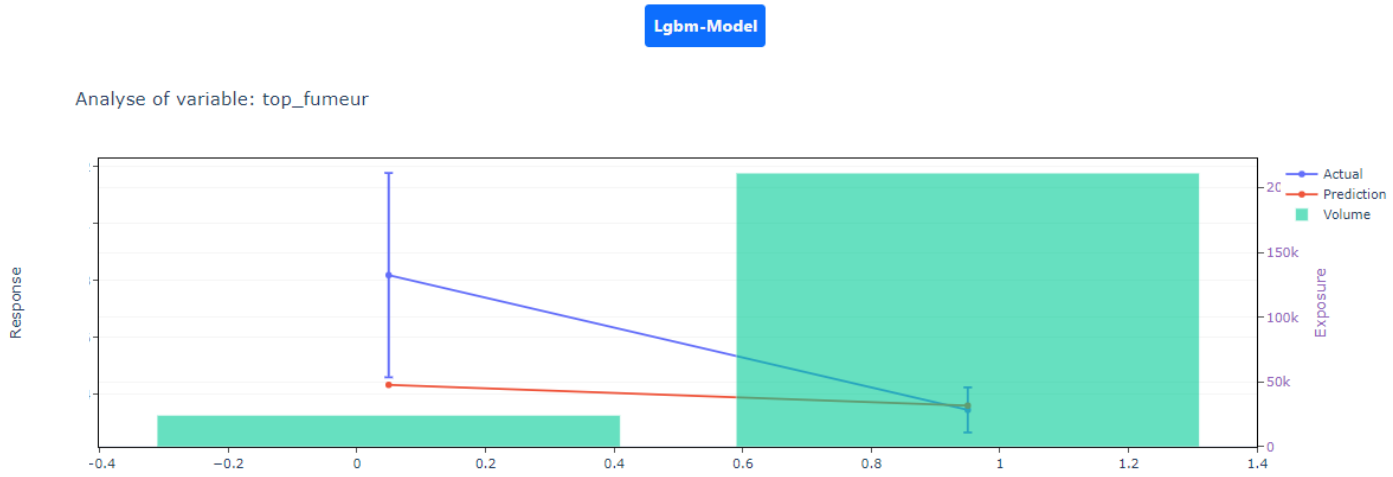


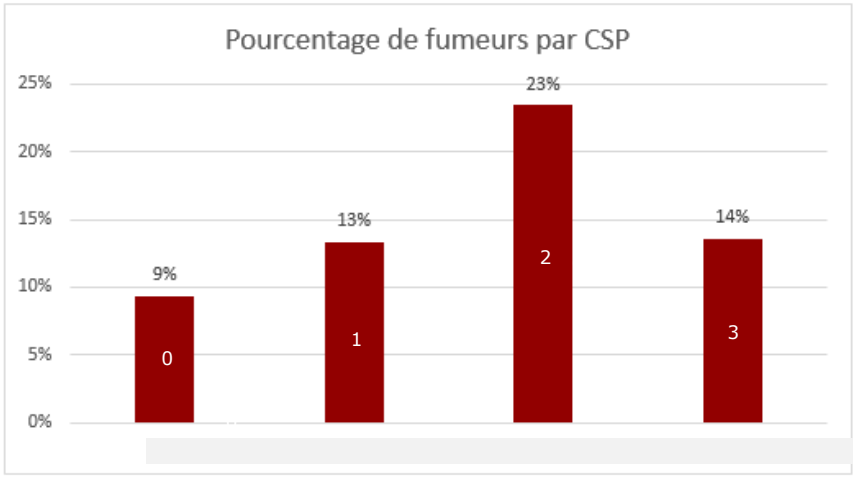
Figure 63 - Analyse de la variable top_fumeur en fonction de l'exposition du modèle individuel

Dans ces deux graphiques, nous observons la mortalité en fonction du statut fumeur et non-fumeur de l'assuré. La variable 0 fait référence à la population des fumeurs. Le graphique indique que les modèles $M_{c(x,y)}$ ainsi que M_i ne donnent pas une forte pertinence à cette variable. Nous observons une légère augmentation du taux de hasard prédit pour la population des fumeurs, mais ceci n'est pas aussi significatif que la sinistralité observée. Ce manque de pertinence vis-à-vis de cette variable peut être lié à la parcimonie des données. En effet, la proportion de fumeurs n'est pas efficiente. En outre, les effets néfastes de la cigarette ne sont pas observables sur les populations jeunes représentatives de ce portefeuille.

S'ajoute à cela le déséquilibre de la variable prédite. En effet, le taux de survie de la population du portefeuille est de 99.8%. Ce qui nous laisse très peu d'observation de décès afin d'obtenir une modélisation pertinente.

Outre la parcimonie des données, nous pouvons voir que l'effet du tabac est très variable lorsque celui-ci est combiné à d'autres informations.

En effet, la distribution de la variable fumeur est différente en fonction de la catégorie socio-professionnelle. Ici nous regardons les contributions de la variable top_fumeur en fonction de la variable CSP (qui donne la catégorie socio-professionnelle de l'individu assuré). Ainsi nous pouvons voir les tendances de la modélisation en fonction de ces deux variables.



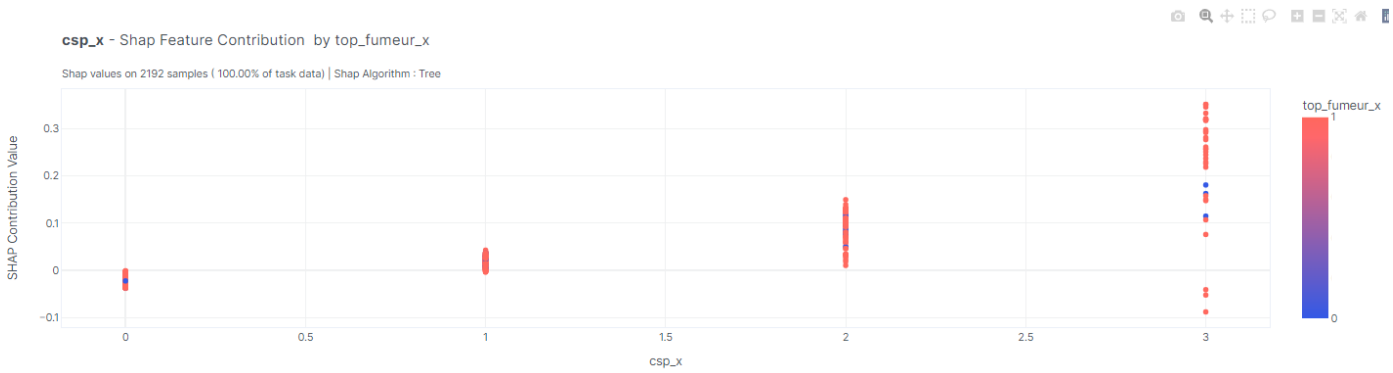


Figure 64 - Corrélation entre CSP et tabac dans $M_{c(x,y)}$

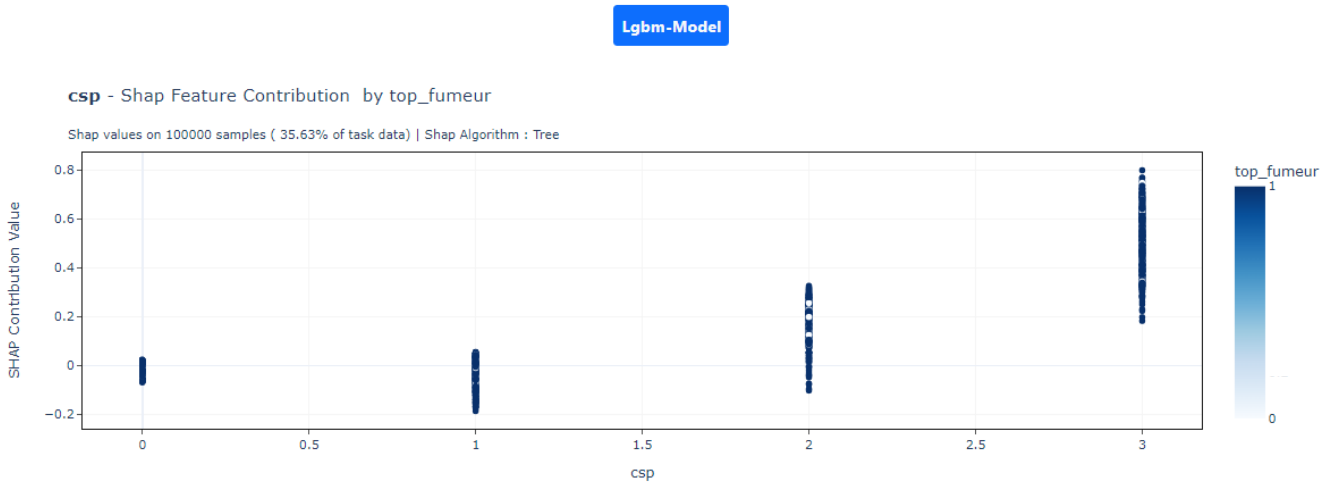


Figure 65 - Corrélation entre CSP et tabac dans M_i

Nous remarquons que les fumeurs de la catégorie 0 ont une sinistralité théorique plus importante dans la population globale par rapport à la population en couple. La pertinence de cette variable est croissante en fonction de la catégorie socio-professionnelle ce qui est cohérent avec le niveau de sinistralité de ces catégories. Nous remarquons que l'effet de la catégorie socio-professionnelle est plus important que l'effet du tabac. Ceci peut se retrouver dans le graphe de statistiques descriptives suivant :

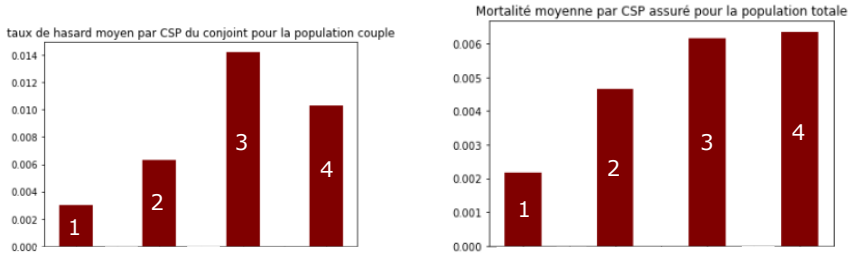


Figure 66- Mortalité moyenne en fonction de la catégorie socio-professionnelle

L'effet de la variable top_fumeur sur le taux de hasard peut également dépendre de l'âge des assurés étudiés. Comme le montrent les graphiques ci-dessous :

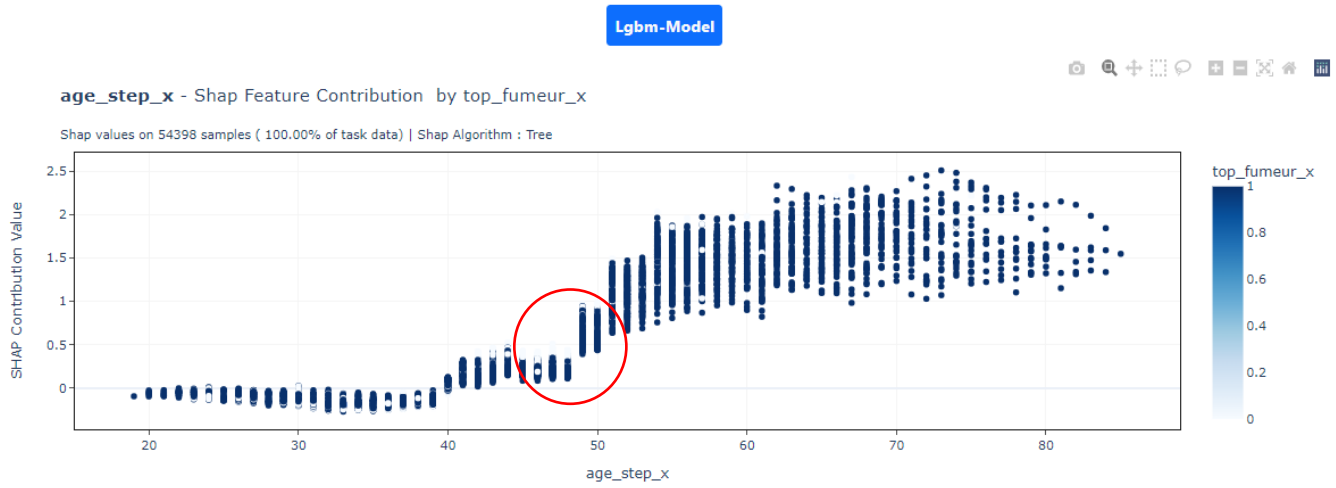


Figure 67 - Corrélation entre l'âge et le tabac dans $M_{c(x,y)}$

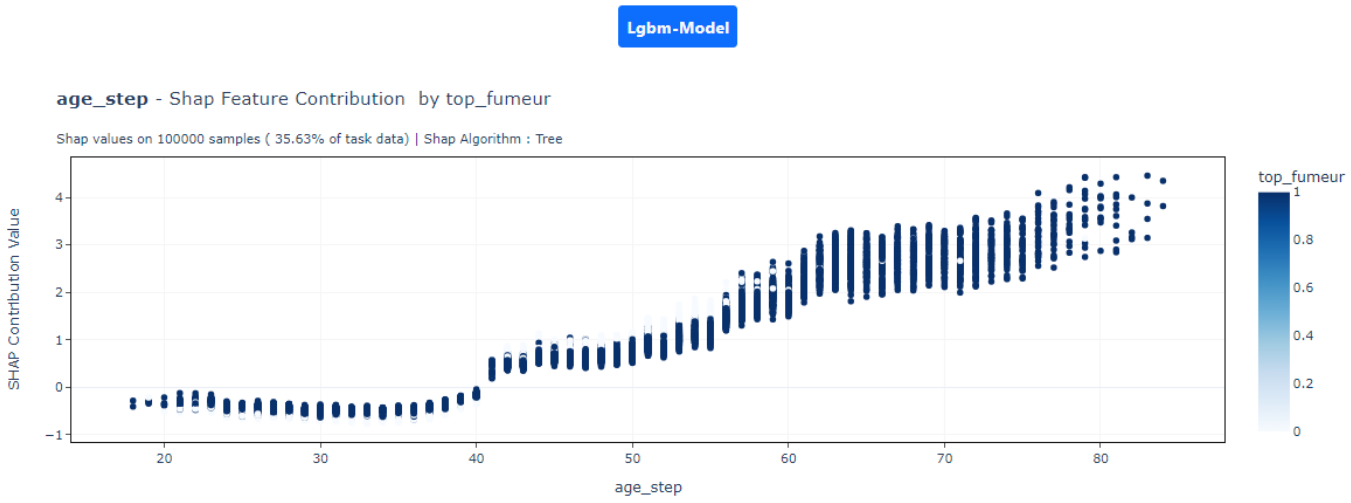


Figure 68 - Corrélation entre l'âge et le tabac dans M_i

En observant ces graphiques, nous pouvons voir que les valeurs de Shapley de cette variable sont croissantes en fonction de l'âge des assurés étudiés. Ceci traduit le fait que le tabagisme agit positivement sur le taux de décès instantané en fonction de l'âge. Cette information est conforme à ce qui est annoncé par la fédération française de cardiologie qui stipule que l'effet du tabac est plus important dans le grand âge du fait de l'augmentation des risques cardio-vasculaires à partir de 55 ans.

Nous retrouvons cet effet avec le modèle $M_{c(x,y)}$ qui montre une fracture nette des valeurs de Shapley à partir de 50 ans, tandis que le modèle M_i suggère une fracture plus nette à partir de 40 ans. De plus, toutes choses égales par ailleurs, les valeurs de Shapley sont moins fortes dans le modèle $M_{c(x,y)}$ que dans le modèle M_i. Ceci peut émaner du fait que la population en couple a un profil moins risqué comme le montre le taux de décès instantané de la population couple par rapport à la population totale. (**Attention** : les graphiques ne sont pas à la même échelle).

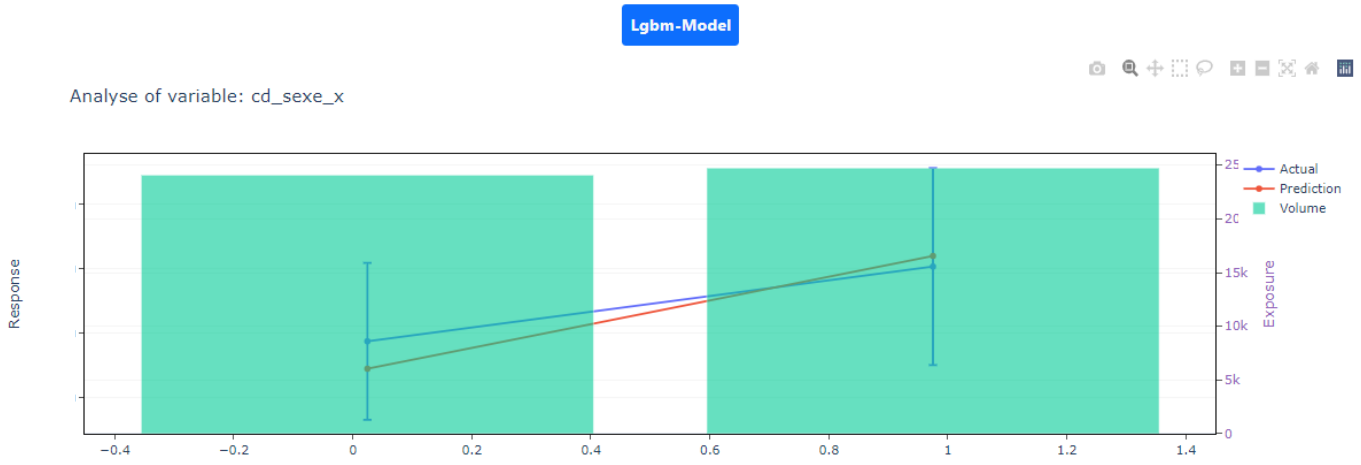


Figure 69 - Analyse de la variable cd_sexe en fonction de l'exposition du modèle couple

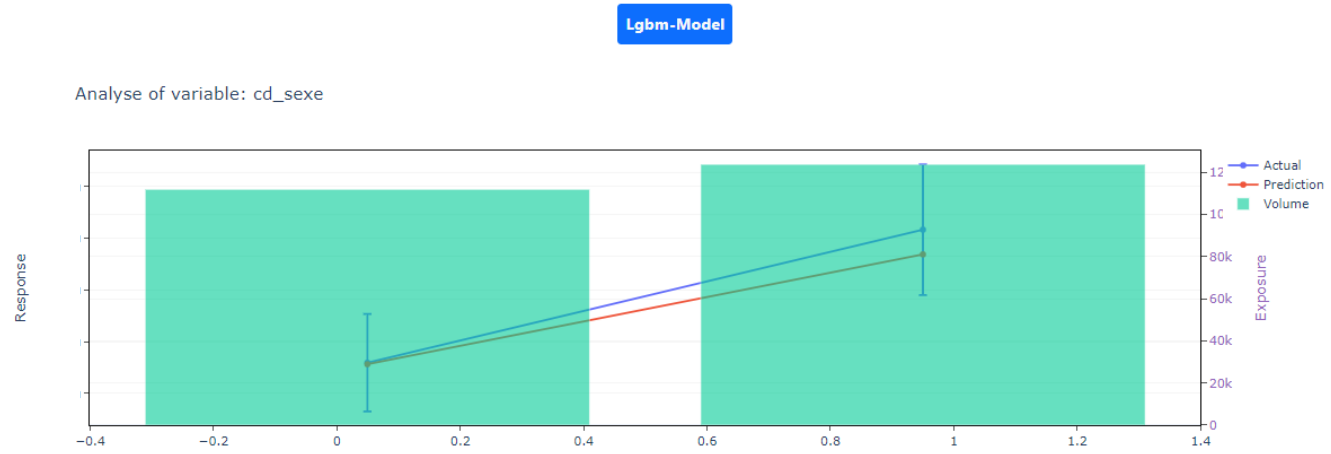


Figure 70 - Analyse de la variable cd_sexe en fonction de l'exposition du modèle individuel

Le modèle couple sous-estime la mortalité des femmes et surestime celle des hommes. Ceci n'est pas tellement important car l'estimation de la prime finale se fait avec un individu de chaque sexe dans 95% des cas (en effet, la base est constituée de 95% de couples de sexes différents). Pour les 5% restants, nous pouvons adapter un modèle équitable qui consiste à choisir la fréquence moyenne parmi les combinaisons suivantes :

Sexe réel	Combinaison 1	Combinaison 2
F	H	F
F	F	H

Sexe réel	Combinaison 1	Combinaison 2
H	H	F
H	F	H

Figure 71 - Combinaison pour les couples de même sexe

Ceci permet d'analyser la tarification en fonction du sexe afin de prendre en compte la structure du portefeuille. De plus, la modélisation restera valable en cas de déformation de celui-ci.

En somme, cette pratique permet de préserver l'équité au sein du couple sans faire de discrimination entre les couples de même sexe et les autres.

La mensualité constitue une information importante dans la modélisation effectuée. En effet, elle occupe la troisième position dans la modélisation $M_{c(x,y)}$ et la cinquième dans la modélisation M_i.

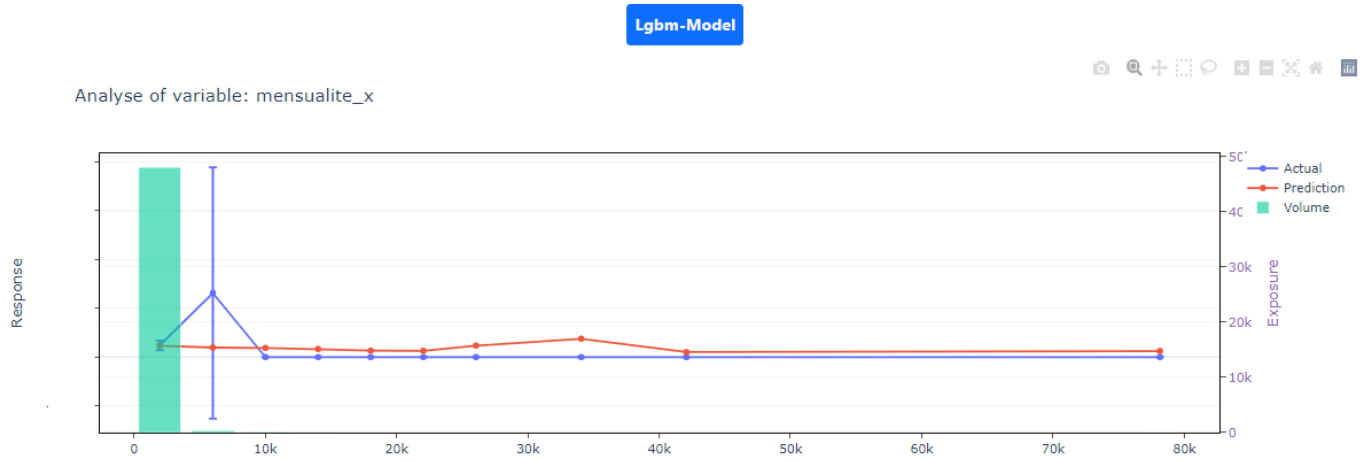


Figure 72 - Analyse de la variable mensualité en fonction de l'exposition du modèle couple

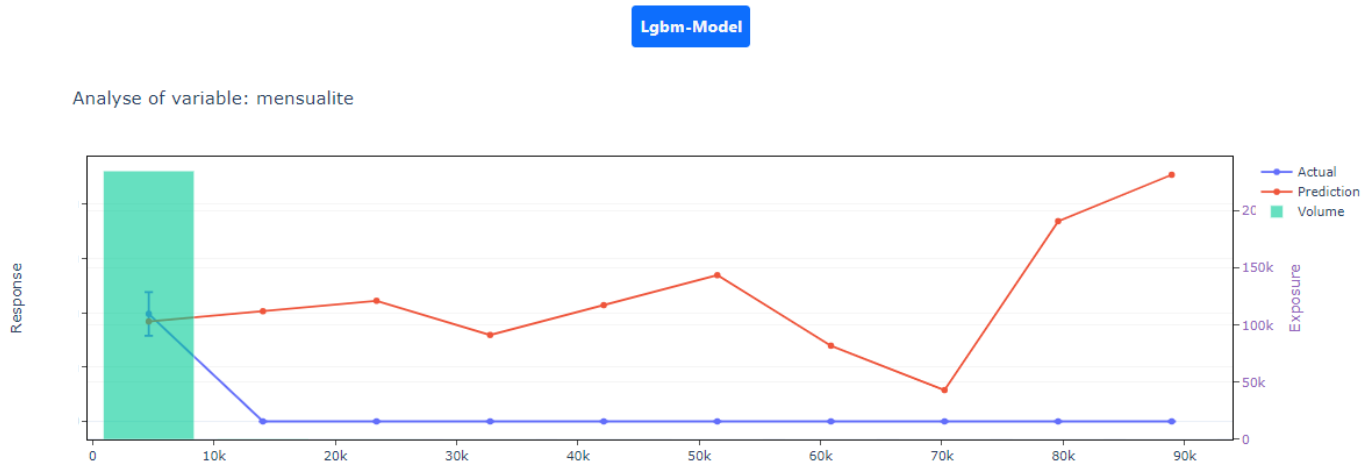


Figure 73 - Analyse de la variable mensualité en fonction de l'exposition du modèle individuel

La sinistralité historique semble indiquer une sinistralité nulle au-delà d'une mensualité supérieure à 10K pour les deux modèles.

Le modèle $M_{c(x,y)}$ propose une adéquation parfaite pour les petites mensualités. Ces prêts concernent plus souvent les populations dont les revenus sont faibles. Or, le salaire est corrélé avec la zone géographique et la facilité ou non de l'accès aux soins.

Si nous croisons cette information avec les lieux des déserts médicaux selon ce graphique :

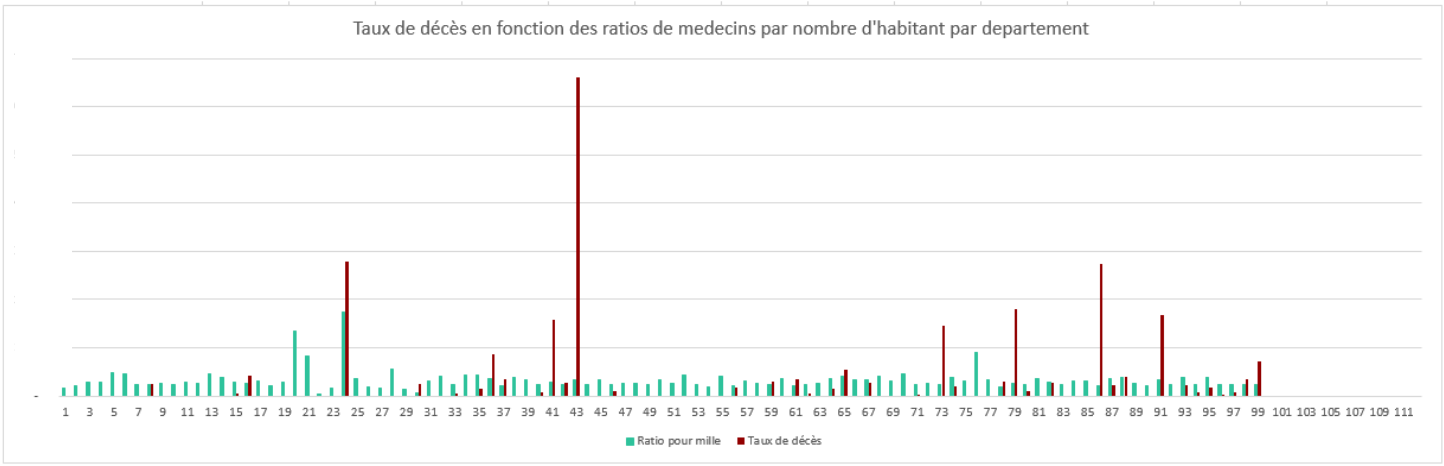


Figure 74- Le taux de décès en fonction des déserts médicaux

Nous déduisons qu'hormis pour la Dordogne, les hauts taux de décès correspondent aux endroits où les professionnels de santé sont manquants.

Si nous enlevons le maximum atteint pour la haute Loire, nous obtenons le graphique suivant :

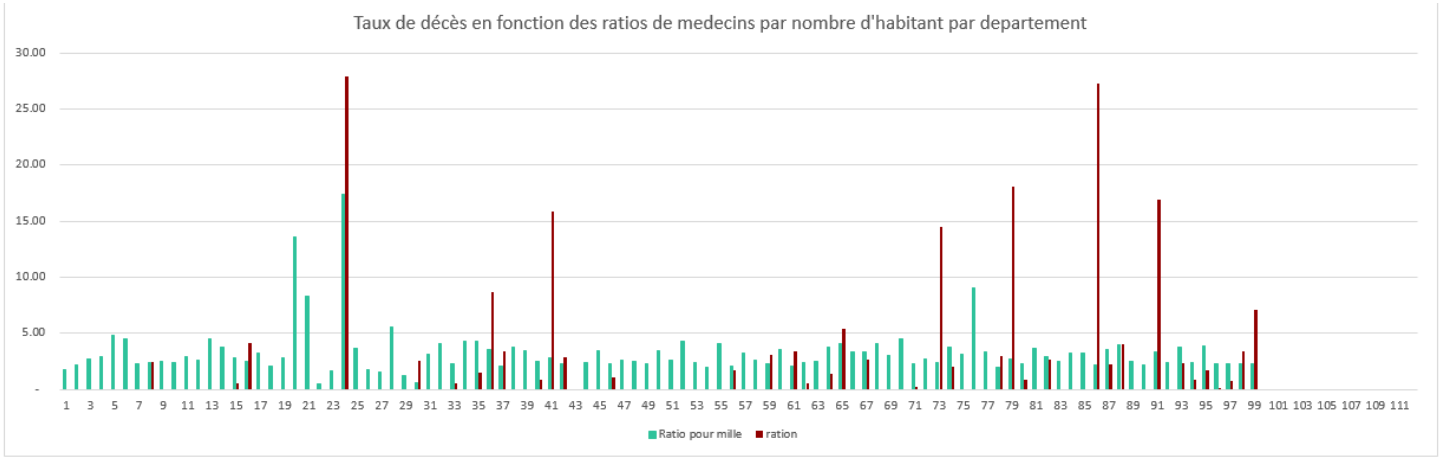


Figure 75- Le taux de décès en fonction des déserts médicaux

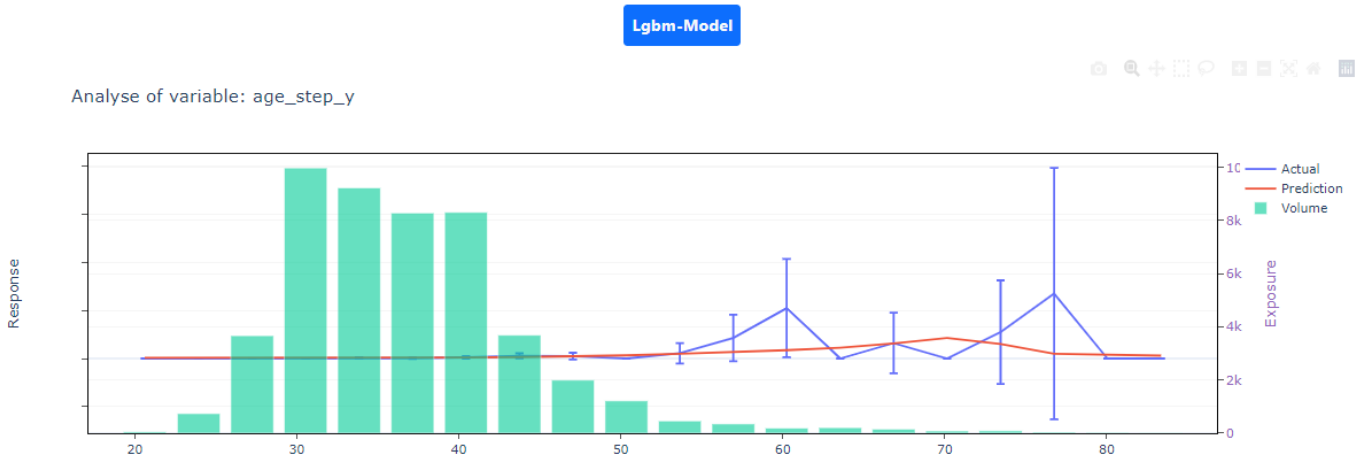


Figure 76 - Analyse de la variable agreste en fonction de l'exposition du modèle couple

La sinistralité historique semble plus importante lorsque le conjoint de la tête assurée est âgé. Le modèle $M_{c(x,y)}$ semble cohérent dans la prédiction du taux de décès instantané en fonction de l'âge du conjoint. Néanmoins la qualité de l'adéquation est dégradée au-delà de 60 ans, ce qui peut s'expliquer par la faible exposition pour ces âges.

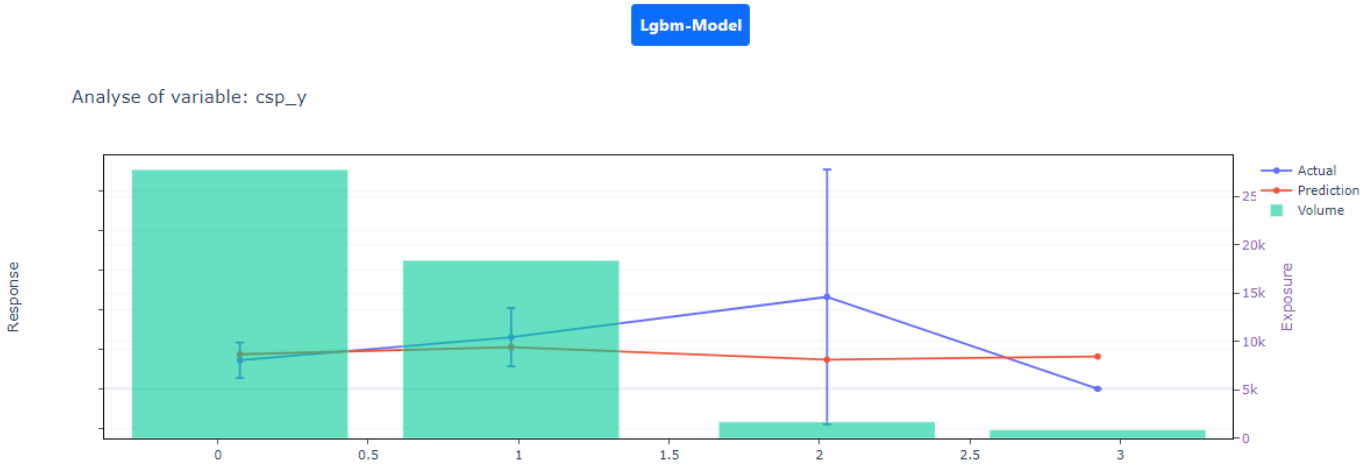


Figure 77 - Analyse de la variable csp_y en fonction de l'exposition du modèle couple

Si nous regardons la modélisation selon la variable qui indique la catégorie socio-professionnelle du conjoint de la tête modélisée, nous pouvons voir une bonne adéquation pour les catégories 0 et 1. Cependant le modèle semble un peu moins fidèle quant aux catégories 3 et 4. Encore une fois, la modélisation commet une erreur dans les catégories avec moins d'exposition.

Cela dit, le modèle $M_{c(x,y)}$ semble plus cohérent que la sinistralité historique au niveau de la catégorie 3.

Si nous observons le graphique du taux de décès moyen observés en fonction du département par catégorie socio-professionnelle, nous pouvons observer une sinistralité plus haute pour la catégorie 3 en Dordogne (département 42). Ceci converge avec le pic du taux de décès par département dans la Figure 159 - Taux de hasard moyen par département pour les couples.

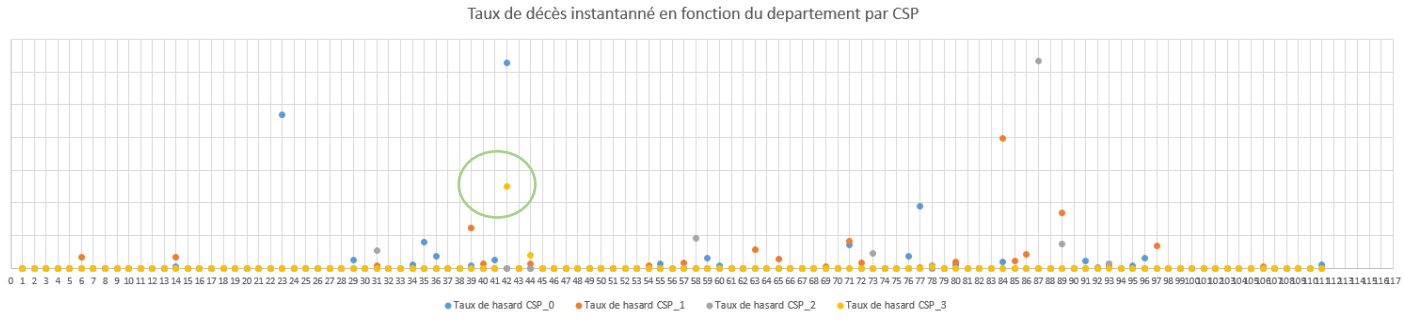


Figure 78 - Taux de décès instantané par département et CSP

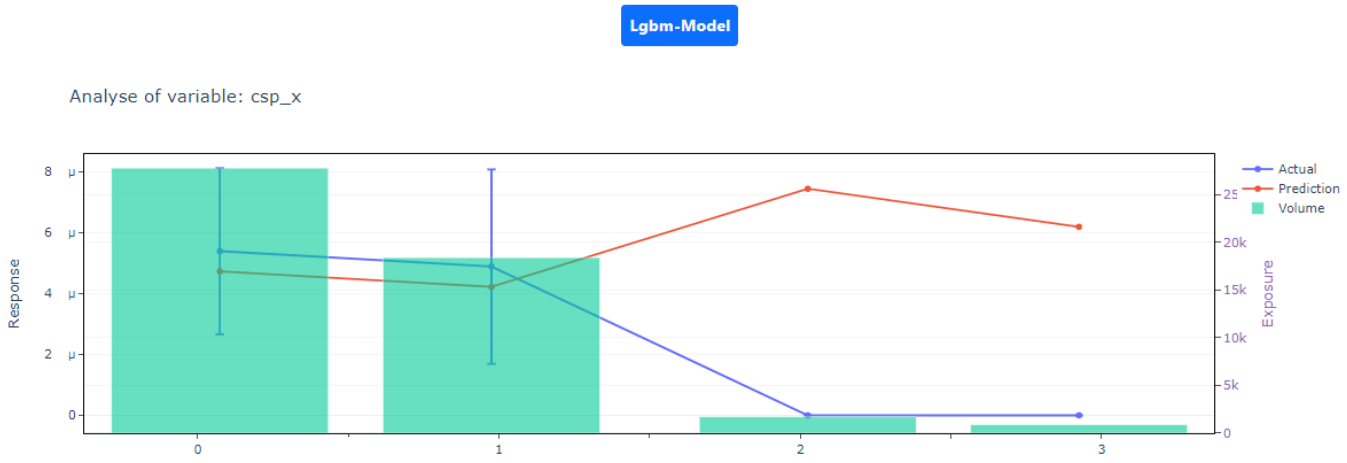


Figure 79 - Analyse de la variable csp_x en fonction de l'exposition du modèle couple

Nous pouvons faire la même observation pour la catégorie socio-professionnelle de l'individu assuré.

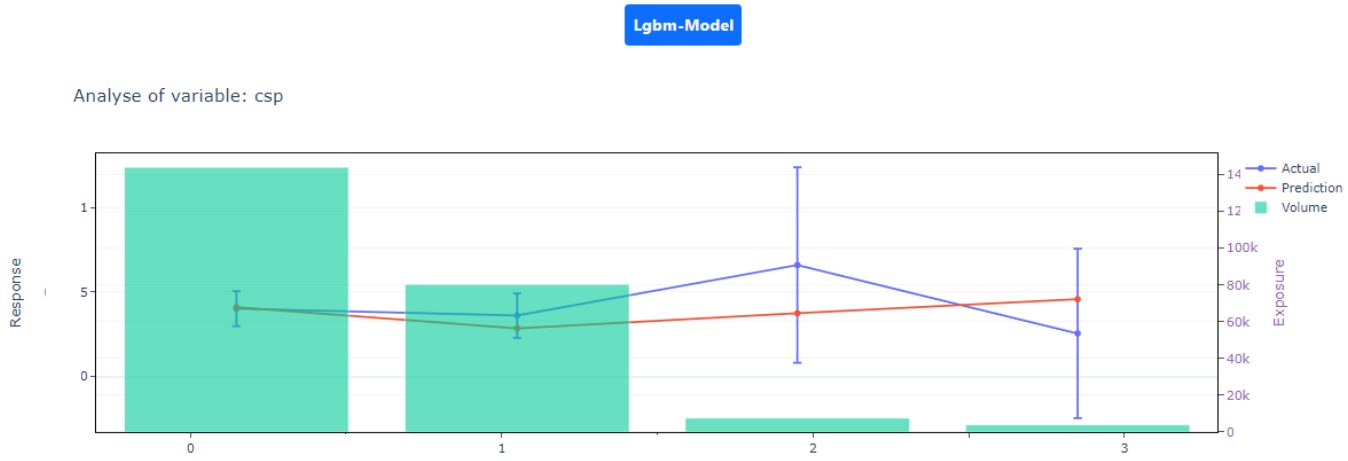


Figure 80 - Analyse de la variable CSP en fonction de l'exposition du modèle individuel

Le modèle M_i semble mieux apprécier la sinistralité en fonction de la catégorie socio-professionnelle de la tête assuré tandis que le modèle $M_{c(x,y)}$ semble plus fidèle à la catégorie socio-professionnelle du conjoint que celle de la tête assurée. Cela rejoint les résultats de la Figure 157 - Taux de hasard moyen en fonction de la CSP du conjoint

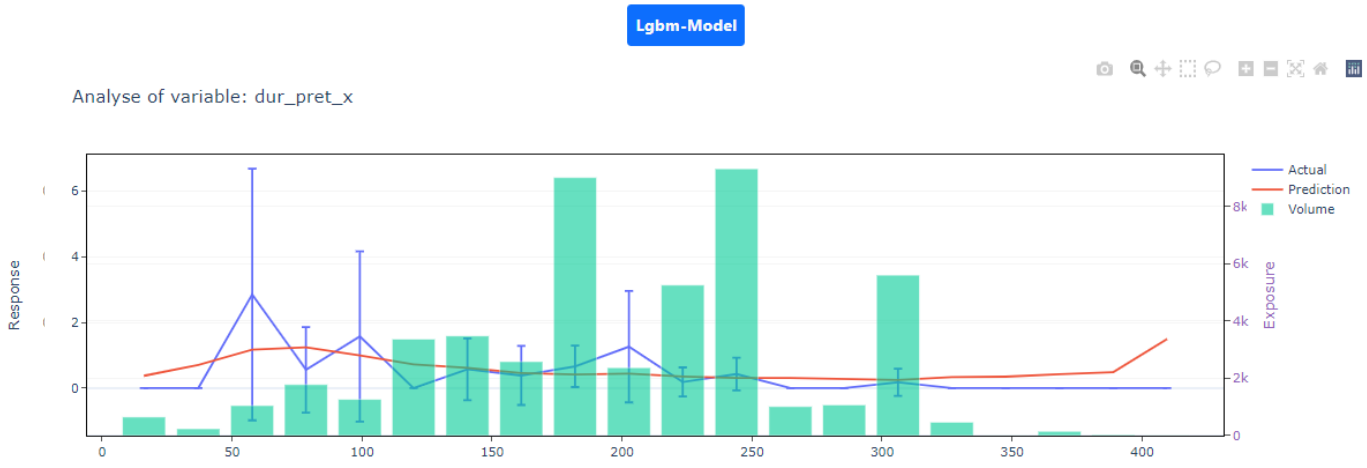


Figure 81 - Analyse de la variable dure_pret en fonction de l'exposition du modèle couple

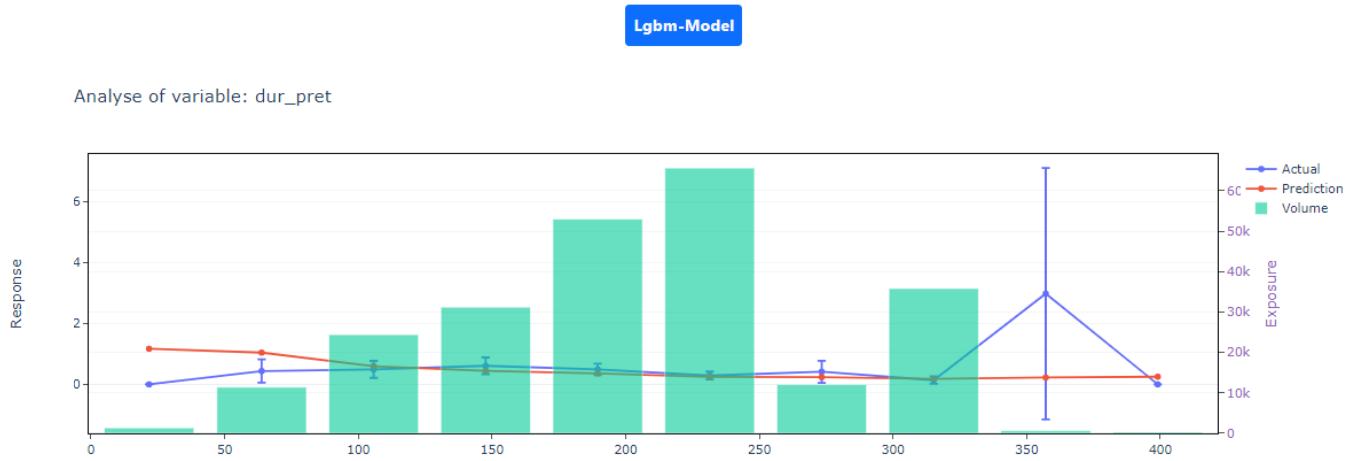


Figure 82 - Analyse de la sinistralité en fonction de la durée du prêt en fonction de l'exposition du modèle individuel

De manière globale, le biais de modélisation est plus important dans la modélisation $M_{c(x,y)}$. Cependant, il faut prendre en considération le fait que M_i est entrainé sur un échantillon beaucoup plus grand. A savoir 600000 lignes contre 100000 pour $M_{c(x,y)}$. Cela est sensé lui conférer une plus grande stabilité.

Malgré cela, l'erreur commise par M_i semble beaucoup trop brutale. De plus $M_{c(x,y)}$ semble surestimer la mortalité dans les zones à faible exposition, ce qui permet d'avoir un modèle plus prudent.

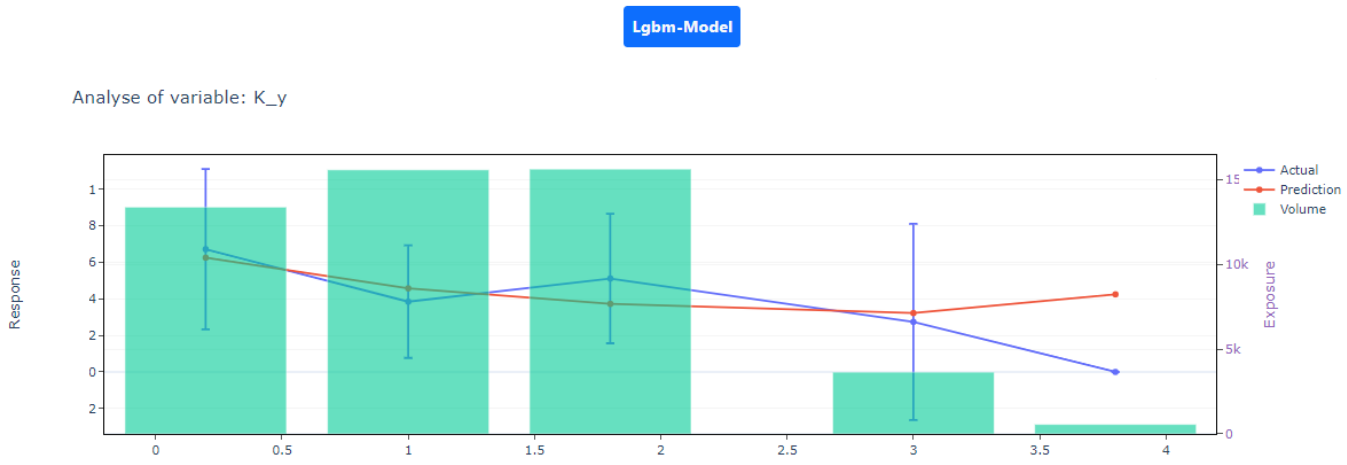


Figure 83- Analyse de la variable K_y en fonction de l'exposition du modèle couple

L'analyse de la sinistralité historique de l'assuré en fonction de la tranche de capital total assuré (tous prêts confondus) de son conjoint semble donner une indication non négligeable sur la corrélation entre le taux de décès instantané de l'assuré et la capacité d'emprunt de son conjoint. Le modèle $M_{c(x,y)}$ semble également capter ce lien et sa tendance reste fidèle à l'observation. Nous observons néanmoins un biais plus fort pour la catégorie 4 due à la faible exposition (nombre d'assurés inférieur à 500 personnes) dans cette tranche.

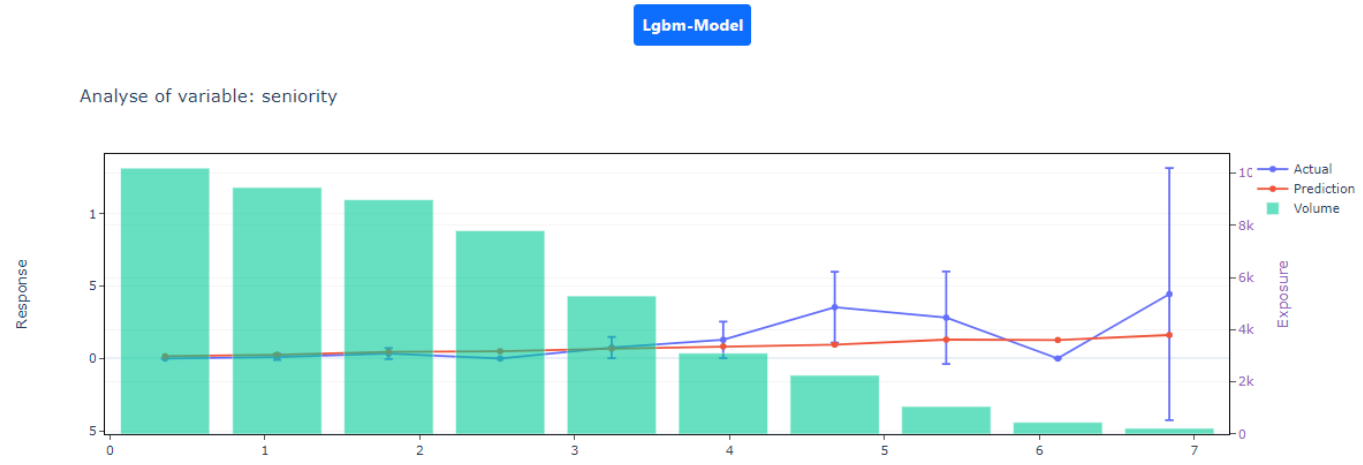


Figure 84- Analyse de la sinistralité en fonction de l'ancienneté du modèle couple

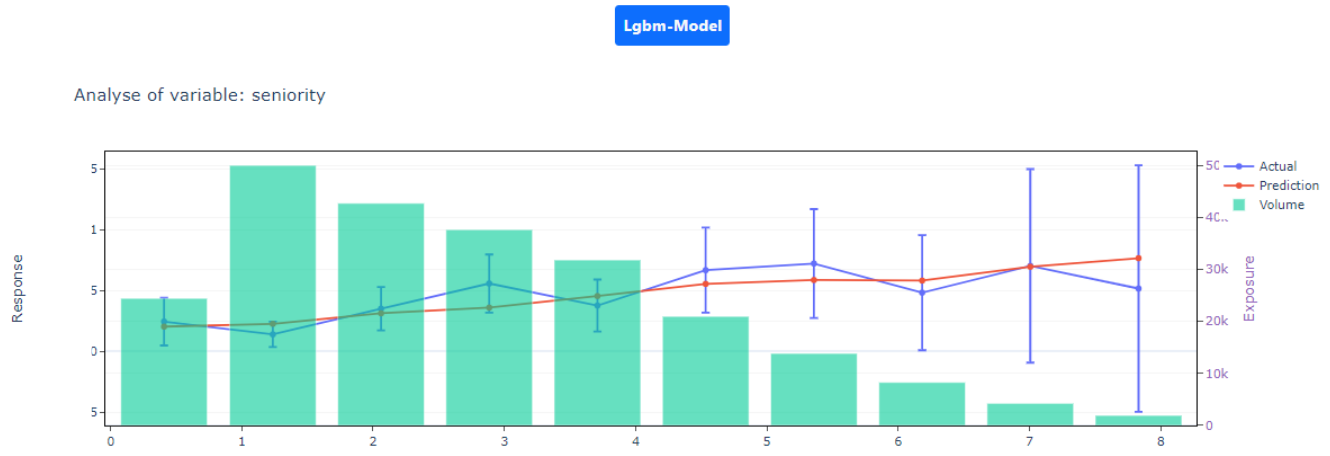


Figure 85 - Analyse de la variable seniority en fonction de l'exposition du modèle individuel

L'étude de la sinistralité en fonction de l'ancienneté de l'assuré semble être mieux estimée par $M_{c(x,y)}$ que par M_i. En effet, ce dernier a un biais global plus important. Cette variable est importante car la sinistralité observée semble indiquer que la mortalité est croissante en fonction de l'ancienneté de l'individu pour tous les âges confondus. Les raisons sous-jacentes sont l'augmentation de l'âge et l'atténuation de l'effet de la sélection médicale.

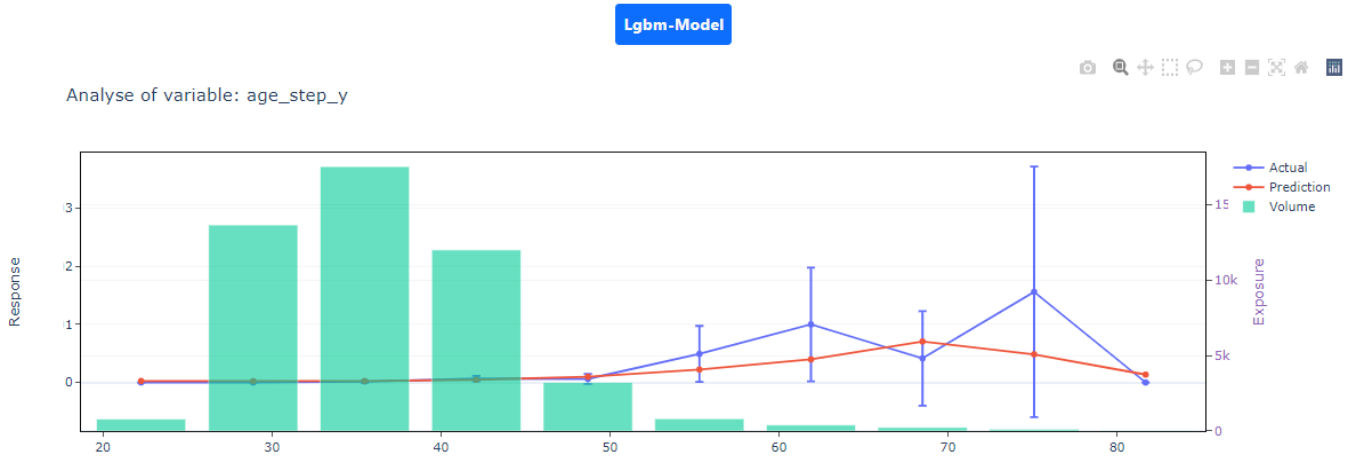


Figure 86 - Analyse de la variable age_step_y en fonction de l'exposition du modèle couple

De prime abord, la sinistralité croissante en fonction de l'âge du conjoint semble triviale car nous avons une corrélation de 85% entre les âges des conjoints. Cependant, les statistiques descriptives consistant à étudier la sinistralité des personnes âgés de plus de 40 ans en fonction de l'âge de leur conjoint ont démontré que les personnes qui sont en couple avec une personne plus âgée ont un taux de décès instantané plus important (cf. Figure 152 - Variance du taux de décès instantané).

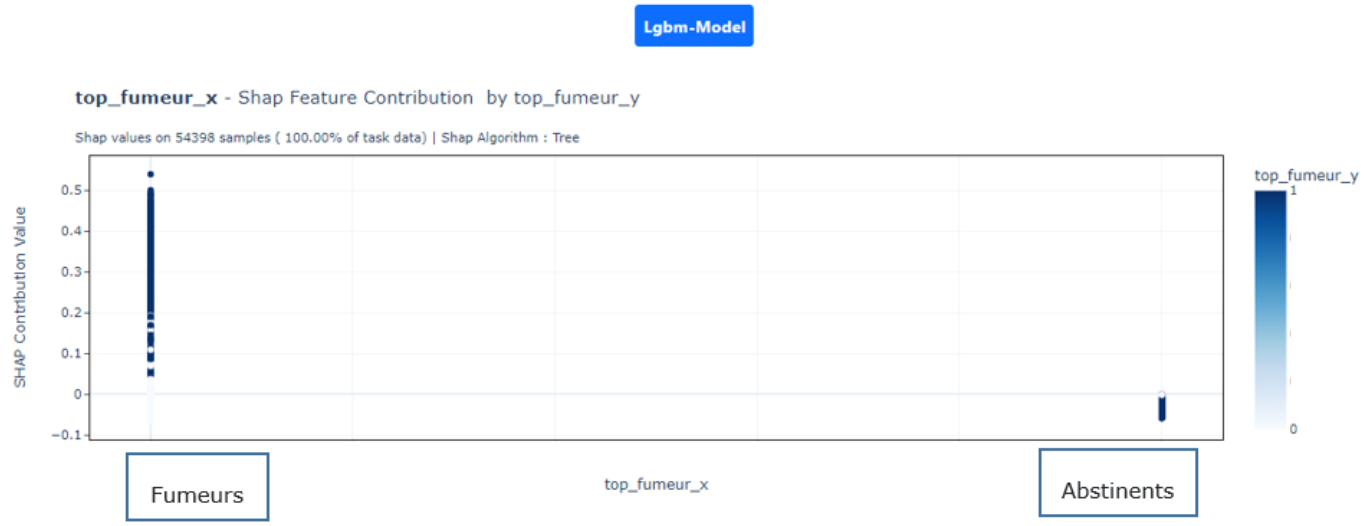


Figure 87- Analyse de la sinistralité en fonction du tabagisme du conjoint du modèle couple

De la même manière, $M_{c(x,y)}$ arrive à saisir la relation entre la mortalité et le statut de tabagisme du conjoint. En effet, les individus vivants avec un conjoint fumeurs semblent avoir une valeur de Shapley positive donc un taux de décès plus important que les personnes vivantes avec un non-fumeurs (valeur de Shapley négative, donc contribution à la baisse).

Afin d'évaluer la cohérence globale des deux modèles, nous utiliserons deux méthodes :

- Permutation importance
- Ration de Shap

Il s'agit d'une méthode qui consiste à permuter aléatoirement les caractéristiques d'une variable afin de quantifier l'impact que cette variable a sur la prédiction d'une valeur cible.

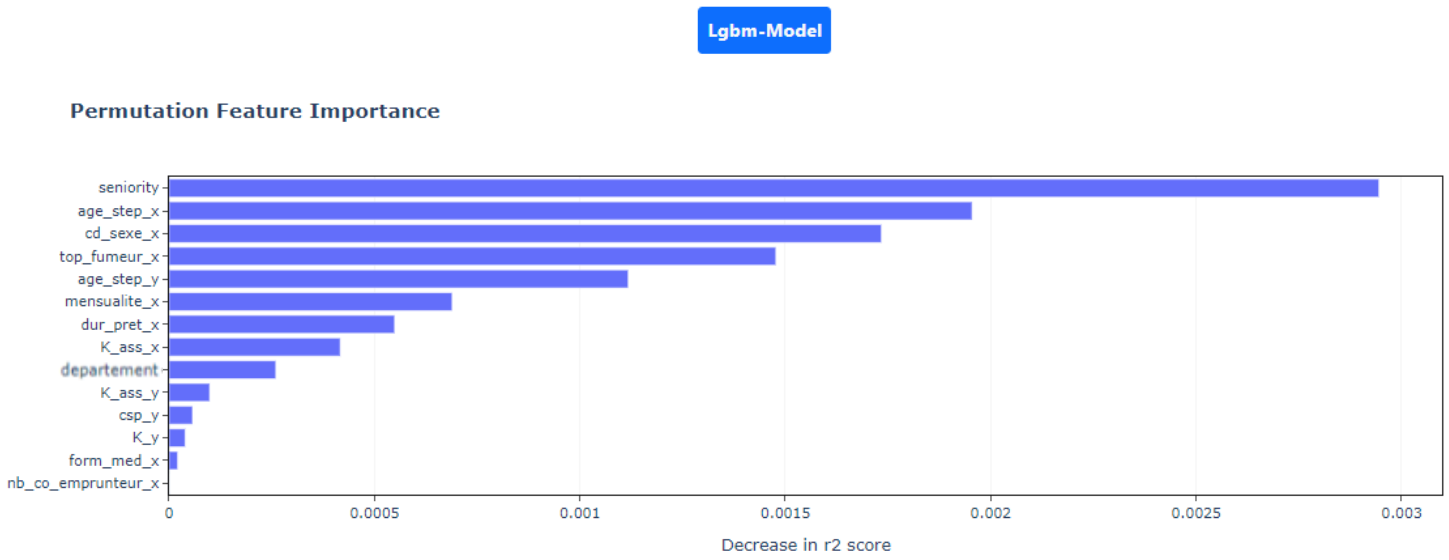


Figure 88- Importance des variables du modèle couple avec ajout des variables du conjoint par la méthode des permutations

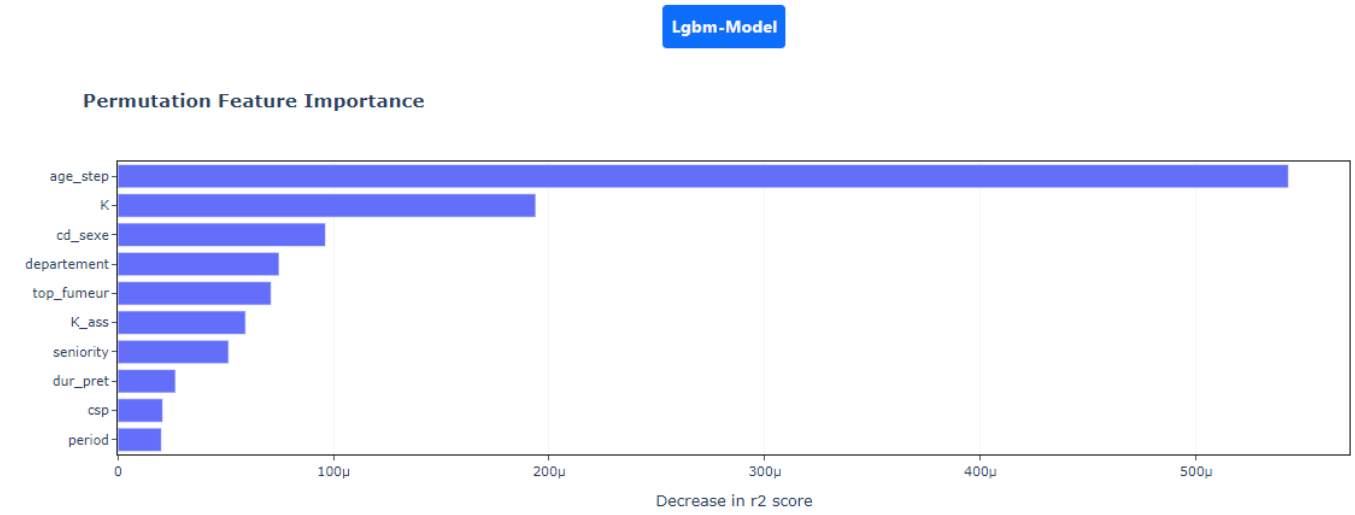


Figure 89 - Importance des variables du modèle individuel avec ajout des variables du conjoint par la méthode des permutations

Nous remarquons que $M_{c(x,y)}$ donne plus d'importance à l'ancienneté de l'individu que le modèle M_i. De plus, il semble indiquer que le taux de décès instantané est croissant avec cette variable, ce qui peut nuire au taux de rétention du portefeuille. Une modélisation dans l'ancienneté est retenue dans la Figure 164 - Taux de hasard en fonction de l'ancienneté.

En seconde position, nous observons la capacité d'emprunt de l'individu. Tandis que $M_{c(x,y)}$ exhibe la mensualité, M_i donne plus d'importance au capital global emprunté (pour tous les prêts confondus).

Sachant que la mensualité tient également compte du taux d'intérêt, il s'agit de ce fait d'une variable qui est plus efficace pour estimer la capacité d'emprunt d'un individu.

Tandis que $M_{c(x,y)}$ souligne l'importance des fumeurs dans l'augmentation de la sinistralité, le modèle M_i lui préfère le département de résidence de la tête assuré. La mortalité est certes différente en fonction du lieu de résidence du fait de la corrélation de cette variable avec le métier exercé ou l'âge de l'individu. Cependant, un modèle qui capte le tabagisme comme étant un facteur aggravant sur la mortalité lui est préférable car ces causes sont prouvées médicalement.

Cela dit, permutation importance ne permet pas de capturer la corrélation non linéaire au sein des variables, donc nous lui préfèrerons l'utilisation des valeurs de Shapley.

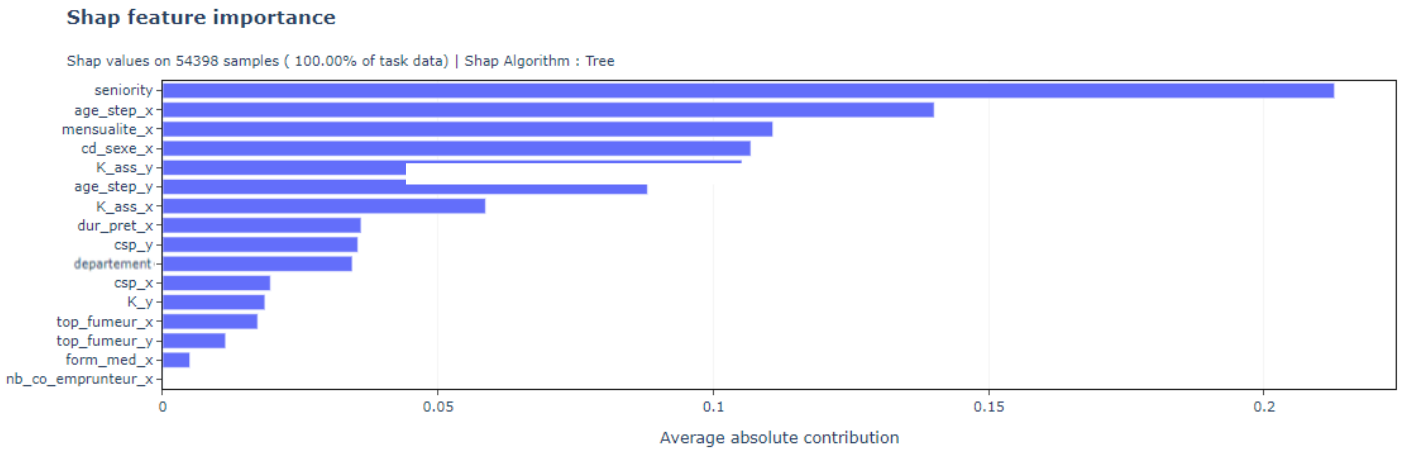


Figure 90- Importance des variables du modèle $M_{c(x,y)}$ en fonction des valeurs de Shapley

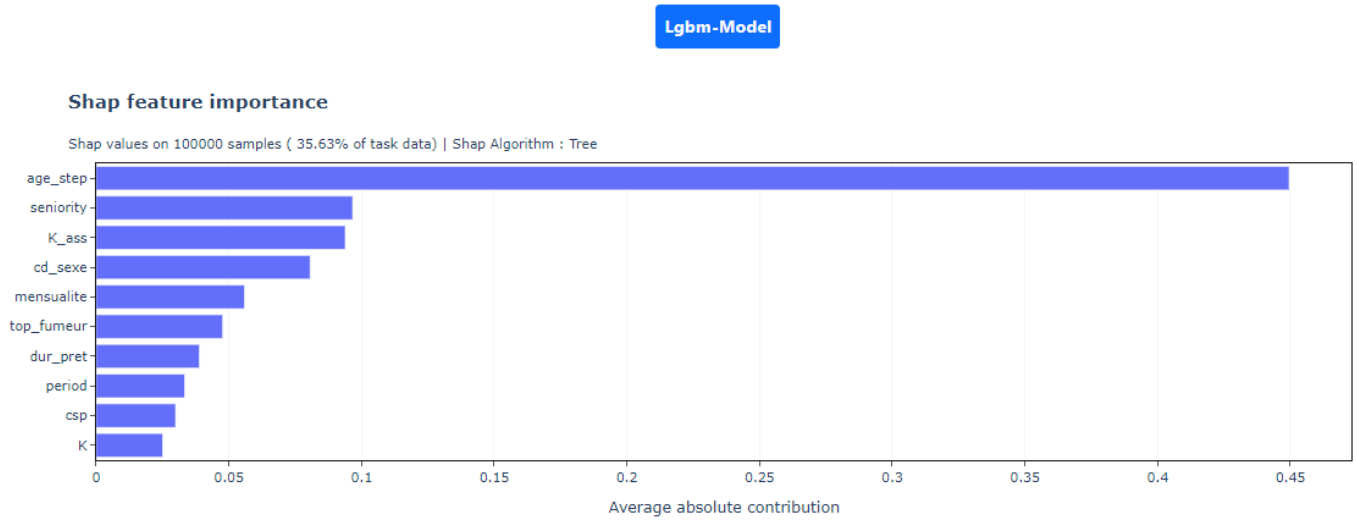


Figure 91- Importance des variables du modèle M_i en fonction des valeurs de Shapley

Hormis la mensualité qui semble plus importante dans le modèle $M_{c(x,y)}$, l'ordre des variables ne diffère pas beaucoup dans les deux modèles.

Nous remarquons malgré tout que $M_{c(x,y)}$ capte l'effet des variables du conjoint sur la mortalité de la tête assurée.

Local contribution plot

Un Local Contribution Plot (ou LCP, appelé également Local Feature Importance Plot) est un graphique qui permet de visualiser l'impact des caractéristiques (features) individuelles sur les prédictions d'un modèle de machine learning pour un point de données spécifique. Les LCP sont utilisés pour comprendre comment chaque caractéristique contribue aux prédictions du modèle pour un exemple particulier. Les LCP sont particulièrement utiles pour les modèles de machine learning à boîte noire, tels que les réseaux de neurones profonds, où il peut être difficile d'expliquer comment les caractéristiques individuelles influencent les prédictions du modèle. Ils permettent d'interpréter le comportement local du modèle et d'expliquer pourquoi il prend une décision spécifique pour un exemple donné.

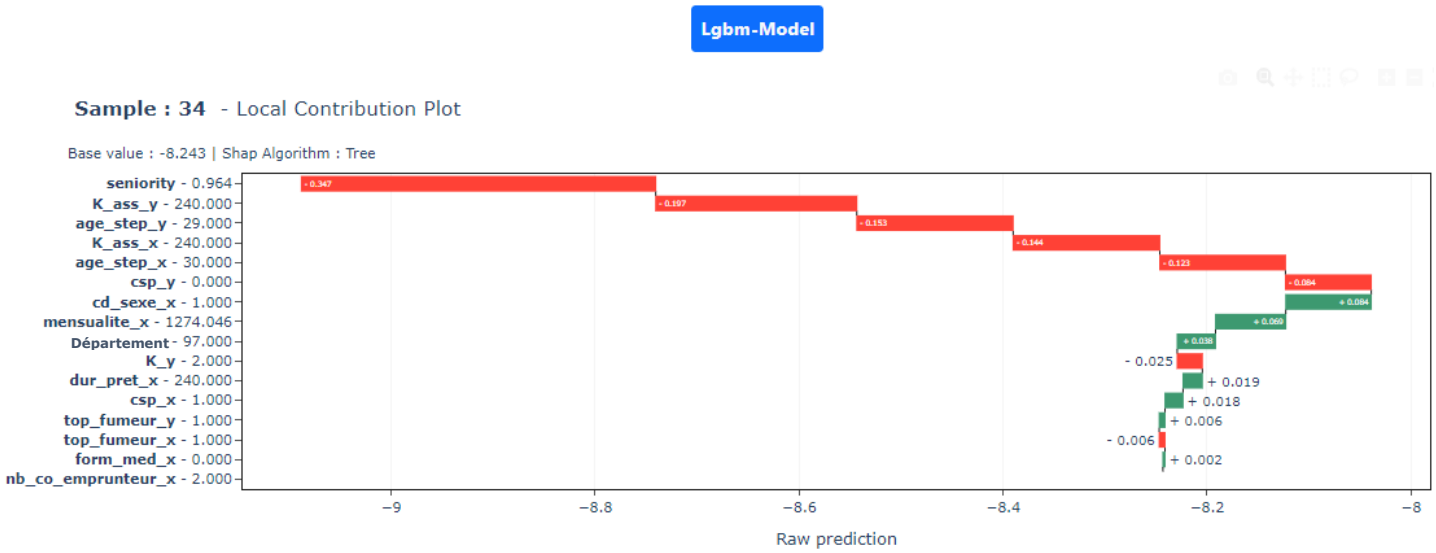


Figure 92 - Vérification locale de la prédiction de la fréquence de l'individu X dans $M_{c(x,y)}$

Dans ce graphique, nous remarquons que les variables expliquent le modèle de manière cohérente. En effet, nous voyons que la contribution de l'âge face à la mortalité est croissante tandis que le fait de s'abstenir de fumer contribue négativement à la mortalité. Cependant, nous relevons une incohérence suite à l'observation d'une influence négative des non-fumeurs face au décès de leur conjoint.

Lorsque nous analysons la structure de l'échantillon d'apprentissage, nous remarquons un déséquilibre de la mortalité face au statut fumeur/abstinent. Le problème est que lors de l'échantillonnage, nous ne pouvons pas faire en sorte d'équilibrer les données en fonction de toutes les possibilités sous-jacentes. De plus, il ne faut pas perdre de vue le fait que la modélisation se fait en couple, et que l'échantillonnage s'exécute par dossier, ce qui rend le contrôle de ce paramètre quasi-impossible.

Pour contrer ce problème, nous avons décidé de diminuer la taille de l'échantillon test et d'échantillonner puis d'utiliser les mêmes paramètres optimaux que précédemment afin d'entrainer de nouveau le modèle. Le résultat est sans équivoque. Nous remarquons que le biais lié à l'échantillonnage a été corrigé et la variable top_fumeur_y agit bien positivement lorsque celle-ci indique le label d'un fumeur (soit la valeur 0).

D'autre part, nous avons entrainé le modèle sur le nouvel échantillon d'entrainement en utilisant une validation croisée afin de comparer les deux modélisations et d'interroger leur stabilité.

Cette correction a été appliquée même dans les parties précedents cette section.

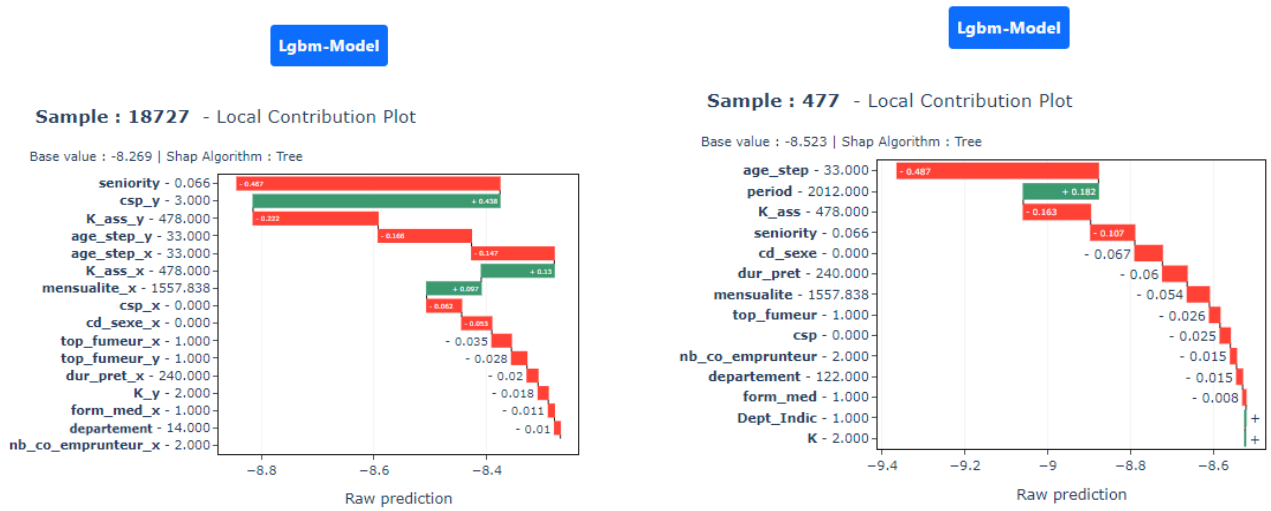


Figure 93- Contribution des variables de l'individu X dans le modèle $M_{c(x,y)}$ et M_i

Dans ces graphiques, nous comparerons les contributions des variables du même individu X aux deux types de modélisations. A savoir le modèle en couple $M_{c(x,y)}$ et le modèle par tête M_i.

Hormis quelques variables, nous remarquons que là plus part des variables agissent dans le même sens dans les deux modélisations.

Toutefois, nous voyons que la mensualité agit négativement dans M_i tandis qu'elle indique un risque plus important dans $M_{c(x,y)}$. Ceci est cohérent ave les formalités médicales passées pat cet assuré. En effet, celles-ci sont légères. De plus, ces informations couplées avec la catégorie socio-professionnelle du conjoint semblent pertinentes. En effet, la catégorie du conjoint est la plus risquée du portefeuille. Le modèle en couple semble tenir compte de la corrélation entre le montant assuré et la catégorie du conjoint.

Enfin, le résultat de $M_{c(x,y)}$ est cohérent avec le montant total assuré (K-ass) qui semble lui aussi agir positivement sur le risque décès.

5.2.5.2 Pour l'entrée en incapacité de travail

Comparaisons des modèles

Dans cette partie, nous comparons les deux modèles en couple avec ajout des variables et individuel, soit la différence entre $M_{c(x,y)}$ et M_i.

Lgbm-Model

metrics	On Task data (100%)
filter data...	
Deviance Explained	0.03659
RMSE	0.07643
MAE	0.00365
TotaPred/TotalObs	0.92388
Gini Index (Normalized)	0.41999
MSE	0.00584
Average Deviance	0.02603
AveragePred	0.00176

Figure 94 - Les métriques du modèle $M_{c(x,y)}$

Lgbm-Model

metrics	On Task data (100%)
filter data...	
Deviance Explained	0.03978
RMSE	0.06973
MAE	0.00659
TotaPred/TotalObs	1.01414
Gini Index (Normalized)	0.38292
Average Deviance	0.03659
MSE	0.00486
AveragePred	0.00334

Figure 95 - Les métriques du modèle M_i.

En effet, lorsque nous analysons le résultat global des deux modélisations, nous remarquons que le fait de limiter l'entrainement à la base couple avec ajout des variables explicatives du conjoint n'a presque pas d'effet sur l'adéquation concernant le risque en incapacité de travail.

Si les erreurs globales sont moindres pour le modèle limité à la base couple, la variable $\frac{Total\_pred}{Total\_obs}$ qui mesure le taux d'adéquation globale est plus proche pour M_i que pour $M_{c(x,y)}$ qui sous-estime la sinistralité. Il serait donc plus prudent de choisir la modélisation individuelle dans laquelle la base est plus grande, ce qui signifie que la modélisation serait plus stable.

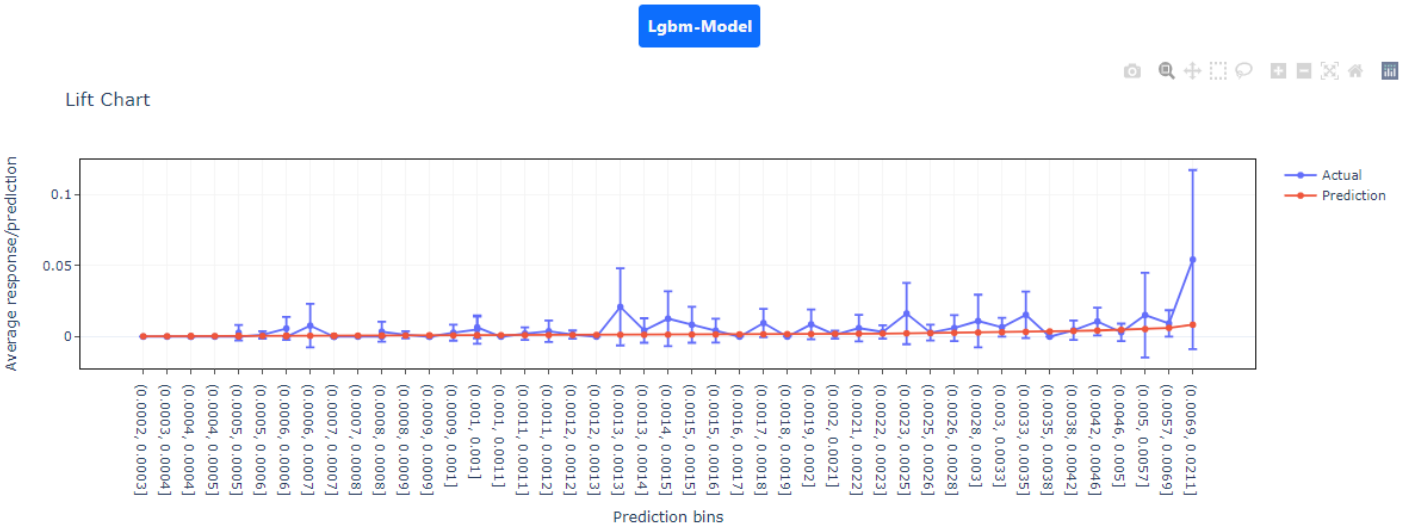


Figure 96 - Le graphique des erreurs de $M_{c(x,y)}$

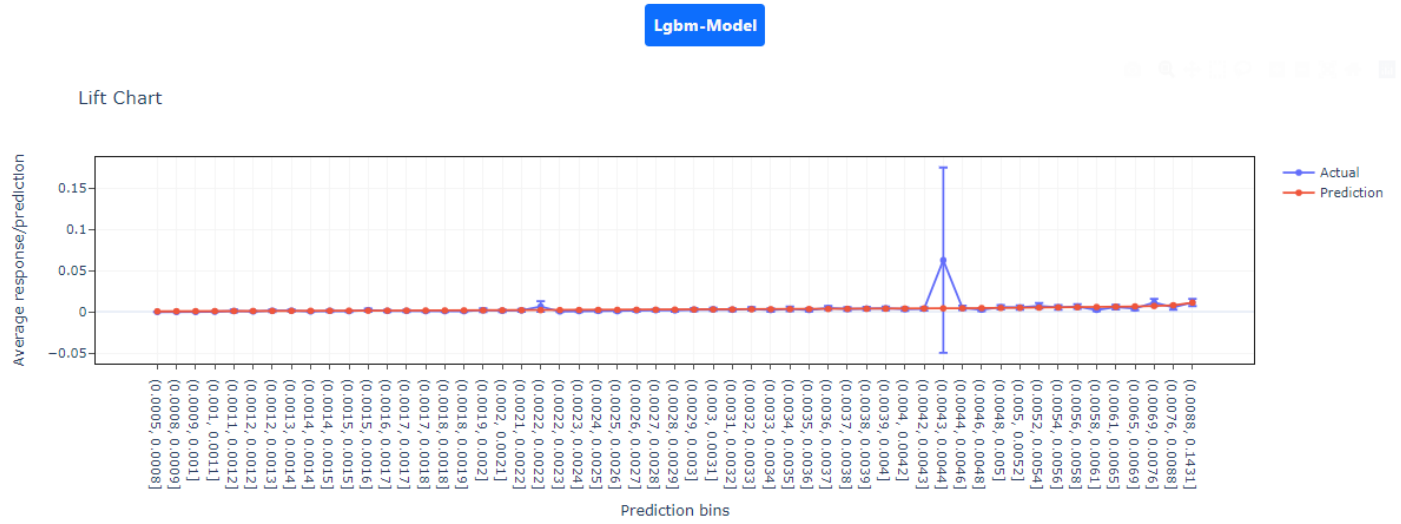


Figure 97 - le graphique des erreurs de M_i

Lorsque nous traçons les lift_chart, nous remarquons que $M_{c(x,y)}$ commet plusieurs erreurs tout au long des prédiction sans jamais dépasser la sinistralité historique. Tandis que le modèle M_i commet une erreur importante mais cette erreur ne concerne que quelques points.

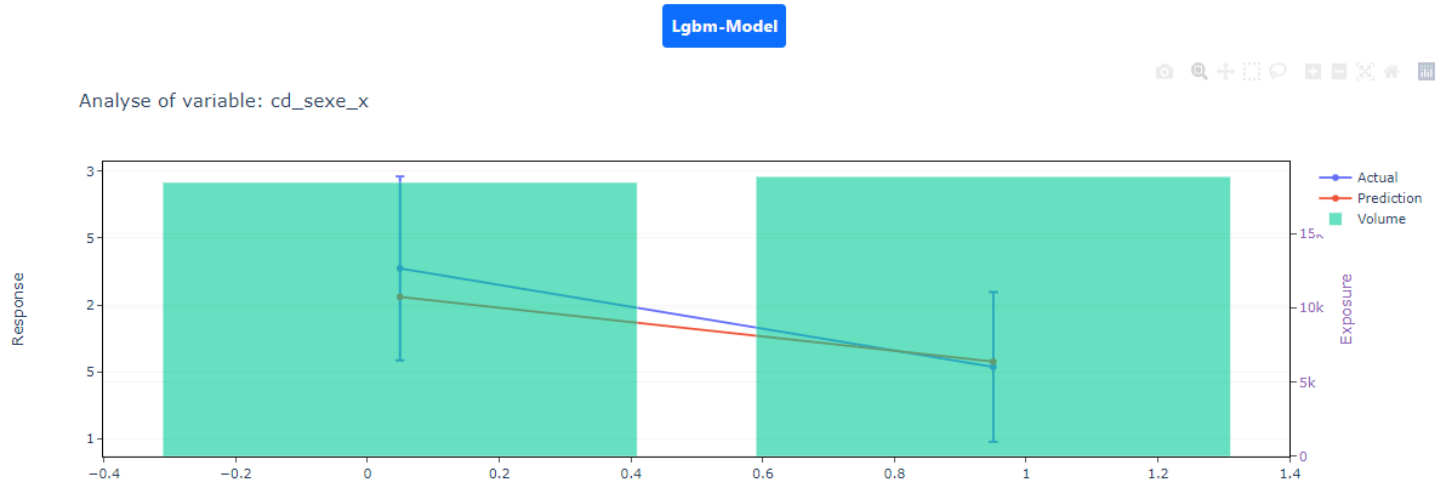


Figure 98 - La sinistralité en fonction du sexe dans $M_{c(x,y)}$

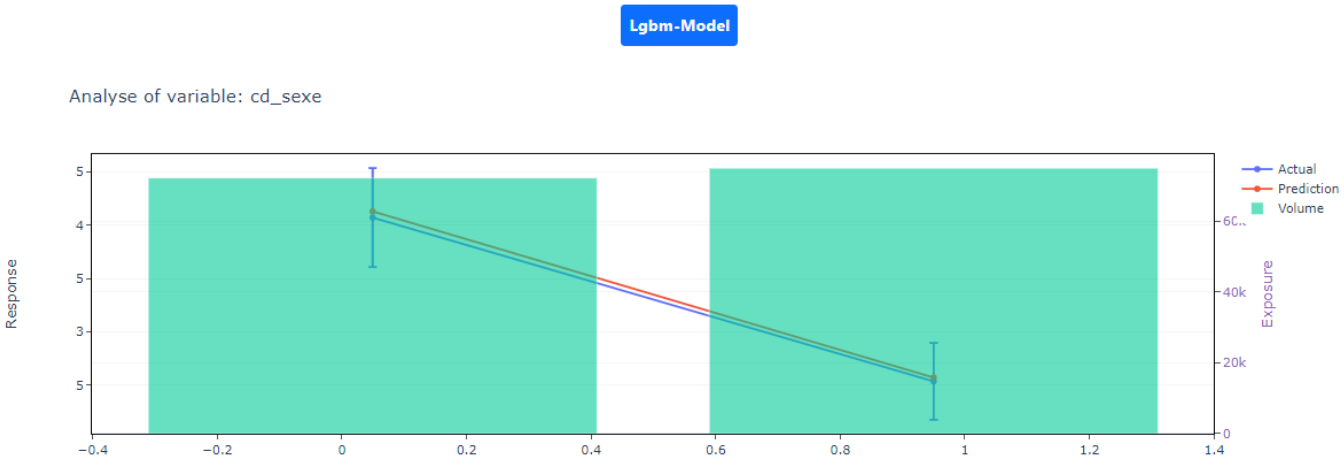


Figure 99 - La sinistralité en fonction du sexe dans M_i

L'analyse de la modélisation en fonction du sexe semble avoir une tendance assez cohérente pour les deux types de modélisations. En effet, les graphiques ci-dessus montrent que le taux d'incidence en arrêt de travail des femmes est supérieur à celui des hommes.

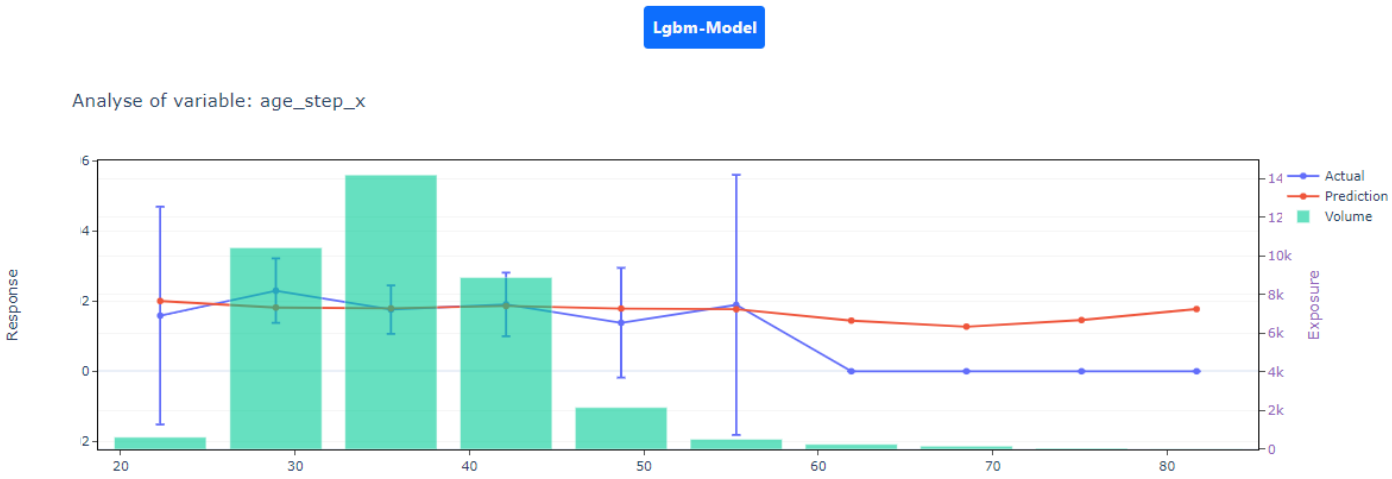


Figure 100 - $M_{c(x,y)}$ en fonction de l'âge de l'assuré

Figure 101 - M_i en fonction de l'âge de l'assuré

Dans l'ensemble, M_i surpasse $M_{c(x,y)}$ pour prédire la sinistralité en fonction de l'âge. Or, il s'agit d'une variable très importante dans la détermination du taux d'incidence pour l'entrée en arrêt de travail. La population du portefeuille étant jeune (exposition faible à partir de 50 ans), le modèle commet malgré tout un biais important. Cependant, c'est un axe à prendre en compte car la modélisation pourrait se stabiliser avec le vieillissement du portefeuille, notamment avec le recul de l'âge de la retraite.

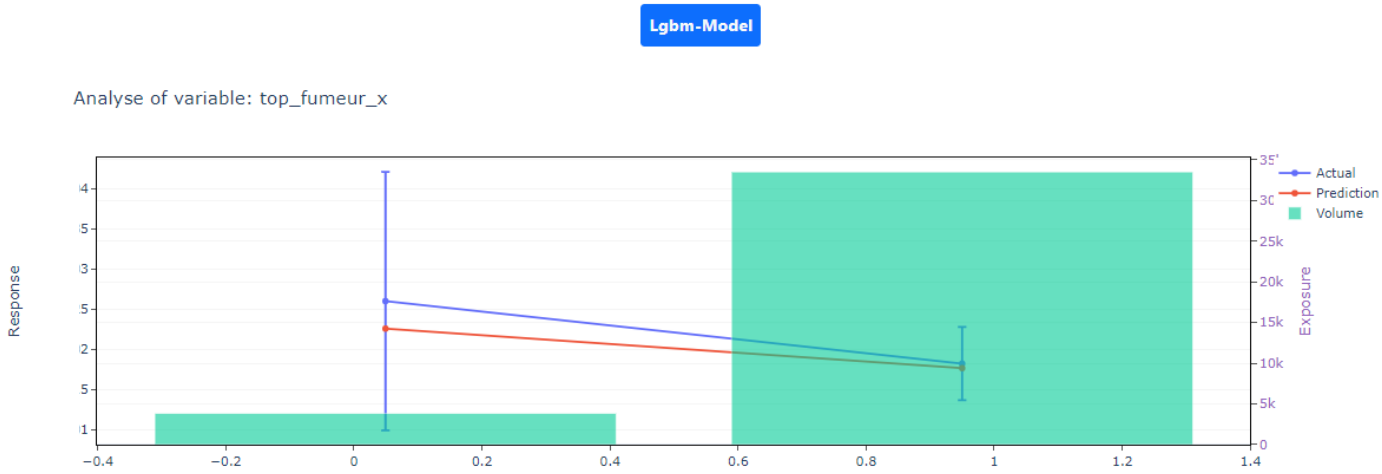


Figure 102 - $M_{c(x,y)}$ en fonction du tabagisme de l'assuré

Figure 103 - M_i en fonction du tabagisme

Les fumeurs en couple ont un taux d'incidence plus important que les non-fumeurs. La modélisation en couple permet de mieux rendre compte de l'influence du tabac sur le taux de hasard. En effet, l'erreur commise par M_i est plus importante que celle commise par $M_{c(x,y)}$. De plus, M_i semble sous-estimer la sinistralité des fumeurs et surestimer celle des abstinents.

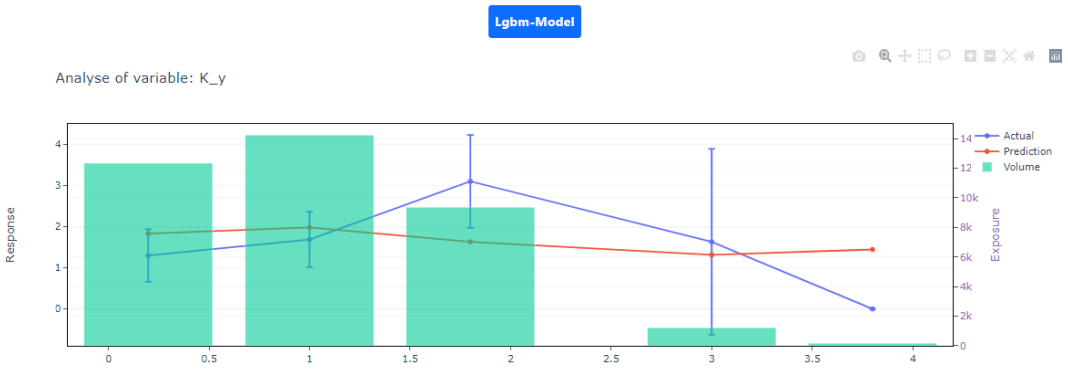


Figure 104 - modélisation en fonction de capital assuré avec $M_{c(x,y)}$.

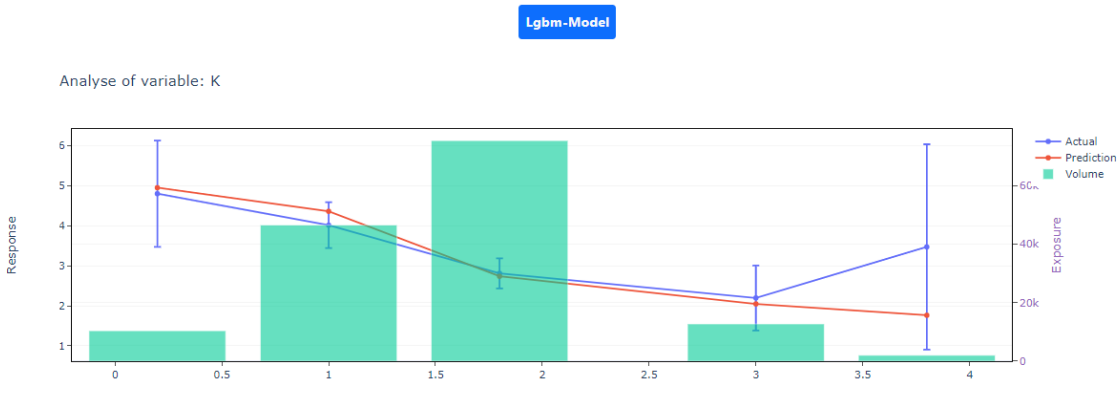


Figure 105 - Modélisation en fonction du capital assuré M_i.

L'étude du taux de hasard (ITT) en fonction du capital total assuré semble proposer une tendance moins risquée pour les individus à forts capitaux. Ceci rejoint les résultats du bilan de la **Drees** sur l'état de santé de la population en France, publié le 21/09/2022. En effet, cette étude témoigne de l'inégalité face aux soins en invoquant une déclaration de diabète 2.8 fois plus important chez les personnes dont le

niveau de vie est inférieur au premier décile. Cette étude a également observé un risque de 1.5 fois plus importante concernant les maladies cardio-vasculaires dans le même sens.

Pour détecter cette tendance, la modélisation en couple avec ajout des variables du conjoint semble mieux adéquate. Ceci permet d'estimer la sinistralité avec une meilleure précision.

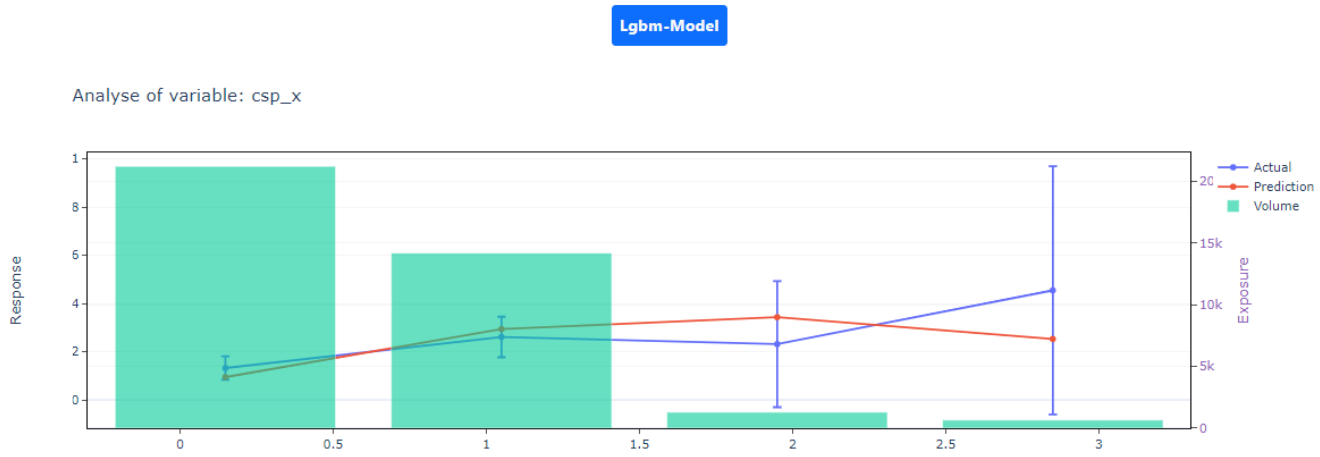


Figure 106 - $M_{c(x,y)}$ en fonction de la catégorie socio-professionnelle de la personne assurée

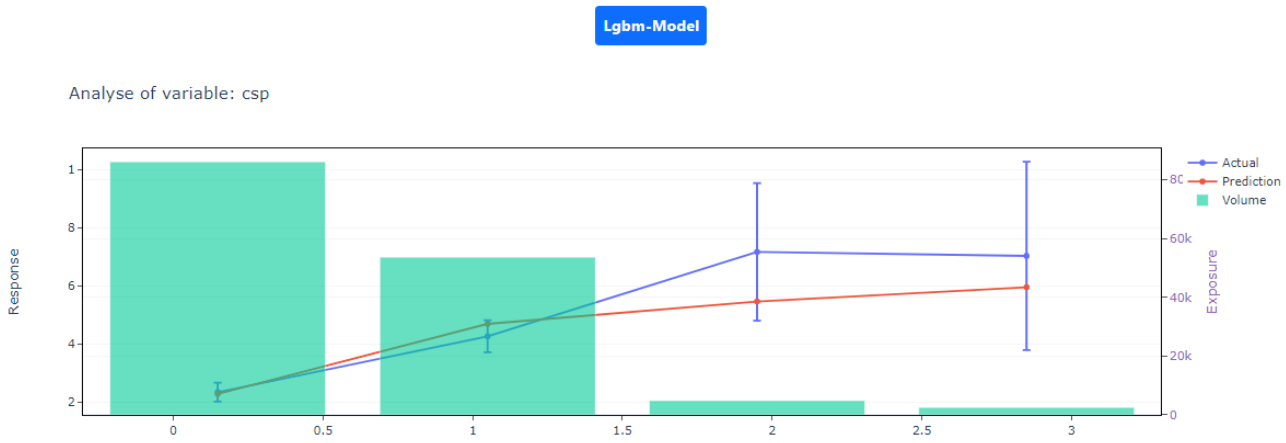


Figure 107 - M_i en fonction de la catégorie socio-professionnelle de la personne assurée

L'importance de la catégorie socio-professionnelle pour analyser l'entrée des individus en arrêt de travail réside dans son rôle crucial en tant que facteur discriminent. En effet, comme nous pouvons le lire dans l'étude « health in the workplace » publiée le 19/07/2019 par Département for Work & Pension, il a été prouvé que les employés travaillants dans l'administration publique, l'éducation et la santé sont plus susceptibles de déclarer une maladie longue durée contrairement aux personnes qui travaillent dans la finance ou la construction. De plus ce graphique extrait de la même étude indique que certaines maladies ne sont pas dispersées de manière uniforme au sein des différentes professions.

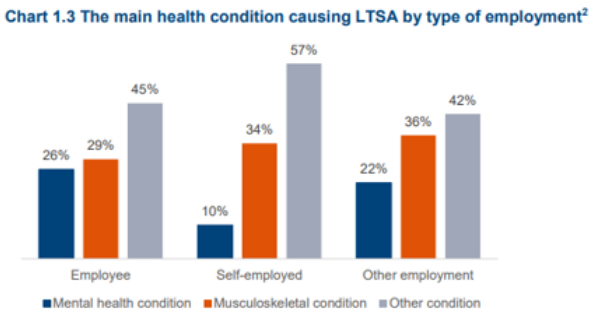


Figure 108- Dispersion des maladies en fonction de la catégorie professionnelle

Nous remarquons ainsi que les maladies mentales sont plus fréquentes chez les employeurs que chez les travailleurs libéraux. Si nous revenons à la modélisation, nous remarquons que $M_{c(x,y)}$ semble décrire la sinistralité de manière plus fidèle selon ce segment. Même si $M_{c(x,y)}$ s'éloigne de l'historique en catégorie 3 et 4, il semble mutualiser entre ces deux catégories, tandis que M_i sous-estime la sinistralité des deux catégories les plus risquées.

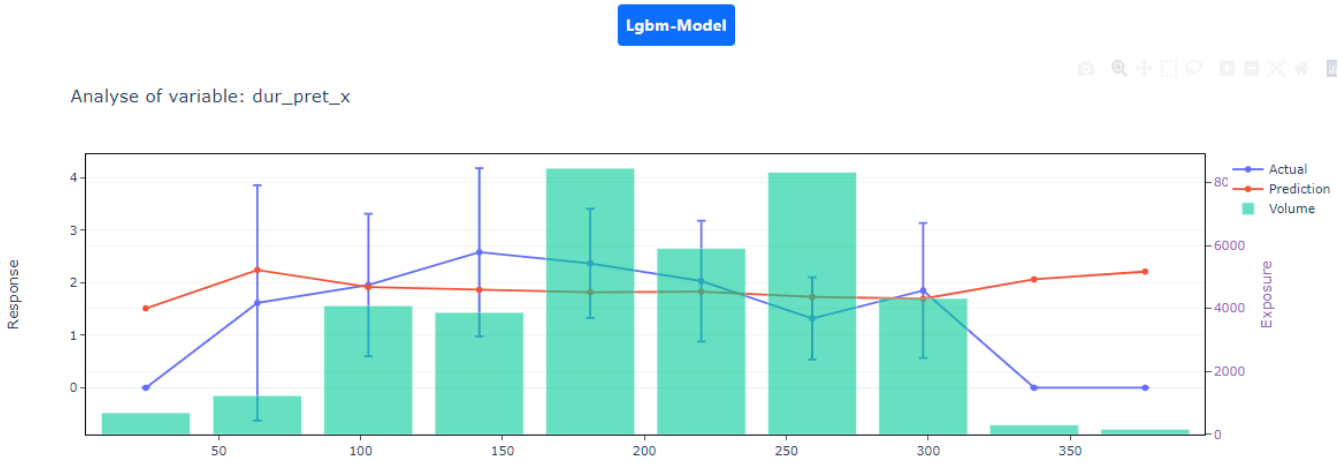


Figure 109 - $M_{c(x,y)}$ en fonction de la durée de prêt

Lors de la souscription, les individus sont moins risqués (dû à la sélection médicale). De plus, l'utilisation d'une franchise de six mois permet d'observer un taux de hasard nul durant les premiers mois. Passé ces délais, les individus commencent à bénéficier d'indemnités. Nous pouvons ainsi observer un premier pic au bout de six ans d'ancienneté.

Nous observons un second pic au bout de 12 ans. Entre 8 et 12 ans le modèle $M_{c(x,y)}$ semble sous-estimer la fréquence pour l'arrêt de travail tandis que celui-ci surestime cette fréquence après 12 ans.

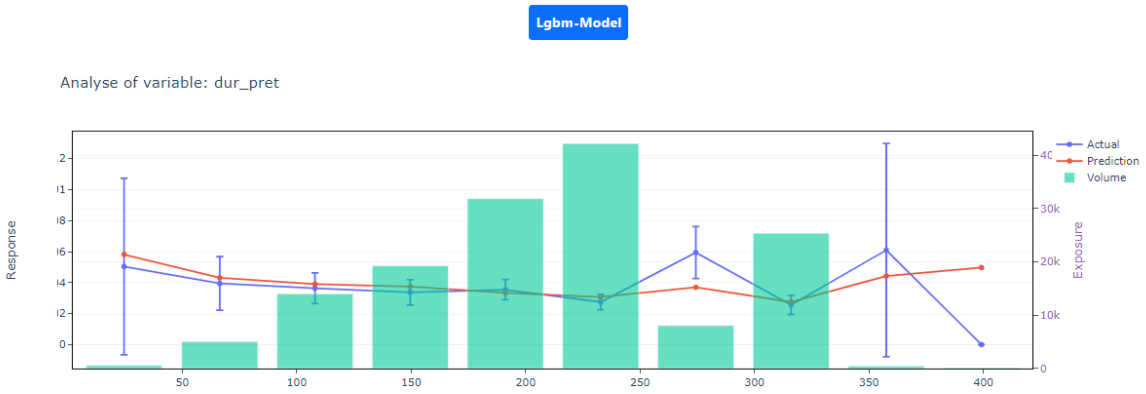


Figure 110 - M_i en fonction de la durée du prêt

M_i présente une très bonne adéquation de 0 à 18 ans d'ancienneté. Cependant, il sous-estime la sinistralité vers les durées les plus longues due à la faible exposition. Cela peut être corrigé par la sinistralité historique comme vu précédemment.

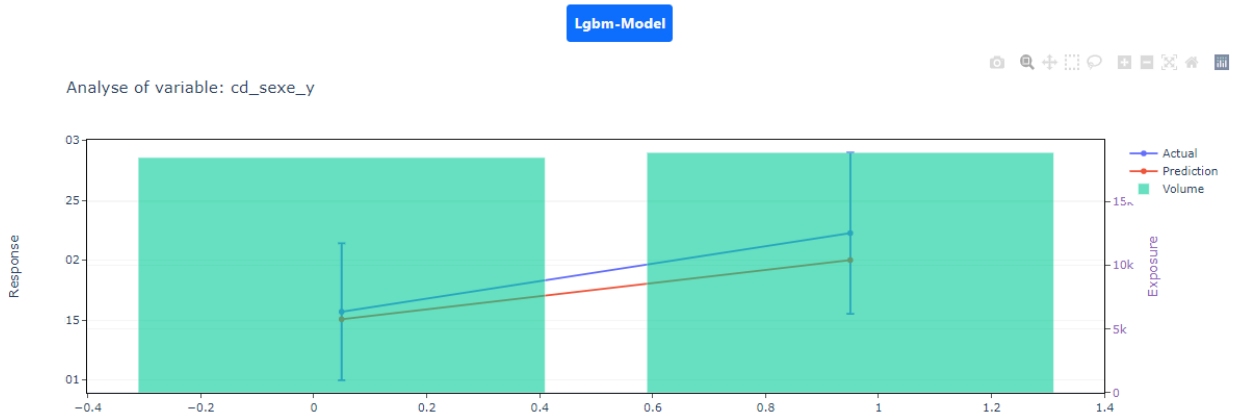


Figure 111 - M_i en fonction du sexe du conjoint de l'assuré

Tout comme pour le modèle décès, $M_{c(x,y)}$ semble capter l'effet du sexe du conjoint sur le taux d'incidence en arrêt de travail.

Sexe 1	Sexe 2	Taux de hasard	Proportion
F	F	0.20%	1%
H	H	0.07%	1%
F	H	0.60%	99%
H	F	0.35%	99%

Figure 112- taux de hasard en fonction de la composition du couple.

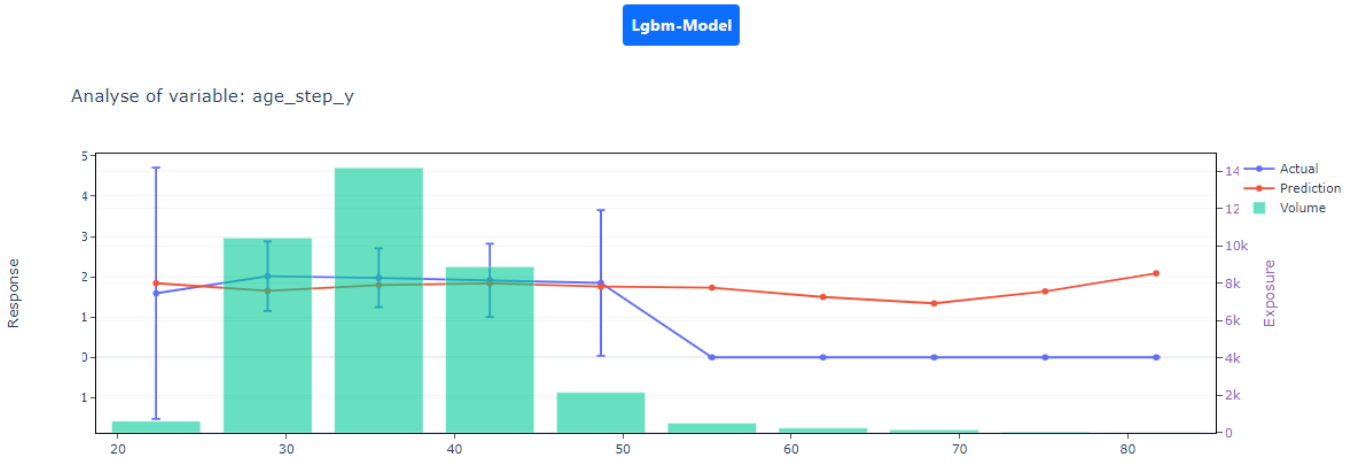


Figure 113 - $M_{c(x,y)}$ en fonction de l'âge du conjoint

La modélisation en fonction de l'âge du conjoint offre une excellente adéquation dans les zones où l'exposition est importante. Néanmoins, le modèle $M_{c(x,y)}$ semble surestimer la sinistralité après 50 ans. Dans ces zones où le nombre d'individus est très faible, nous pouvons incomber cette erreur à la parcimonie des données.

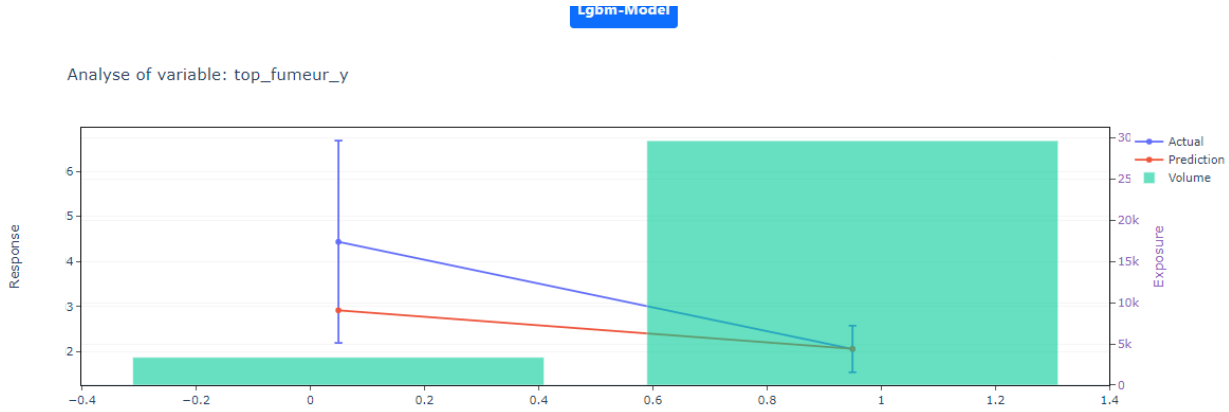


Figure 114 - $M_{c(x,y)}$ en fonction du tabagisme du conjoint

Les variables historiques indiquent que les conjoints fumeurs agissent positivement sur l'entrée en incapacité de travail de la tête assurée, tandis que le modèle indique capte un peu moins cette propriété. Il faut savoir que la proportion de fumeur n'est pas pertinente, le taux d'entrées en incapacité qui est de extrêmement faible, enfin le portefeuille est très jeune (12 années d'observation). Il est de ce fait extrêmement difficile de capter le signal avec des données aussi parcimonieuses. Cependant la modélisation LightGBM arrive à capter une partie de l'information. En effet, nous pouvons observer l'effet du tabagisme passif qui est susceptible d'accroitre les maladies cardio-vasculaires de 25%. (Étude « Epidémiologie » en 2005).

Si nous regardons la sinistralité en fonction du tabagisme au sein des couples, nous remarquons que les taux de hasard décrivent une tendance aggravante sur les non-fumeurs en fonction du tabagisme du conjoint sur le taux d'incidence en arrêt de travail.

D'autre part, une étude INSEE menée par Mélanie Vanderschelden et parue le 01/05/2006 indique que le choix du conjoint est influencé par sa catégorie socio-professionnelle. D'autre part, lorsque nous regardons la composition des couples du portefeuille, nous remarquons que les tendances sont à l'harmonisation des unions (cf. Figure 169 - Taux de hasard moyen en fonction de la CSP du conjoint).

Figure 115- Contribution de la catégorie socio-professionnelle du conjoint sur le taux de hasard

Cette étude semble se confirmer si nous regardons les valeurs de Shapeley de la variable indiquant la catégorie socio-professionnelle du conjoint assuré. En effet, les conjoints aux métiers les plus risqués agissent positivement sur le taux d'incidence en arrêt de travail de la tête assurée.

Si nous combinons ces informations avec le graphique ci-dessous, nous déduisons que l'environnement familial est un élément non négligeable dans l'analyse de la sinistralité en incapacité de travail.

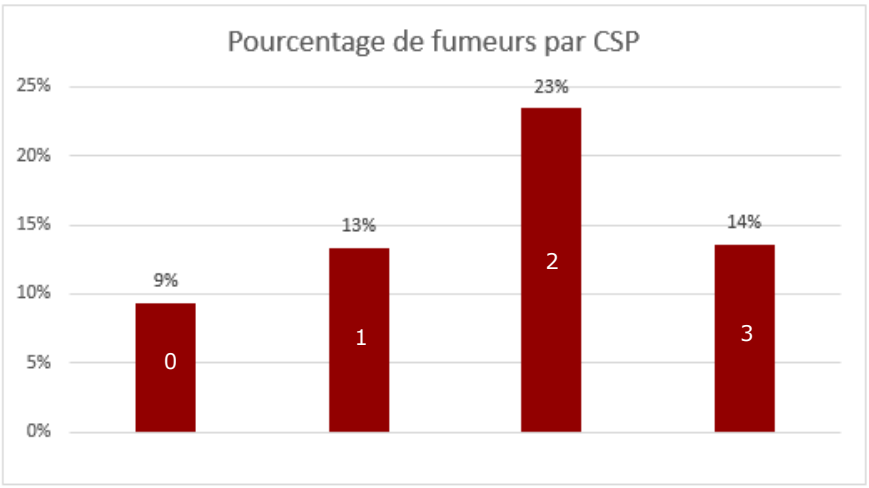


Figure 116- dispersion des fumeurs par catégorie socio-professionnelle

Nous remarquons ainsi que les catégories 0 et 1 manifestent moins d'intérêt pour le tabac tandis que la catégorie 3 affiche un score significativement élevé. Celle-ci se rapproche plus du taux national qui est de 25% selon les statistiques de santé publique.

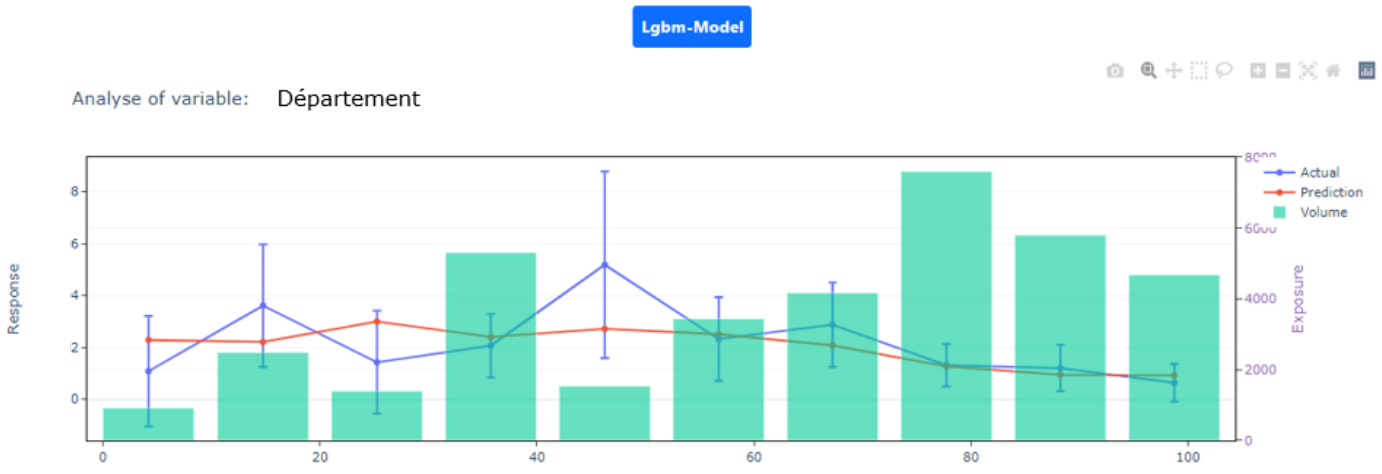


Figure 117- Taux d'incidence en arrêt de travail pour $M_{c(x,y)}$

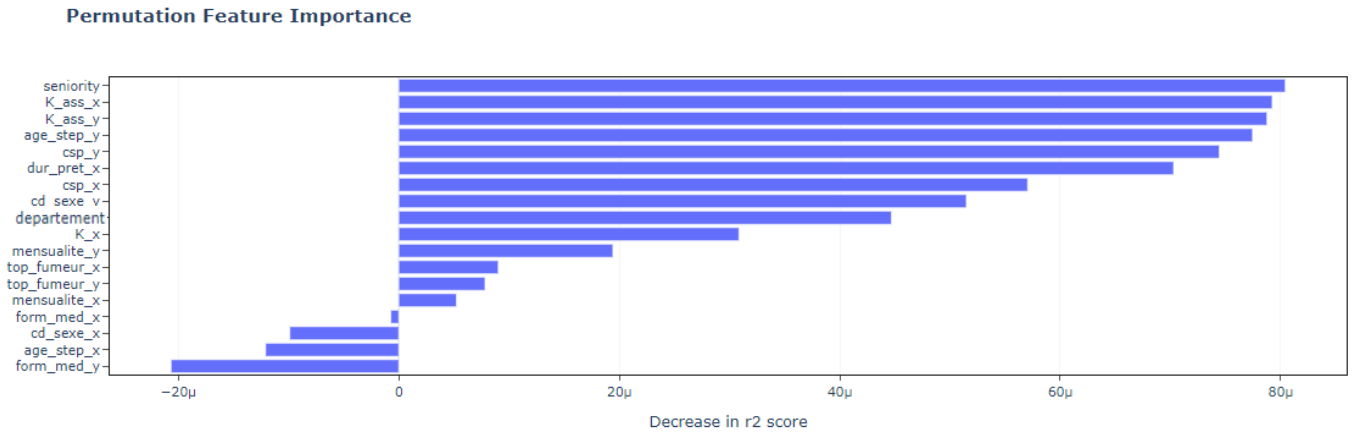


Figure 118 - Graphique permutation importance $M_{c(x,y)}$

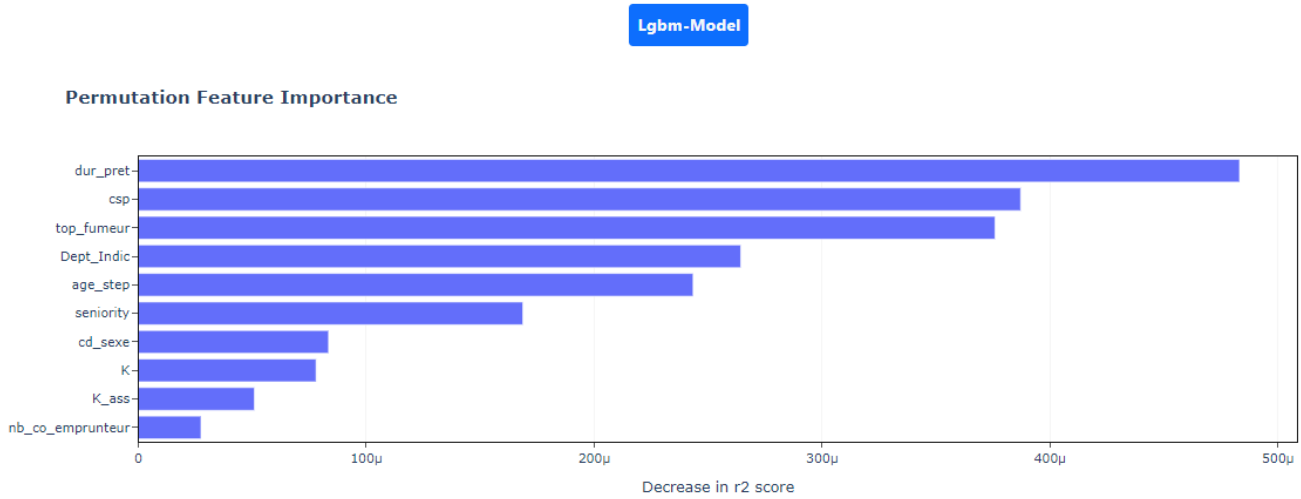


Figure 119 - Graphique feature importance M_i

Dans la modélisation $M_{c(x,y)}$, nous remarquons que l'âge de l'individu est beaucoup moins important que dans la modélisation M_i . Si nous comparons ceci aux statistiques descriptives, nous comprenons que la dispersion de la population du portefeuille « couple » par âge est beaucoup moins hétérogène que dans le portefeuille de la population globale.

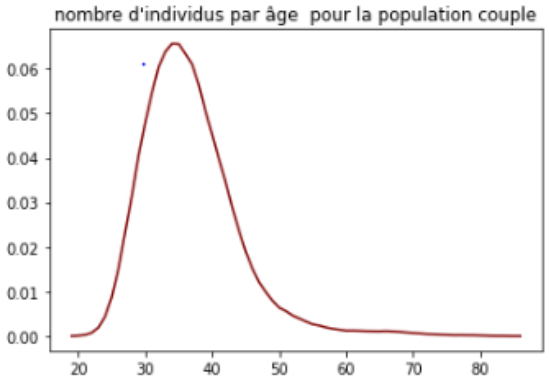


Figure 120 - Dispersion de la population du portefeuille par âge

Lgbm-Model

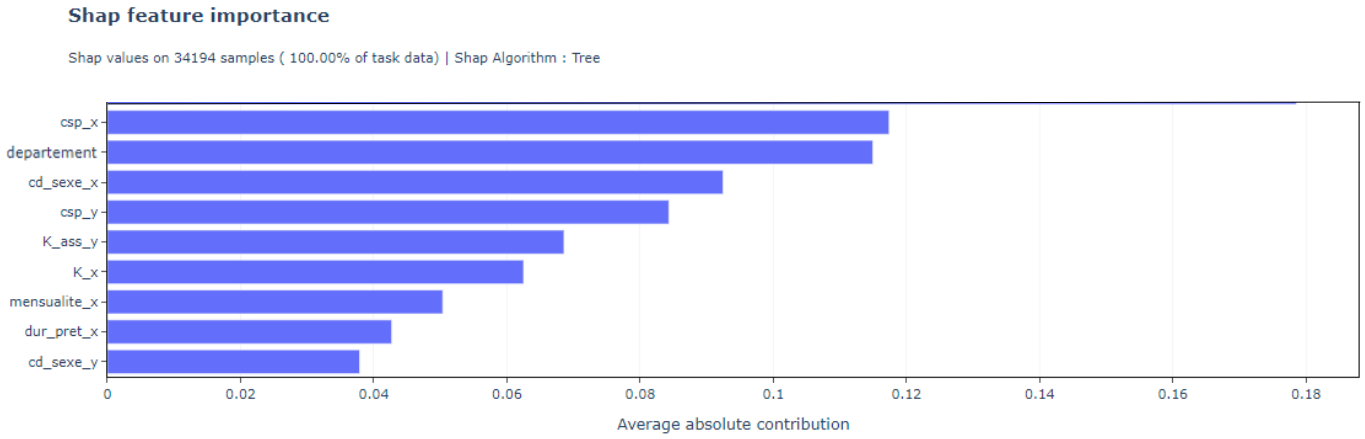


Figure 121- Intensité des variables sur la prédiction du modèle en couple

Lgbm-Model

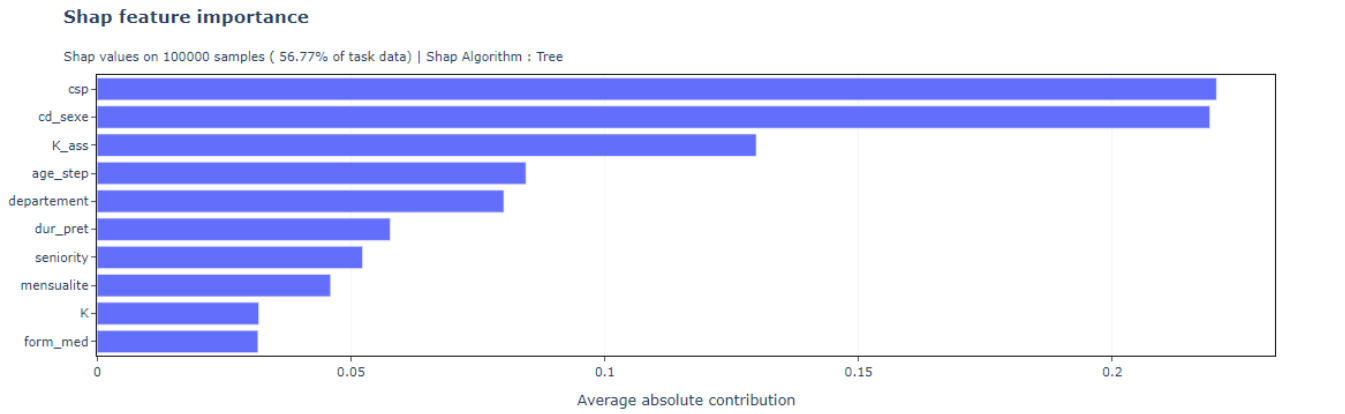


Figure 122- Intensité des variables sur la prédiction à partir du modèle individuel

Nous remarquons en calculant les valeurs de Shapley que dans les deux échantillons, la catégorie socio-professionnelle contribue de la même manière au le taux d'incidence en arrêt de travail pour les deux populations. Cependant, la variable sexe ainsi que l'âge ont été devancés par le capital assuré ainsi que

le département des individus en couple. En effet, comme nous le voyons dans la présentation ci-dessous, les écarts entre les taux des individus en couple en fonction du sexe sont moins importants en comparaison à la population totale. De la même manière, la variance de la sinistralité en fonction des âges agit dans le même sens.

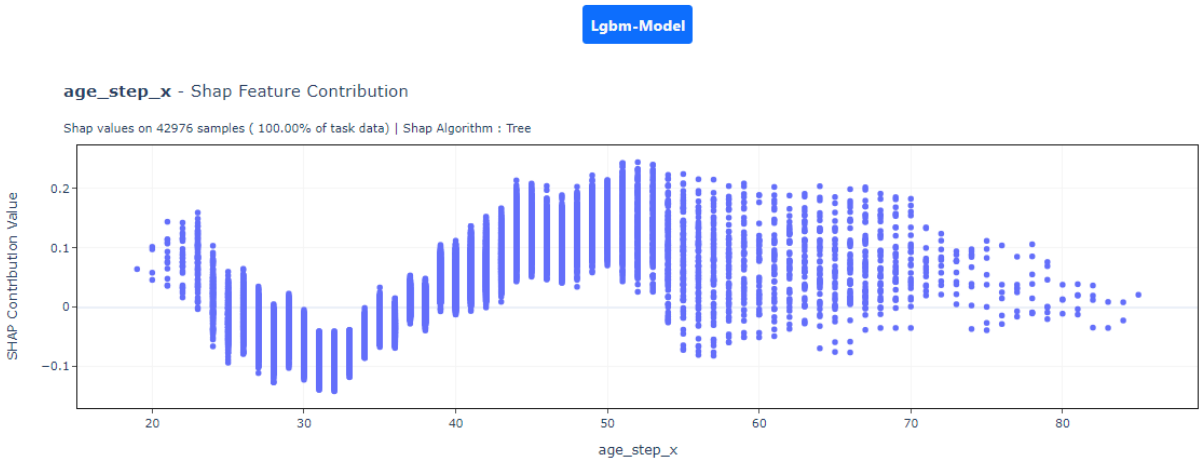


Figure 123- l'influence de l'âge dans le modèle en couple

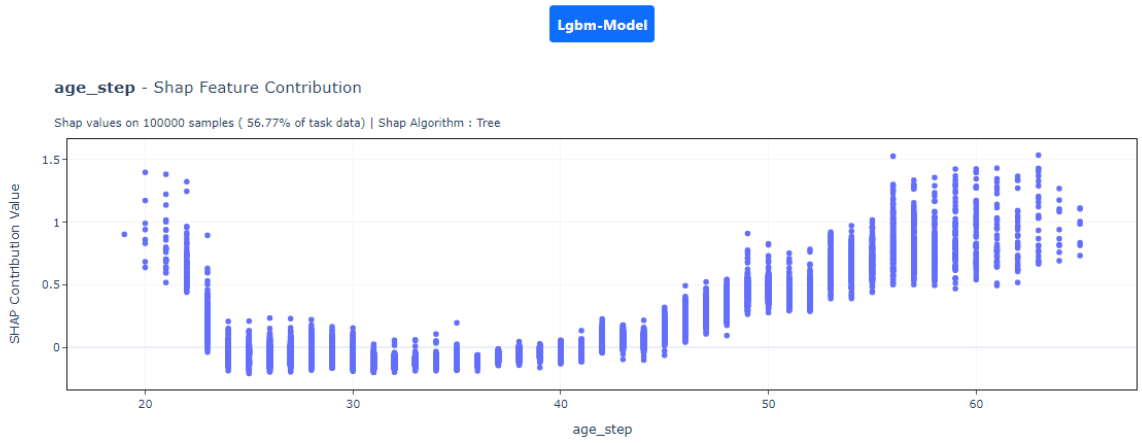


Figure 124- l'influence de l'âge dans le modèle individuel

Ces deux graphes montrent que la variance des valeurs de Shapley en fonction de l'âge est plus importante au sein de la population globale.

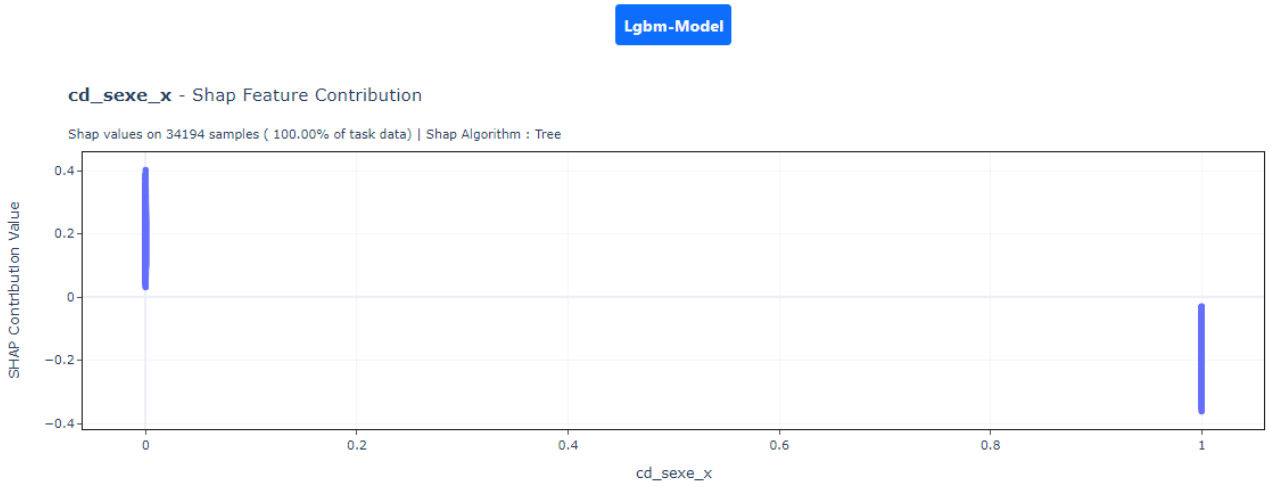


Figure 125- l'influence du sexe dans le modèle en couple

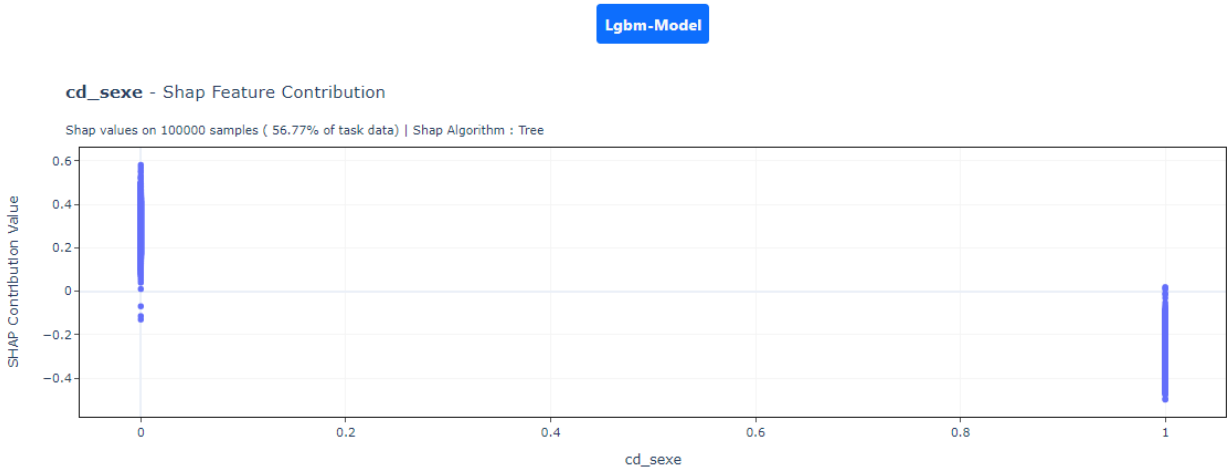


Figure 126 - l'influence du sexe dans le modèle individuel

De la même manière que l'âge, nous remarquons que l'influence du sexe agit dans le même sens sur le taux de hasard en incapacité de travail.

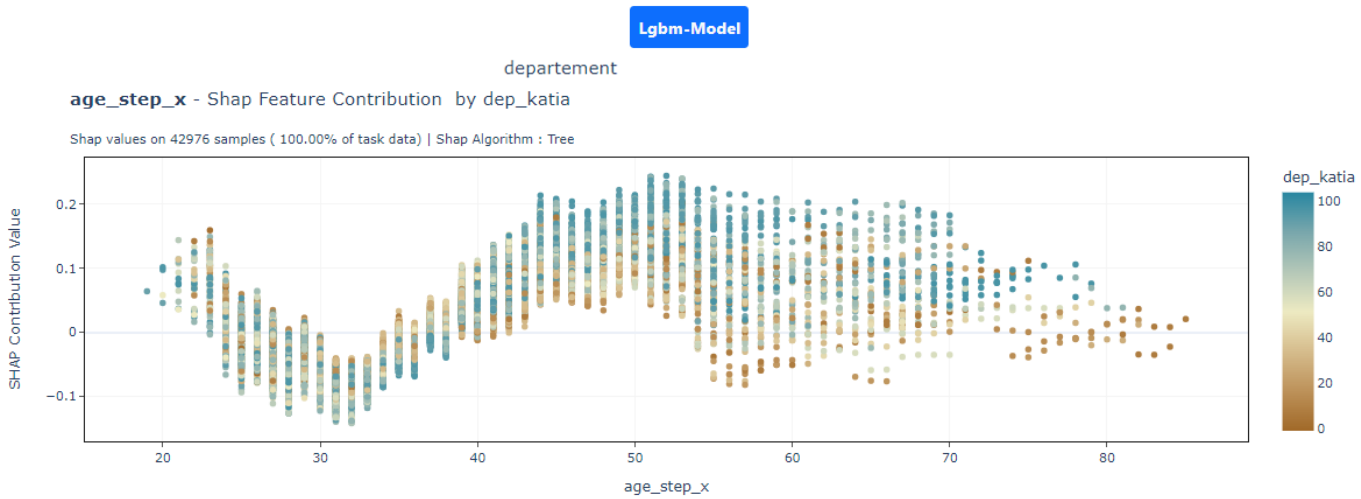


Figure 127- l'influence de l'âge par département dans le modèle en couple pour l'incapacité de travail

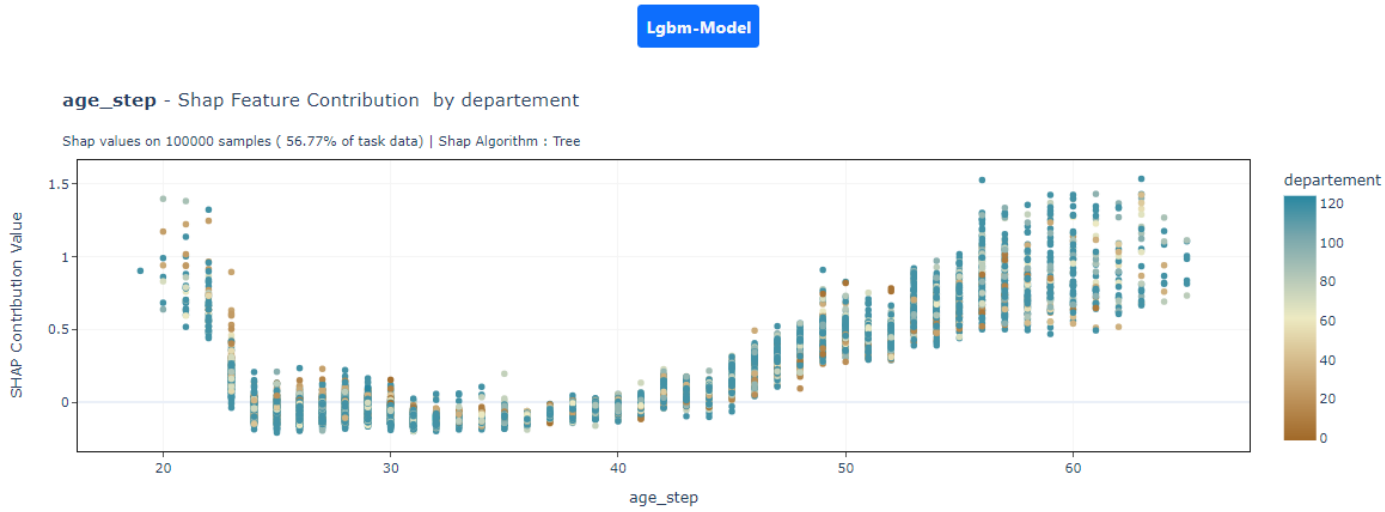


Figure 128 - L'influence de l'âge par département dans le modèle individuel pour l'incapacité de travail

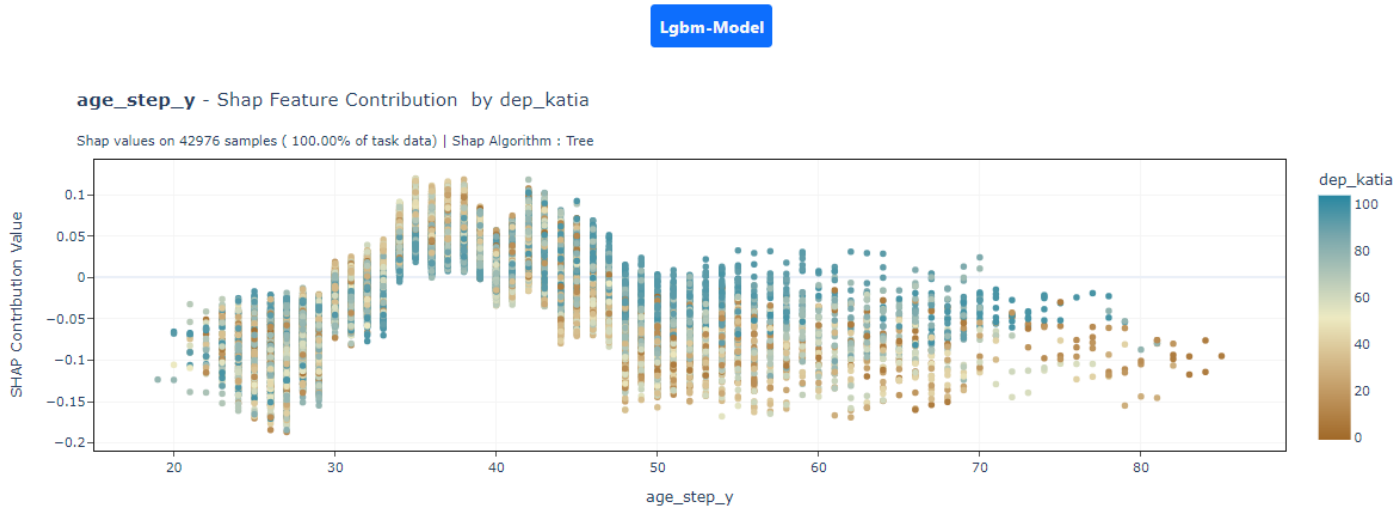


Figure 129- l'influence de l'âge du conjoint par département dans le modèle couple pour l'incapacité de travail

Si nous traçons les valeurs de Shapley en fonction de l'âge par département, nous remarquons que dans le modèle pour couple, l'âge influence le taux de décès instantané. Mais l'influence varie en fonction de l'âge et du département. A titre d'exemple, une personne jeune vivant en Ile de France aura un taux de survie plus important qu'une personne jeune vivant dans le Cher. En revanche, ce constat est inversé pour les populations les plus agées.

Dans le modèle global ou individuel, il l'influence est moins marquée.

D'autre part, nous voyons également que les tendances de la personne principale et celles de son conjoint sont antagonistes entre 20 et 30 ans. Plus le conjoint principal est jeune, plus nous observons une décroissance franche du taux de hasard dans cette zone, tandis que l'inverse est observé lorsque nous regardons plutôt le groupe des conjoints.

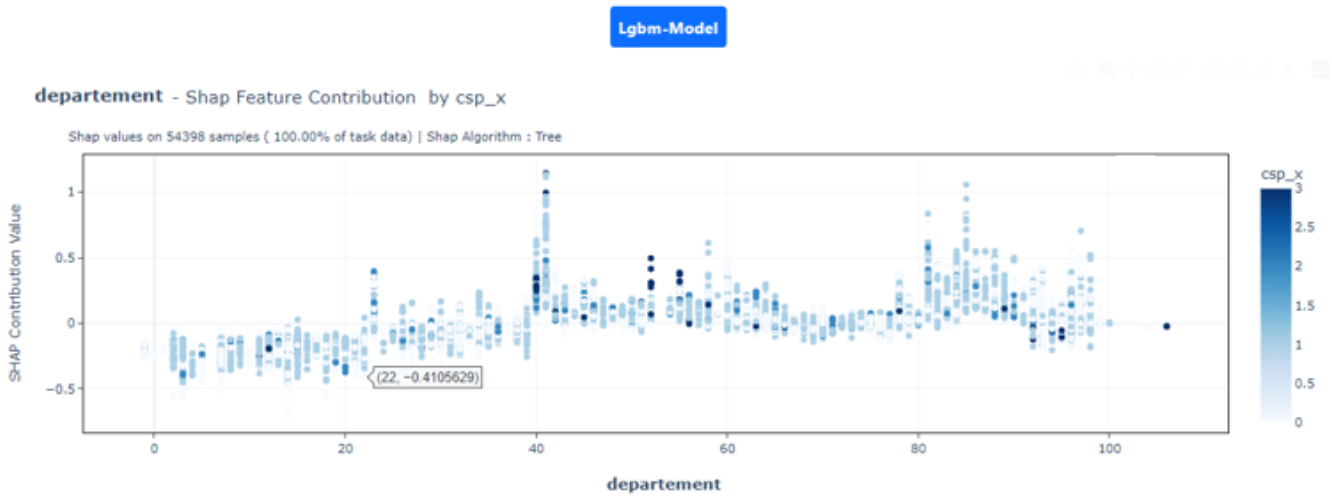


Figure 130- l'influence et corrélation entre la CSP et le département dans le modèle couple pour l'incapacité de travail

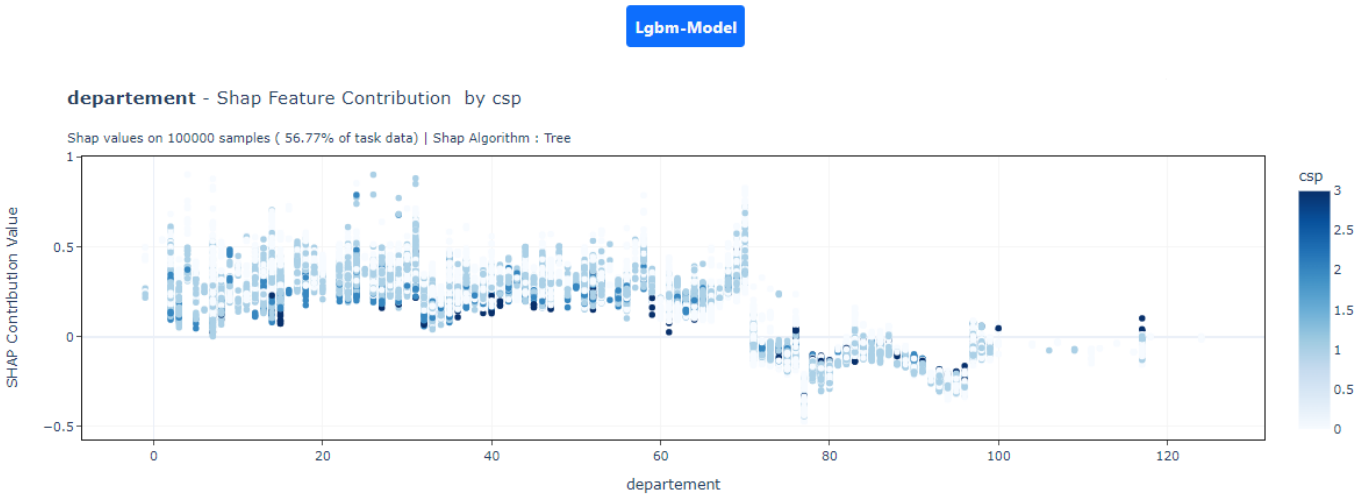


Figure 131- l'influence et corrélation entre la CSP et le département dans le modèle individuel pour l'incapacité de travail

Les graphiques ci-dessus indiquent que la catégorie socio-professionnelle 3 (CSP 3) influence positivement sur le taux d'entrées en incapacité parmi les couples, tandis qu'elle est plus dans l'inhibition de ce sinistre lorsque nous considerons la population globale. Nous remarquons d'autre part que la CSP est beaucoup plus corrélée au département dans le modèle en couple. Cela est cohérent avec les statistiques descriptives qui classe la catégorie 3 comme la plus risquée en terme de taux d'incidence en arrêt de travail. Le modèle en couple capte donc plus le risque en fonction des CSP.

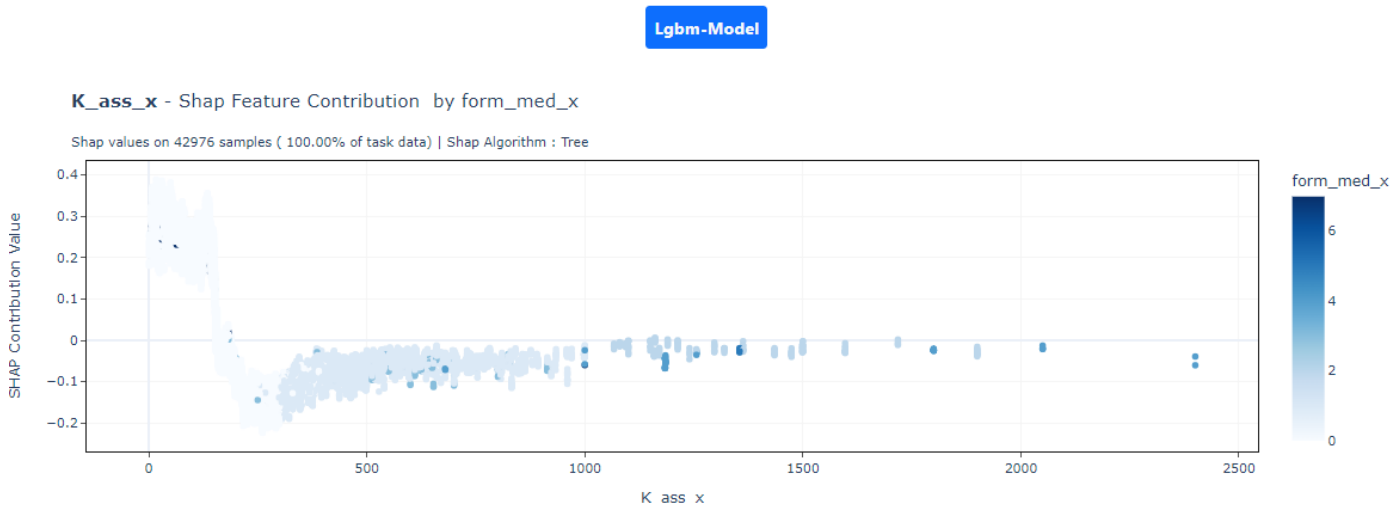


Figure 132- Corrélation entre capital assuré et formalités médicales et l'influence de cette combinaison sur l'incapacité de travail en couple

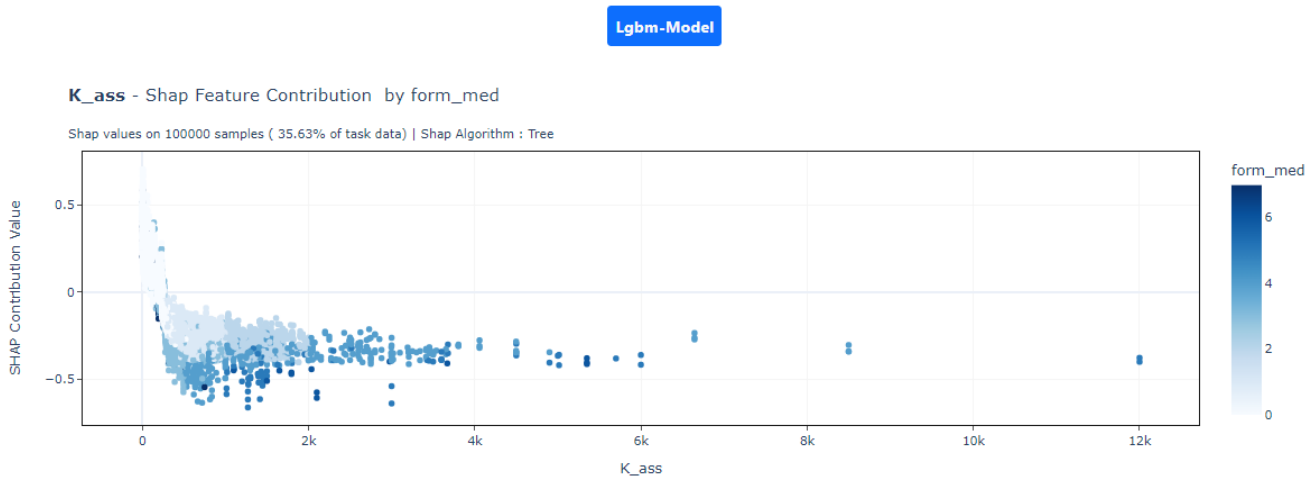


Figure 133- Corrélation entre capital assuré et formalités médicales et l'influence de cette combinaison sur l'incapacité de travail (individuel)

De manière générale, le dégradé de couleur indique que plus le capital assuré est important, plus les formalités médicales imposées à l'assuré sont importantes. De plus les valeurs de Shapley indiquent que les formalités les plus importantes agissent négativement sur le taux de hasard. Ceci est cohérent avec les études internes indiquant que la vérification de la santé des individus du portefeuille concorde avec une acceptation plus prudentes des risques souscrits. Ceci a pour conséquence l'observation de capital assuré parmi les variables explicatives du taux de décès instantané.

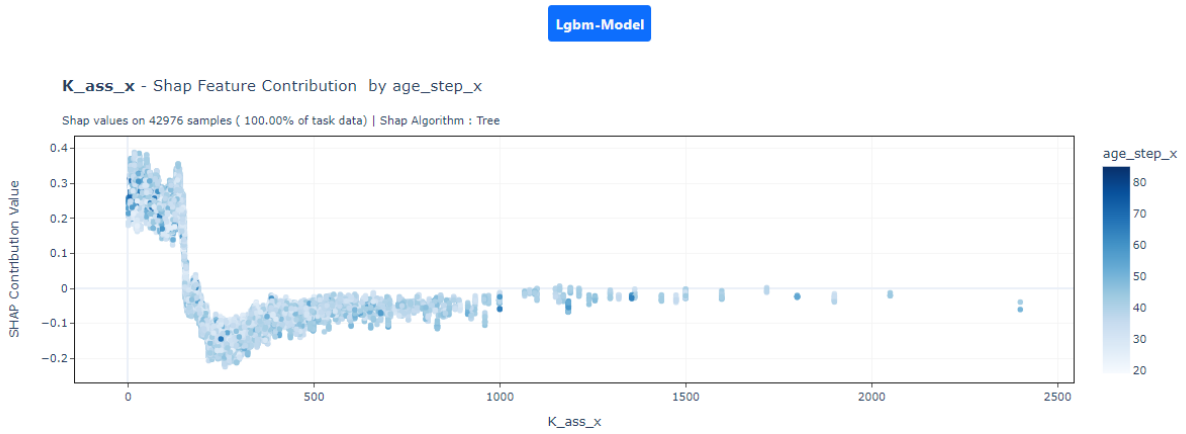


Figure 134- Corrélation entre capital assuré et âge médicales et l'influence de cette combinaison sur l'incapacité de travail (couple)

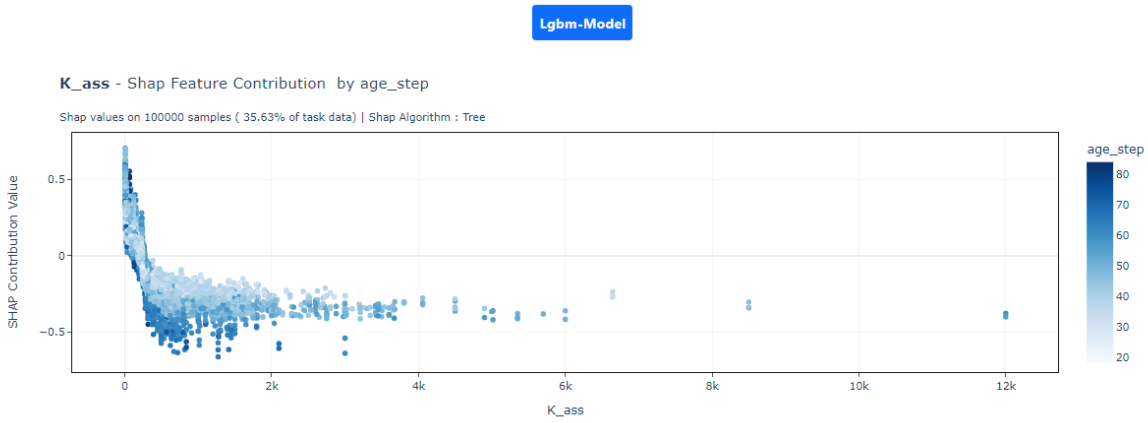


Figure 135- Corrélation entre capital assuré et âge médicales et l'influence de cette de cette combinaison sur l'incapacité de travail (individuel)

Lorsque nous analysons le capital assuré, nous remarquons que les niveaux d'emprunts sont beaucoup plus dispersés au sein de la population totale qu'au sein de l'échantillon limité aux couples. Dans les deux cas, les individus les plus âgés ont tendance à emprunter des capitaux plus faibles que les individus les plus jeunes. Ce constat permet de mettre en évidence que les personnes jeunes ont certes plus de capital mais moins de risque, contrairement aux personnes les plus âgées.

La structure du portefeuille en fonction de ces segments est toutefois différente. Nous remarquons une structure plus homogène au sein des couples d'assurés. C'est pour cela que le modèle qui apprend sur l'ensemble de la population donne plus d'importance à la variable K_ass (capital assuré) du modèle ayant appris sur les couples.

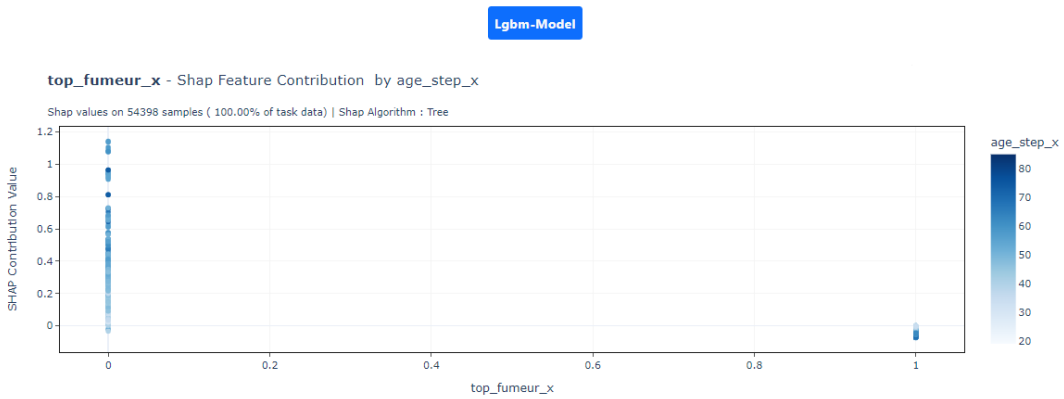


Figure 136- Corrélation entre le tabac et âge médicales et l'influence de cette de cette combinaison sur l'incapacité de travail (couple)

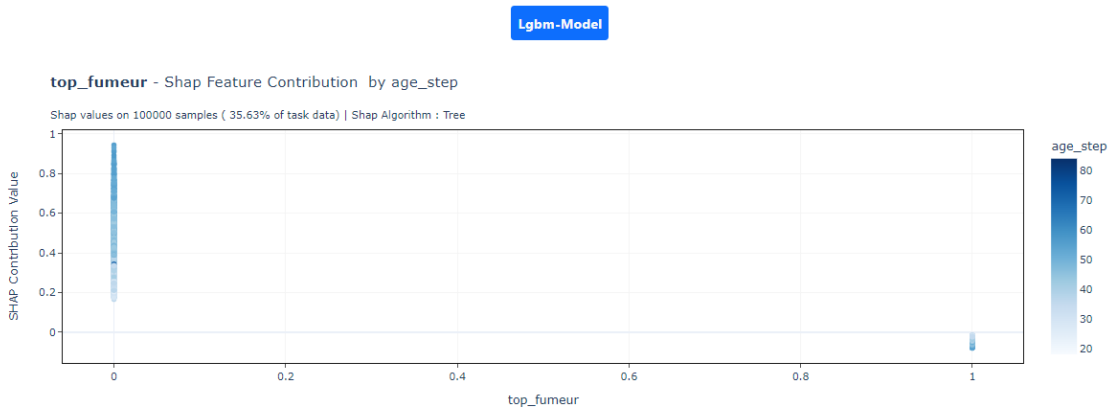


Figure 137- Corrélation entre le tabac et âge médicales et l'influence de cette de cette combinaison sur l'incapacité de travail (individuel)

Les valeurs de Shapley indiquent un taux de hasard accru avec l'âge. Dans le modèle couple tout comme dans le modèle global, elles indiquent une mortalité plus importante si l'individu est fumeur que si celui-ci est non-fumeur. Nous observons d'autre part que l'impact des individus fumeurs est plus fort dans le modèle couple que celui qui modélise les assurés individuellement. Le taux de hasard des abstinents au tabac est toutefois le même dans les deux modélisations.

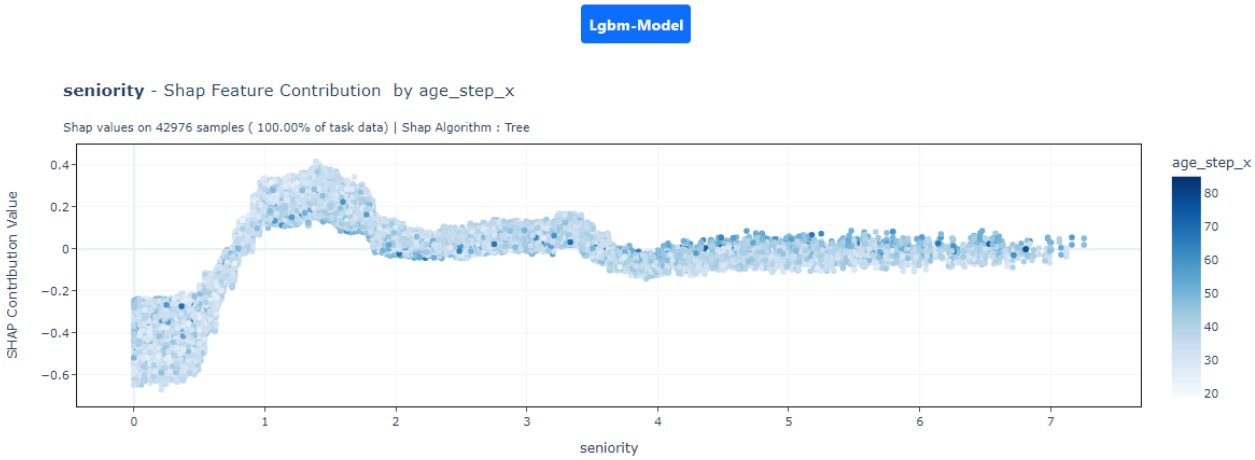


Figure 138- Corrélation entre l'âge et l'ancienneté et l'influence de cette combinaison sur l'incapacité de travail (couple)

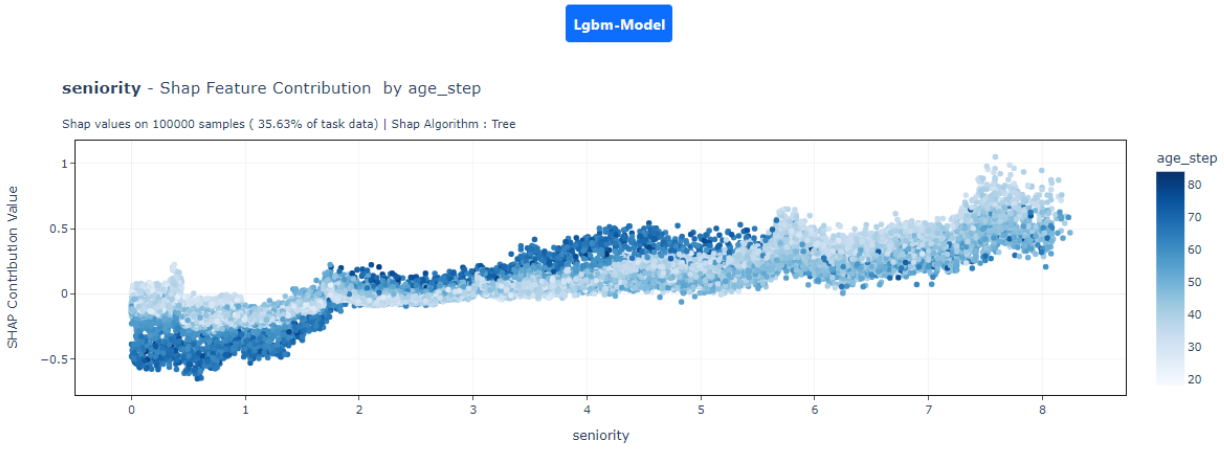


Figure 139- Corrélation entre l'âge et l'ancienneté et l'influence de cette combinaison sur l'incapacité de travail (individuel)

Dans les deux modèles, nous observons une augmentation du taux de hasard en fonction de l'ancienneté de l'individu dans le portefeuille. Si nous regardons cela en fonction de l'âge, nous remarquons que les individus les plus âgés sont moins risqués durant les deux premières années par rapport aux individus les plus jeunes. Le pivot se fait après 2 ans d'ancienneté où nous observons une inversion de la tendance.

Si nous croisons cette information avec le type de formalités médicales exigées aux assurés, nous y trouvons une explication.

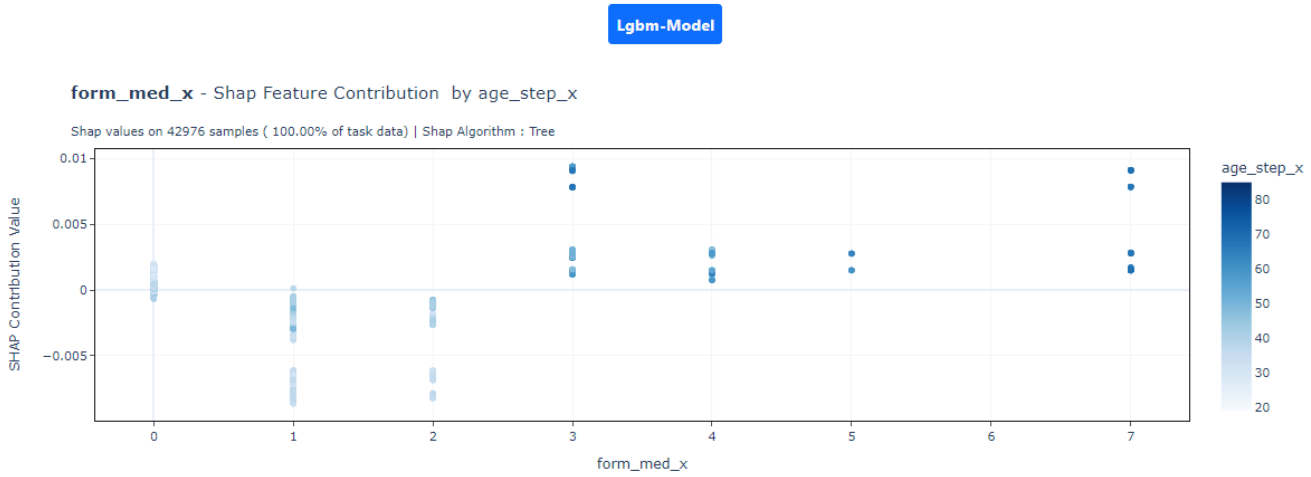


Figure 140- Corrélation entre l'âge et les formalités médicales et l'influence de cette combinaison sur l'incapacité de travail (couple)

En effet les formalités médicales les plus poussées sont plus souvent demandées aux personnes les plus âgées.

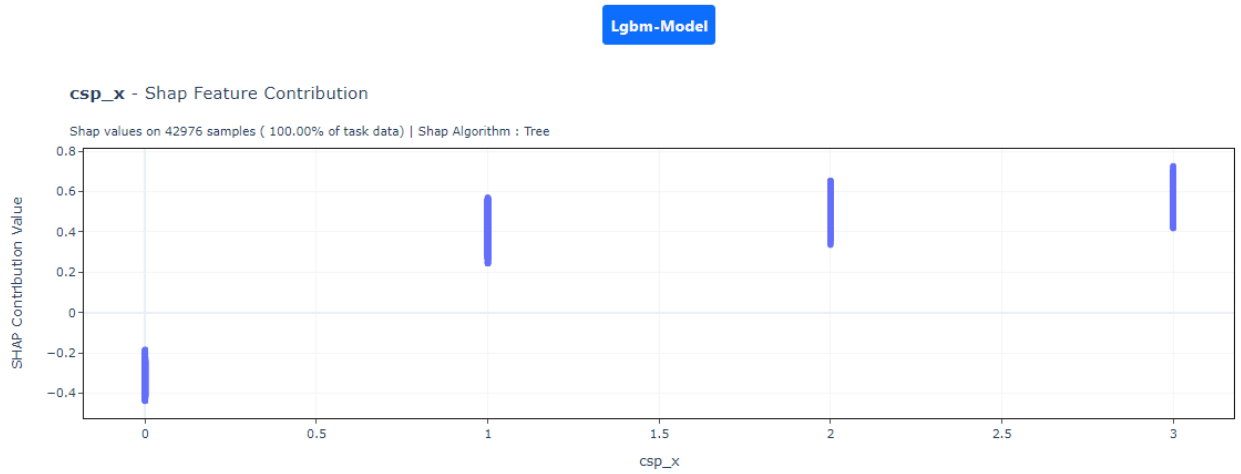


Figure 141- Influence de la catégorie socio-professionnelle sur l'entrée en incapacité de travail (couple)

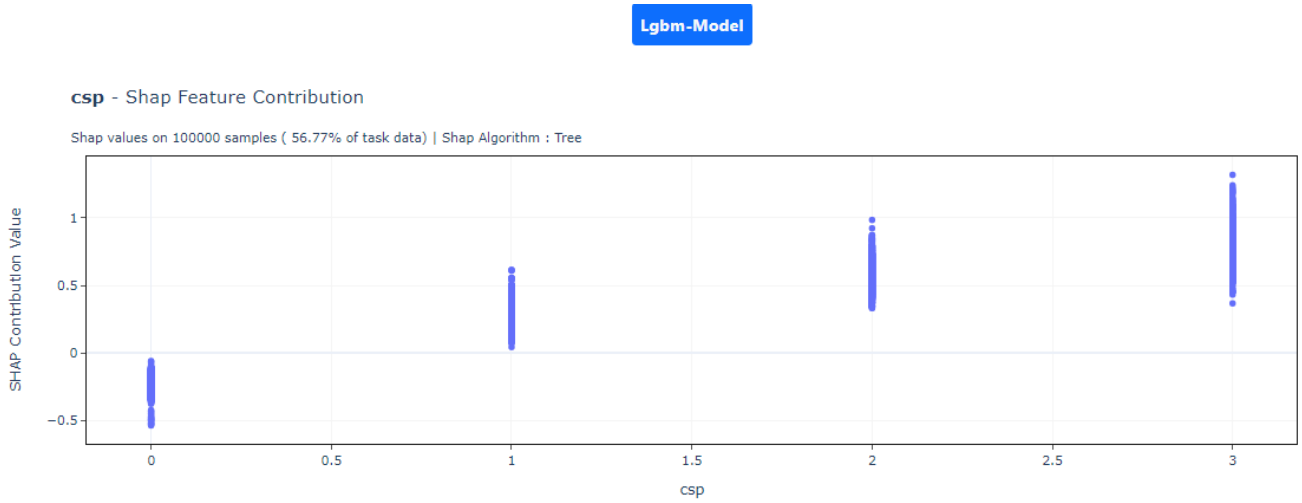


Figure 142 - Influence de la catégorie socio-professionnelle sur l'entrée en incapacité de travail (individuel)

L'influence de la catégorie socio-professionnelle est plus importante dans la population globale que dans la modélisation en couple. En effet, comme observé dans le graphique, certaines catégories sont également influencées par celle du conjoint (cf. Figure 122- Intensité des variables sur la prédiction à partir du modèle individuel)

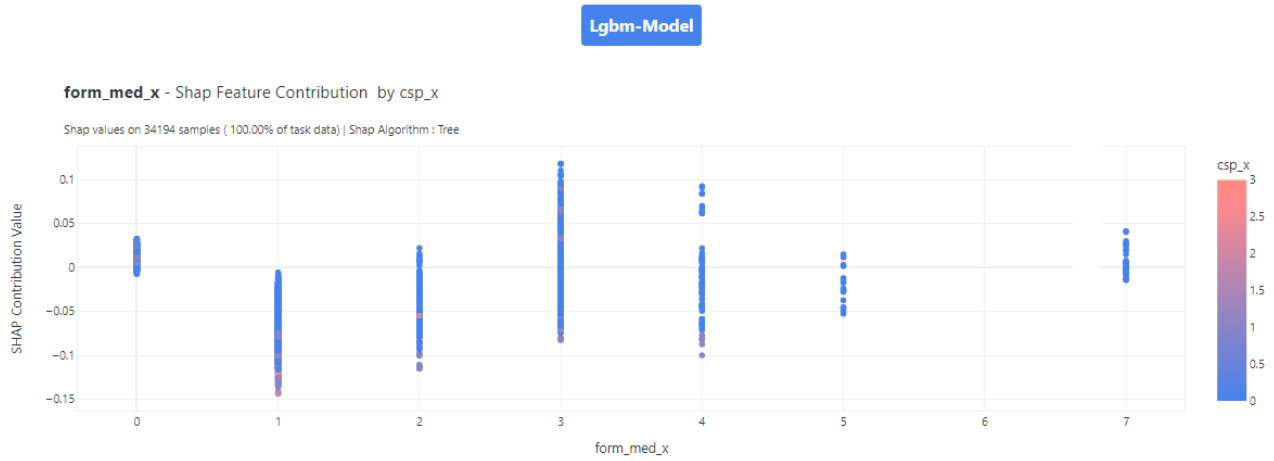


Figure 143 - Lien entre formalités médicales et CSP (couple)

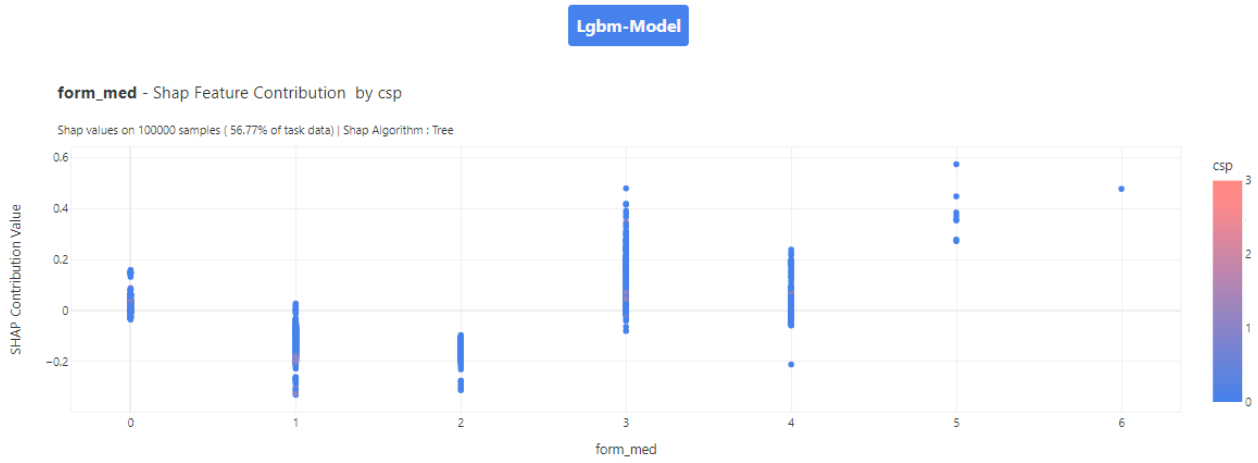


Figure 144 - Lien entre formalités médicales et CSP (individuel)

Si nous regardons les formalités médicales en fonction de la catégorie socio-professionnelle des assurés, nous remarquons que le modèle en couple est moins sensible aux formalités les plus poussées.

Conclusion

La modélisation par l'algorithme LightGBM est cohérente, il peut donc être utilisé pour l'analyse des risques étudiés.

Local contribution plot

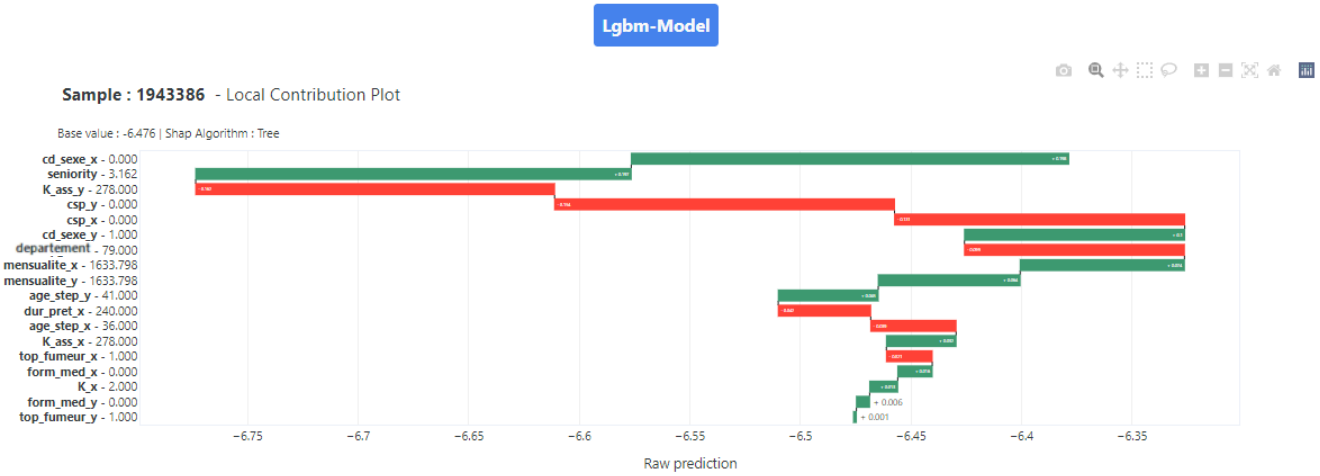


Figure 145- Contribution des variables de l'assuré X dans le modèle couple

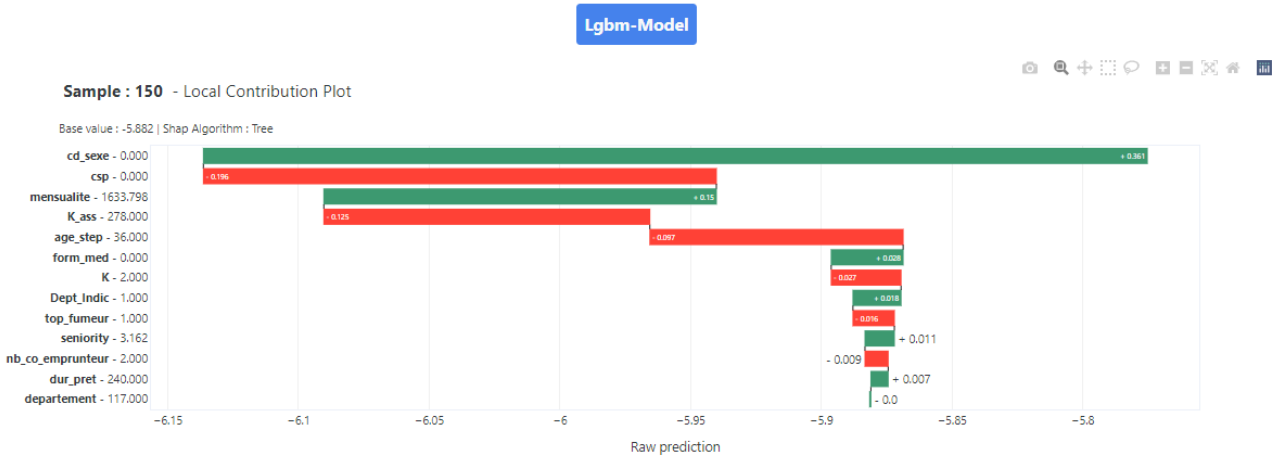


Figure 146- Contribution des variables de l'assuré X dans le modèle individuel

Ces deux graphiques mettent en exergue la contribution des caractéristiques d'un même individu dans les modèles M_i et $M_{c(x,y)}$.

La plupart des variables existants dans les deux modèles agissent dans le même sens, comme l'âge de l'individu ou son sexe. Nous observons cependant deux différences.

Tandis que dans $M_{c(x,y)}$ la durée du prêt présente moins de risque, celle-ci est considérée comme risquée par le modèle M_i.

Le capital assuré quant à lui est présenté comme plus risqué par $M_{c(x,y)}$ que par M_i. Pour cet individu, la prime obtenue par le modèle global sera beaucoup plus importante que celle obtenue après la modélisation en couple. La seule différence entre ces deux modélisations c'est que $M_{c(x,y)}$ semble avoir pris en compte les variables du conjoint de l'individu X qui semble également peu risqué étant donnée ses caractéristiques.

5.3 Validation de la cohérence du modèle

5.3.1 Les valeurs de Shapley (SHAP)

Shap (Shapley Additive Explaination) Explainer est un package de la bibliothèque SHAP de python. C'est un outil qui permet d'expliquer les modèles de machine Learning considérés jusqu'alors comme des « boites noires ». Shap est fondé sur la théorie des valeurs de Shapley que nous trouverons dans la théorie des jeux par exemple. Le but de cette méthode est d'attribuer des contributions équitables aux différentes variables de modélisation afin de mesurer leur impact réel sur la variable de prédiction.

Ainsi pour chaque variable, Shap Explainer compare l'impact de la variable étudiée à la prédiction moyenne sur l'ensemble des valeurs d'entrainement.

Shap calcule ces valeurs de manière individuelle, ce qui donne l'impact marginal de chaque caractéristique à la prédiction du modèle. Mais les valeurs peuvent être agrégées pour donner une interprétation globale. Nous pouvons ainsi utiliser :

- Summary Plot() : Il affiche les valeurs de Shap moyenne pour chaque variable explicative et les trie dans l'ordre décroissant de leur importance.
- Calculer la somme des valeurs absolues des valeurs de shap afin d'obtenir l'impact global de chaque variable.
- Calculer l'importance relative : ceci est obtenue en divisant la somme des valeurs absolues des valeurs de shap par la somme des valeurs de shap en valeurs absolue de toutes les variables.
- Comparaison avec les valeurs de référence : Comparer les valeurs de shap moyennes pour chaque caractéristique par rapport à une valeur de référence. Nous pouvons penser à la valeur moyenne du modèle.

Dans cette partie, nous enlevons la variable seniority du modèle couple et nous analysons le comportement du modèle en utilisant les variables de Shapley. Pour rappel, seniority fait référence à l'ancienneté de l'individu dans le portefeuille.

5.3.2 Variables explicatives pour le décès

Dans le modèle limité aux couples, les variables dont le suffixe et « x » représentent les variables explicatives de l'individu modélisé tandis que les variables finissant par « y » représentent les variables explicatives du conjoint.

Dans cette partie, nous enlèverons la variable « seniority » dans le modèle limité aux couples. En effet, celle-ci indique l'ancienneté de l'individu dans le portefeuille pour tous les prêts confondus. D'autre part, elle influence le modèle positivement, or nous ne souhaitons pas pénaliser la fidélité d'un client. Nous analysons donc la différence entre ce nouveau modèle et le modèle individuel.

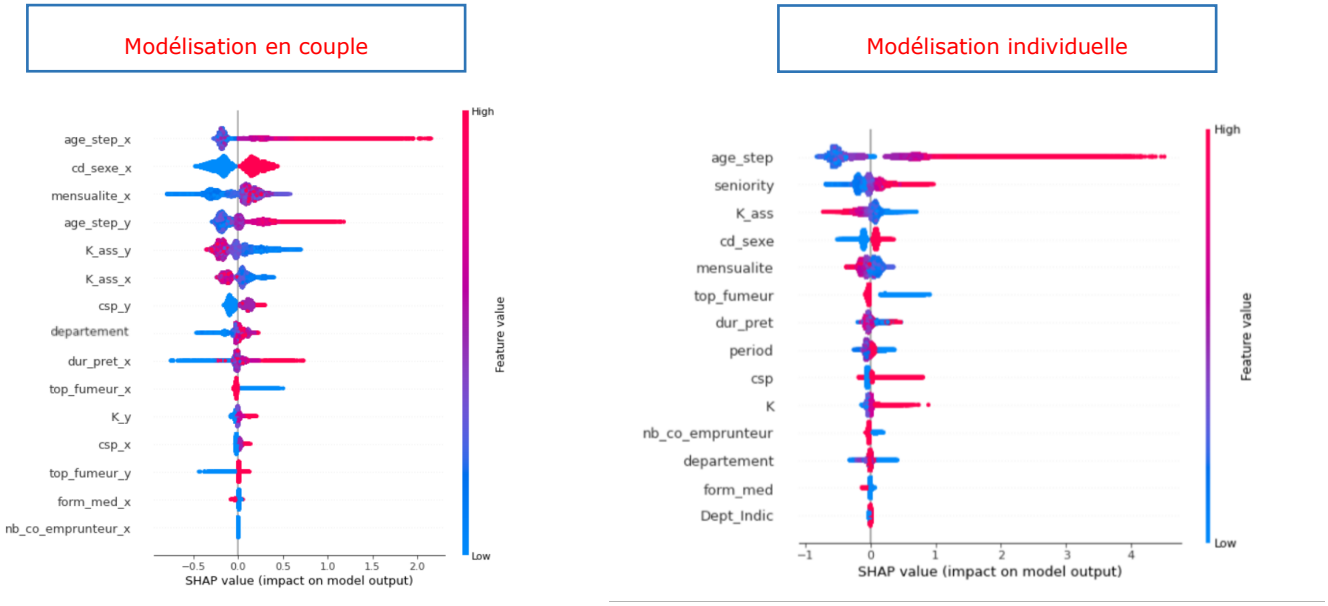


Figure 147 - Shap Value individuelle et en couple

Le ratio de Shap indique l'importance des variables au sein du modèle de machine Learning considéré. Il indique le rang de significativité des variables explicatives et leur niveau d'influence sur les prédictions.

L'ordre des variables n'est donc pas le même dans les deux modèles. Cependant il n'y a une différence dans le nombre de variables. Nous remarquons que pour le modèle individuel tout comme le modèle entrainé sur les couples, l'âge de l'individu est la variable dont le niveau de significativité est le plus important.

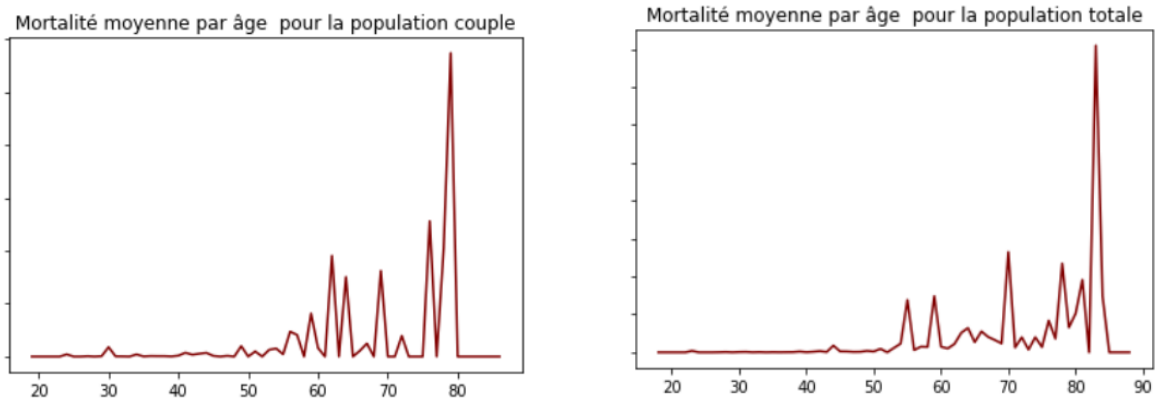


Figure 148 - Mortalité moyenne par âge

Sur ces graphiques bidimensionnels, nous remarquons que l'âge a la même tendance dans les deux cas. Le taux de décès semble malgré tout plus important pour la population générale que pour la population des couples (l'échelle des graphiques n'est pas la même).

Il est ensuite suivi de la variable Seniority qui mesure l'ancienneté de la personne dans le portefeuille, dans le modèle individuel tandis que cette variable ne fait pas partie de la modélisation dans le modèle limité aux couples. En effet, cette variable n'améliore pas l'adéquation lors de la modélisation.

En seconde position dans ce dernier modèle, nous observons le sexe de l'individu, tandis qu'elle n'apparait qu'en quatrième position dans la population totale.

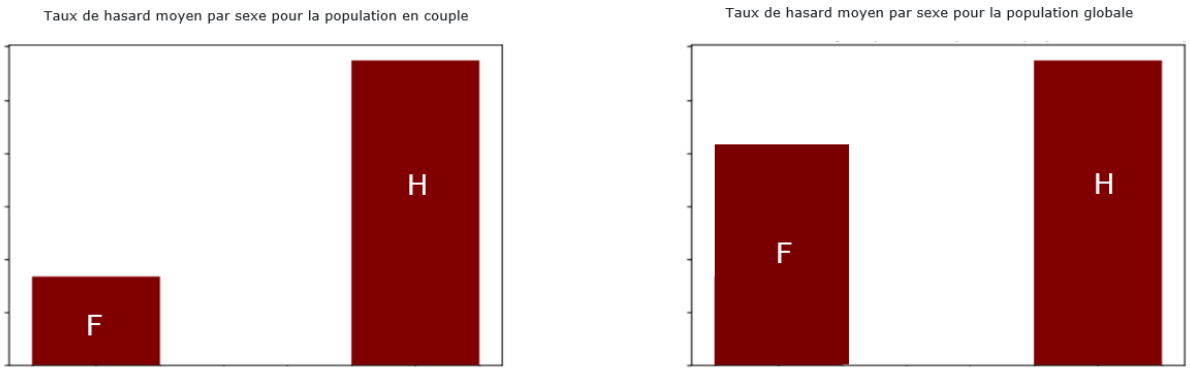


Figure 149 – Taux de hasard moyen par sexe

Contrairement à l'échantillon limité aux couples dans lequel le taux de hasard masculin est trois fois plus important que le féminin, nous observons dans la population que les hommes décèdent plus souvent proche de la date d'anniversaire de leur contrat.

En termes de nombre brut, nous observons la même tendance chez les couples par rapport à la population totale.

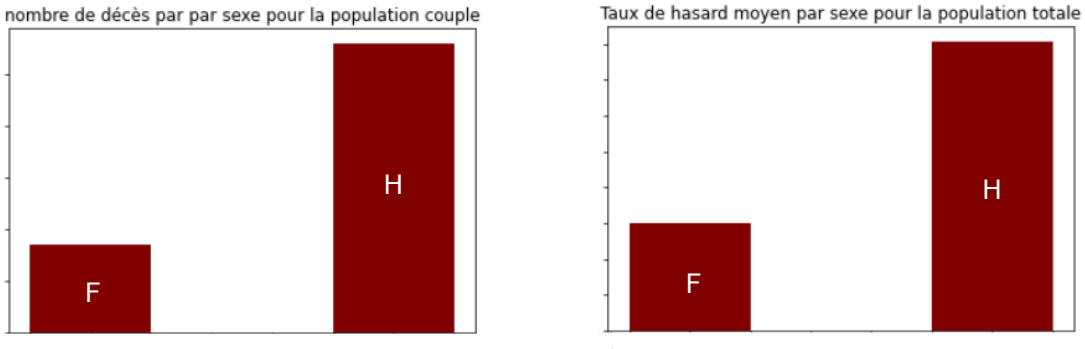


Figure 150 - Comparaison nombre de décès et taux de hasard par sexe

Nous observons malgré tout une différence dans les proportions qui sont trois à quatre fois plus importantes dans la population totale que dans la l'échantillon des couples.

En somme, la variance des taux de décès par classe (homme/femme) au sein des couples est beaucoup plus importante qu'au sein de la population totale. C'est pour cela que les valeurs de shap sont plus importantes pour le premier groupe.

En troisième position, Shap nous propose la variable K_ass dans le modèle global. Cette variable fait référence au capital assuré. Tandis qu'il propose la mensualité de la tête assurée dans le modèle pour couples.

Il faut savoir que la mensualité est un indicateur de richesse de l'assuré qui est plus juste que le capital assuré. En effet, la mensualité prend en compte le taux d'intérêt et la durée du prêt, tandis que le capital assuré ne prend en compte que la proportion du capital emprunté afférent à la tête assurée.

Or, emprunter un montant M avec un taux $r < r'$, n'est pas la même chose qu'emprunter le même montant avec un taux r'. La mensualité présente une sensibilité de 2,5% par rapport à la variation de 1 point du taux nominal.

La correction du calcul de la mensualité a donc nettement amélioré la modélisation.

Après la mensualité, les variables les plus importantes dans l'échantillon limité aux couples sont dans l'ordre :

- Age_step_y: qui est l'âge du conjoint de la tête modélisée.
- K_ass_y : Le capital assuré du conjoint de la tête modélisée.
- K_ass_x : Le capital assuré de la tête modélisée.

Lorsque nous analysons le taux de hasard moyen en fonction de l'âge du conjoint, nous remarquons que ce taux se comporte de la même manière que l'âge de la tête assuré.

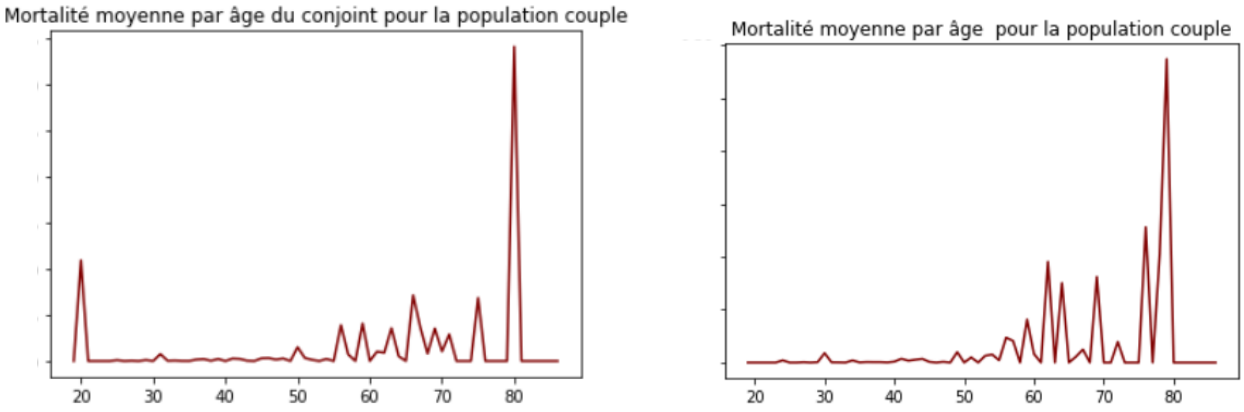


Figure 151 - Mortalité moyenne par âge

Si nous traçons les dispersions des individus de plus de 40 ans et plus âgés que leur partenaire, nous remarquons que la variance du taux de décès instantané est beaucoup plus importante que pour les individus de plus de 40 ans qui seraient plus jeunes que leur partenaire (l'échelle n'est pas la même).

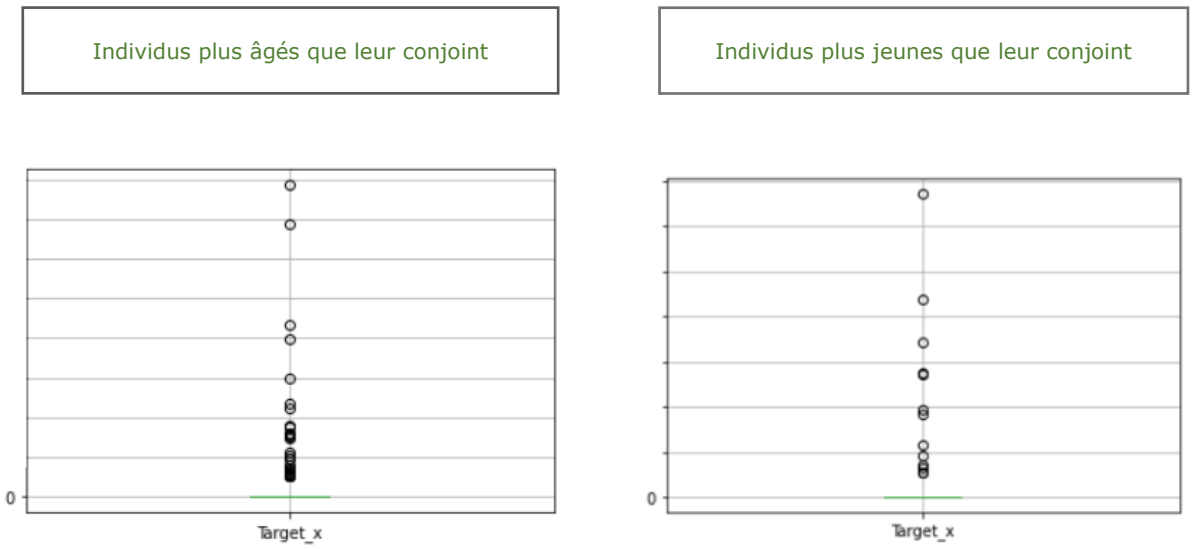


Figure 152 - Variance du taux de décès instantané

En effet, la plupart des personnes âgées de plus de 40 ans dont le conjoint est plus jeune ont des taux de hasard moyens de 1,5 fois plus petit que ceux dont les conjoints sont plus âgés. Il semblerait que vivre avec un conjoint plus jeune augmenterait l'espérance de vie de la tête modélisée.

Nous concluons donc que l'âge du conjoint joue un rôle significatif dans la modélisation du taux de décès instantané de l'assuré.

Après la pertinence de l'âge du conjoint, nous avons le capital assuré du conjoint suivi tout de suite après de celui de la tête assurée. Si nous regardons les deux boxplot suivants, nous pouvons voir que les deux graphes sont très similaires, mais les points sur le graphique de gauche sont légèrement moins dispersés que ceux du graphique de droite.

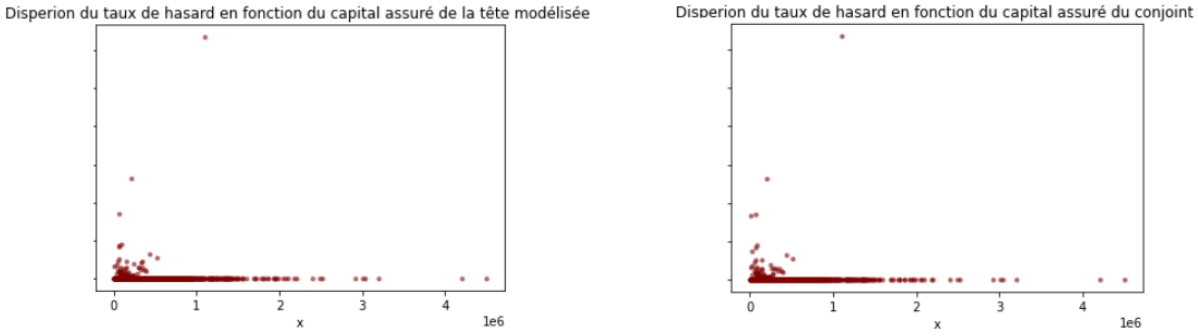


Figure 153 - Dispersion du taux de hasard en fonction du capital assuré

Si nous enlevons les valeurs extrêmes, nous pouvons voir une dispersion un peu plus franche pour la mortalité en fonction du capital du conjoint (les échelles sont différentes).

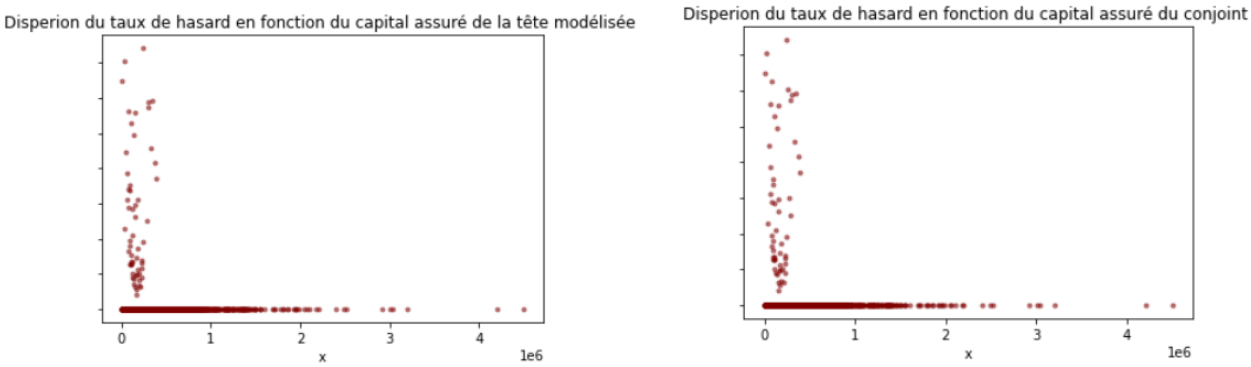


Figure 154 - Dispersion du taux de hasard en fonction du capital assuré sans valeurs extrêmes

Nous pouvons également constater une convergence des statistiques descriptives avec le graphe de Shap qui indique que plus le capital assuré (de la tête modélisée ou de son conjoint) est important, plus de taux de hasard est bas.

Nous voyons une nouvelle fois que la catégorie socio-professionnelle du conjoint de l'assuré souligne son importance en se plaçant devant la variable top-fumeur qui indique le rapport au tabac (0 pour les non-fumeurs et 1 sinon).

Le graphe de Shap indique par ce classement que le taux de hasard est beaucoup plus sensible à la catégorie socio-professionnelle de la tête étudiée qu'au tabac.

Nous tenons malgré tout à rappeler que cette étude est menée sur une population qui contient peu de fumeurs. Ce résultat peut être biaisé par la parcimonie de cette modalité. D'autre part, la population étant composée principalement d'une population active, le rapport au tabac est potentiellement moins capté par le modèle que si l'on avait à faire à une population vieillissante. En effet, l'action néfaste du tabac se manifeste plus souvent dans les grands âges. L'étude fait donc face à une censure. Ceci peut jouer un rôle non négligeable sur la différence entre la population du portefeuille par rapport à la population nationale.

En somme avec un âge médian beaucoup plus jeune pour la population des couples, nous pouvons observer que les âges sont moins dispersés dans cet échantillon. D'autre part, les caractéristiques d'une telle population et son risque face au décès ne peuvent pas converger avec celles des tables de mortalité classiques. Enfin, le peu de fumeurs présents dans notre base ne peut pas saisir toutes l'information

nécessaire par rapport au risque du tabagisme. Nous ne pouvons ni isoler son effet sur la population totale, ni son effet dans le temps à cause du jeune âge dans le portefeuille.

De plus, si l'on regarde les statistiques descriptives, nous observons une rupture assez franche en fonction des catégories socio-professionnelles du conjoint des individus étudiés :

Figure 155 - Taux de hasard et mortalité moyenne par CSP

Le schéma ci-dessous démontre une certaine sensibilité à la catégorie socio-professionnelle dans le choix des conjoints.

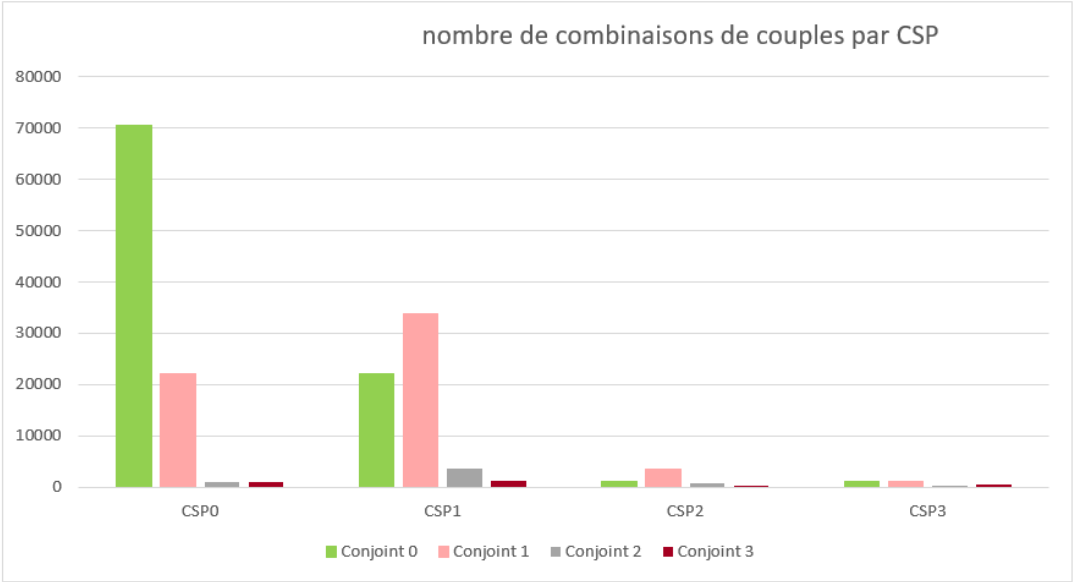


Figure 156 - Nombre de combinaisons de couple par CSP

Notons que les personnes de la catégorie 0 ont tendance à s'unir avec des personnes de la même catégorie. Les assurés de la catégorie 1 semblent également avoir une préférence pour les personnes de leur catégorie en premier puis celles de la catégorie 0.

Si nous nous intéressons maintenant aux taux de hasard moyen par combinaisons des catégories socio-professionnelles, nous remarquons que les comportements ne sont pas du tout les mêmes.

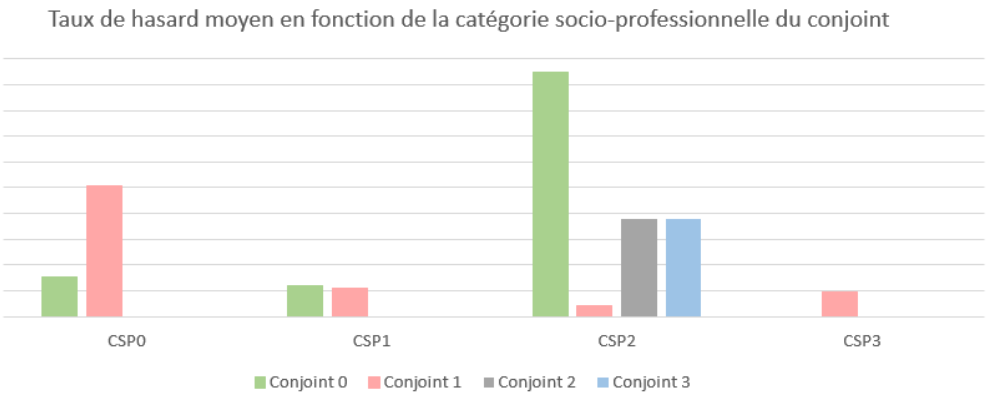


Figure 157 - Taux de hasard moyen en fonction de la CSP du conjoint

Nous remarquons par exemple que les personnes de la catégorie 0 ont un taux de décès instantané plus important lorsqu'ils sont unis à une personne de la catégorie 1 par rapport à une personne de leur catégorie. Tandis que la catégorie 2 est plus sensible à la catégorie 0.

Ces variances un peu brutales font en sorte que le modèle LightGBM soit sensible à la catégorie socio-professionnelle du conjoint assuré.

Cela est suivi de la variable département qui est plus importante dans la population restreinte aux couples. En effet si nous regardons les graphes ci-dessous, nous remarquons une réelle différence entre la population totale et le second graphe qui trace le taux de hasard moyen par département pour la population restreinte aux paires d'assurés.

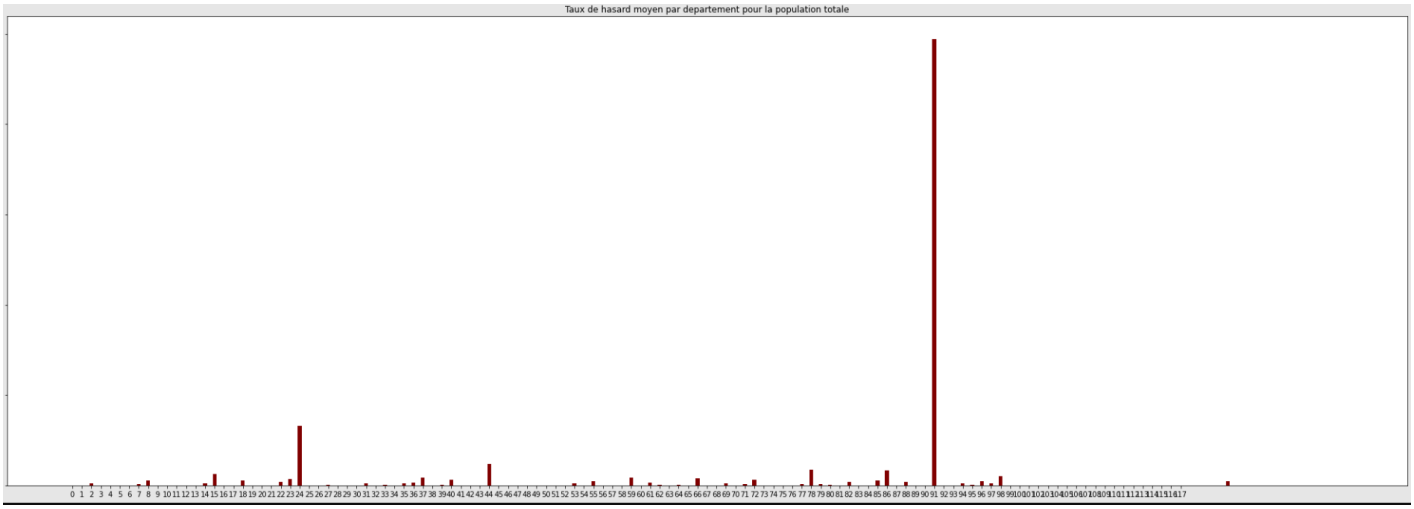


Figure 158 - Taux de hasard moyen par département pour la population totale

Le pic du taux de hasard moyen est atteint dans le 91 pour la population totale, tandis qu'il est atteint pour les couples dans la Loire. Le maximum du taux de décès moyen est quatre fois plus important pour la population totale que pour les couples.

Le second taux le plus important est atteint dans le 24 pour la population totale tandis que pour les couples, le même taux est atteint dans le 24 et 82.

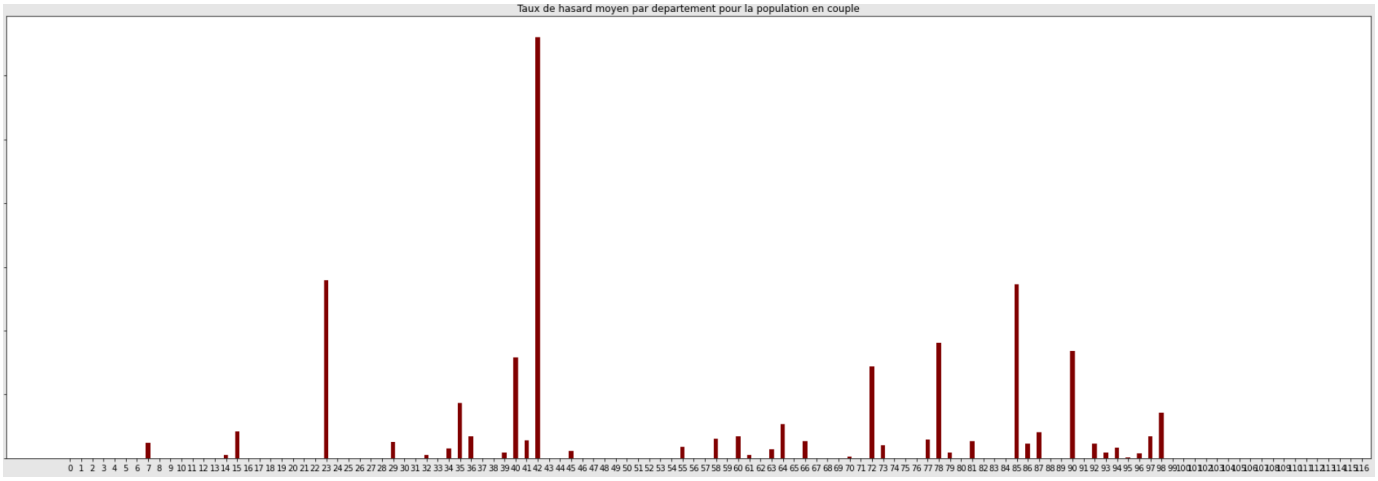


Figure 159 - Taux de hasard moyen par département pour les couples

Enfin, notons que la dispersion globale n'est pas tout à fait la même et que la dispersion est beaucoup plus tranchée dans la population limitée aux couples.

La durée de l'emprunt est aussi importante pour la population restreinte aux couples que pour la population globale. Même si l'effet des valeurs extrêmes biaise un peu ce point de vue, lorsque nous les enlevons nous remarquons une homogénéité des deux échantillons en fonction de cette variable.

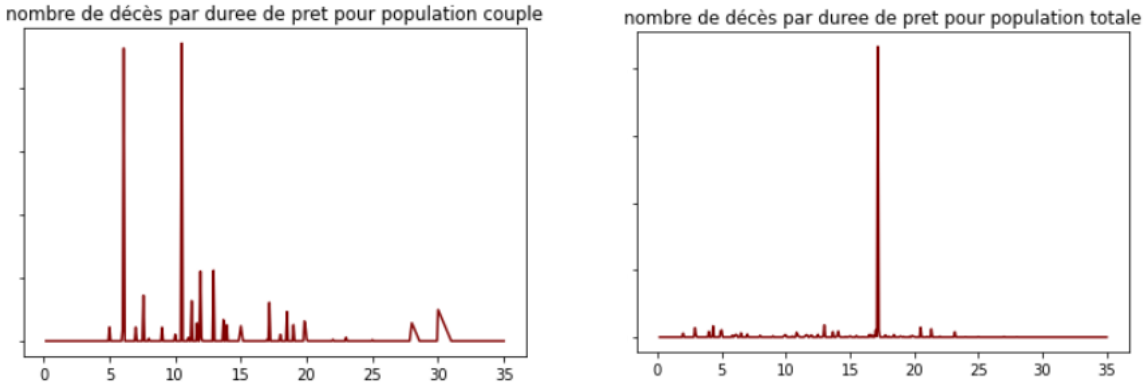


Figure 160 - Nombre de décès par durée de prêt

En effet, sans les valeurs extrêmes, nous remarquons de manière globale que le taux de hasard en fonction de la durée est moins dispersé au sein de la population totale qu'au sein de la population couple.

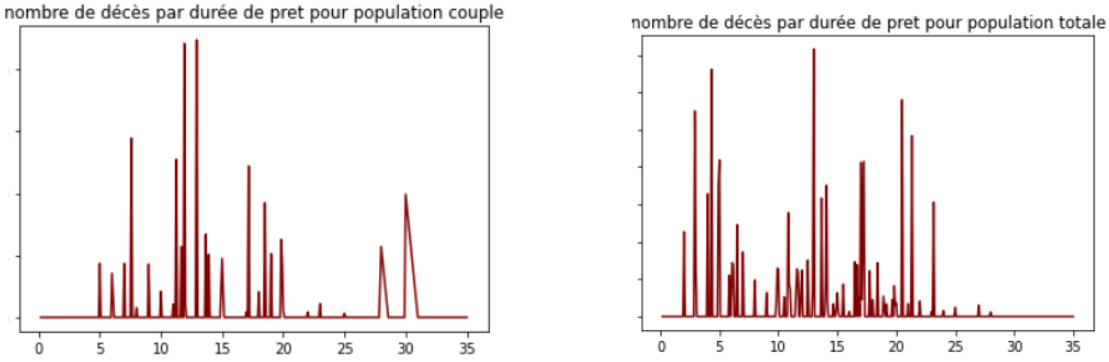


Figure 161 - Nombre de décès par durée de prêt sans valeurs extrêmes

Enfin le squatter-plot ci-dessous rejoint notre remarque sur le fait que la variable « durée du prêt » influence les deux modèles avec la même intensité.

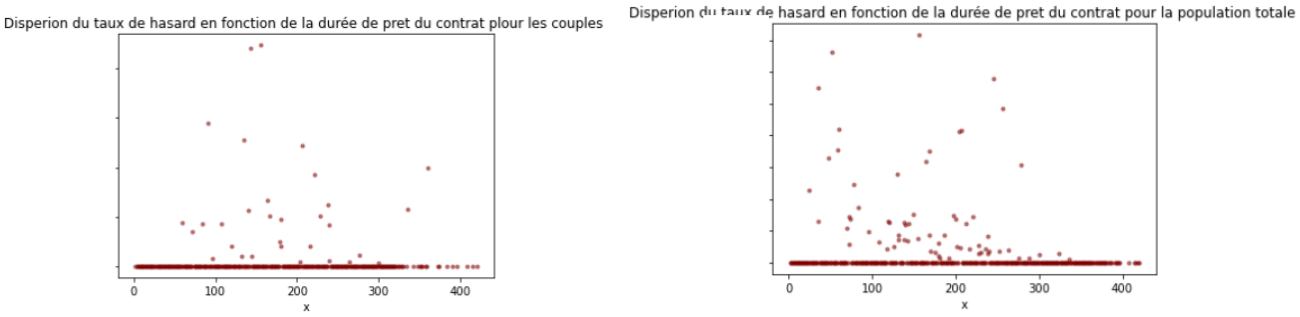


Figure 162 - Dispersion du taux de hasard en fonction de la durée de prêt

La variable top_fumeur_y (état de tabagisme du conjoint) prendra ainsi la $6eme$ position. Cette information est cruciale car nous avons ici une confirmation de notre hypothèse concernant le tabac. Même si le graphe de Shap ne tranche pas la question de manière claire, il indique toutefois sans équivoque le fait que les couples non-fumeurs constituent une catégorie d'assurés moins risqués. En effet, la partie rouge afférente aux fumeurs est proche de 0, nous ne pouvons pas lier la consommation de tabac du conjoint à la mortalité de manière claire. Cependant, nous pouvons suspecter la parcimonie des données vu qu'ils ne présentent qu'une petite portion concernant les fumeurs. En effet, ceci crée un déséquilibre au sein de notre échantillon, ce qui peut provoquer un amoindrissement de l'importance d'une telle information. De plus, l'effet du tabagisme actif ou passif sur la mortalité se constate à des âges bien plus élevés par rapport à l'âge médian de la population du portefeuille.

Concernant l'ordre donné par Shap, les graphes ci-dessous, montrent que le taux de hasard est plus variable en fonction du statut face au tabagisme dans la population globale que dans la population limitée aux couples.

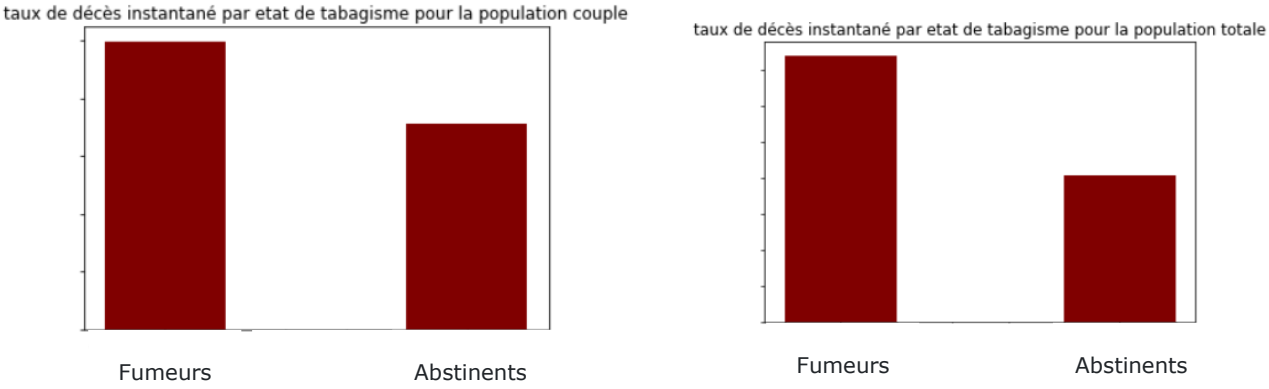


Figure 163 - Taux de décès instantané par état de tabagisme

La différence entre le taux de hasard des fumeurs et des abstinents dans la population totale est plus notable qu'au sein des couples.

Globalement les fumeurs de la population totale ont un taux de décès instantané de 32% plus important par rapport aux fumeurs en couple.

La différence entre fumeurs et non-fumeurs au sein des couples est de 28% tandis que cette différence est de 45% parmi la population totale.

Le classement de la variable top fumeur (qui est plus importante dans le modèle global que pour les couples) semble totalement justifié. Nous pouvons soupçonner un effet du tabagisme passif qui pourrait réduire cette différence au sein des couples.

Le modèle couple quant à lui présente un ordre plutôt classique lorsque nous excluons la variable « seniority » dans la modélisation.

En effet, le graphe ci-dessous présente les trois premières variables les plus pertinentes dans le même ordre que la population nationale. Nous observons ainsi :

- L'Age de la tête modélisée
- Le sexe de la tête modélisée
- La mensualité bancaire de la tête modélisée

Si nous nous focalisons sur la mensualité, cette variable est entièrement corrélée au capital assuré.

Cependant la mensualité indique la capacité d'emprunt de l'assuré, de manière plus efficace. En effet, la mensualité normalise un peu plus les niveaux de richesses des individus car elle prend en compte le taux nominal dans son calcul.

Donc à mensualités égales, nous avons à minima le même niveau de richesse.

Le graphique de Shap nous indique les petites mensualités présentent des petits taux de hasard.

Ces taux commencent à devenir un peu plus importants vers le milieu pour finir par diminuer à partir d'un certain seuil S.

Ces informations sont confirmées avec le graphique du taux de hasard en fonction de l'ancienneté. En effet, les grands capitaux sont souvent couplés avec une ancienneté plus importante.

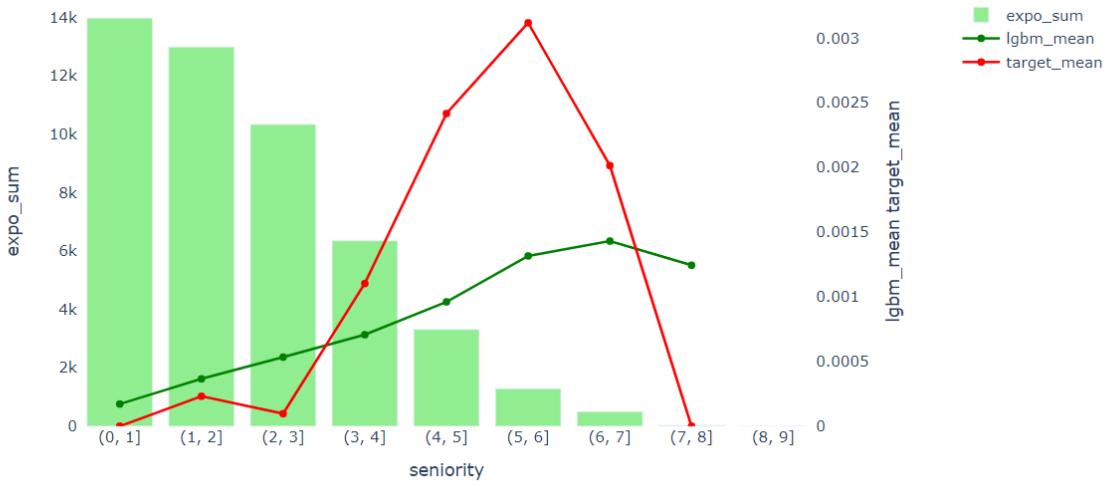


Figure 164 - Taux de hasard en fonction de l'ancienneté

Nous remarquons en effet que les taux de décès sont croissants en fonction de l'ancienneté jusqu'à 5 ou 6 ans. Ces derniers finissent par décroitre par la suite.

Tandis que le capital assuré (K_ass) est beaucoup plus sensible au taux d'emprunt.

Nous remarquons que l'âge du conjoint (age_step_y) joue sur la mortalité de la tête assurée. Cette variable parait beaucoup plus importante que le capital assuré en brut et la catégorie socio-professionnelle de ce dernier.

La variable age_step_y joue globalement dans le même sens que l'âge de la tête assuré. Plus ce dernier est âgé, plus le taux de hasard de l'assuré est important.

5.3.3 Variables explicatives pour l'arrêt de travail

Ces graphiques listent dans l'ordre d'importance les variables explicatives des deux modélisations (individuel et avec lien couple) pour le risque arrêts de travail.

Modélisation individuelle

Modélisation en couple

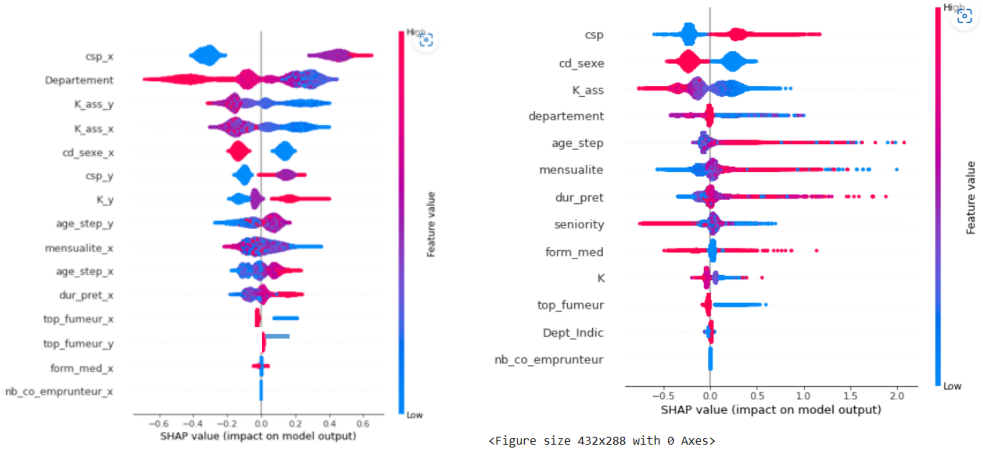


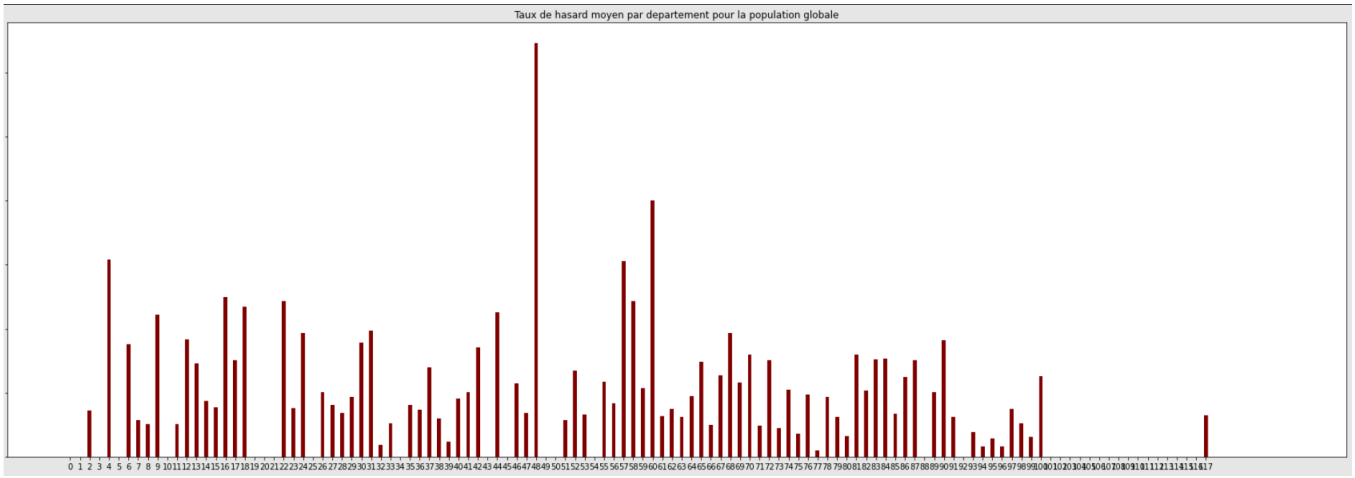
Figure 165 – les valeurs de Shapley du modèle

Dans le premier graphique, les variables plus pertinentes qui contribuent dans la modélisation de l'arrêt maladie sont d'abord la catégorie socio-professionnelle et ensuite le sexe de l'individu.

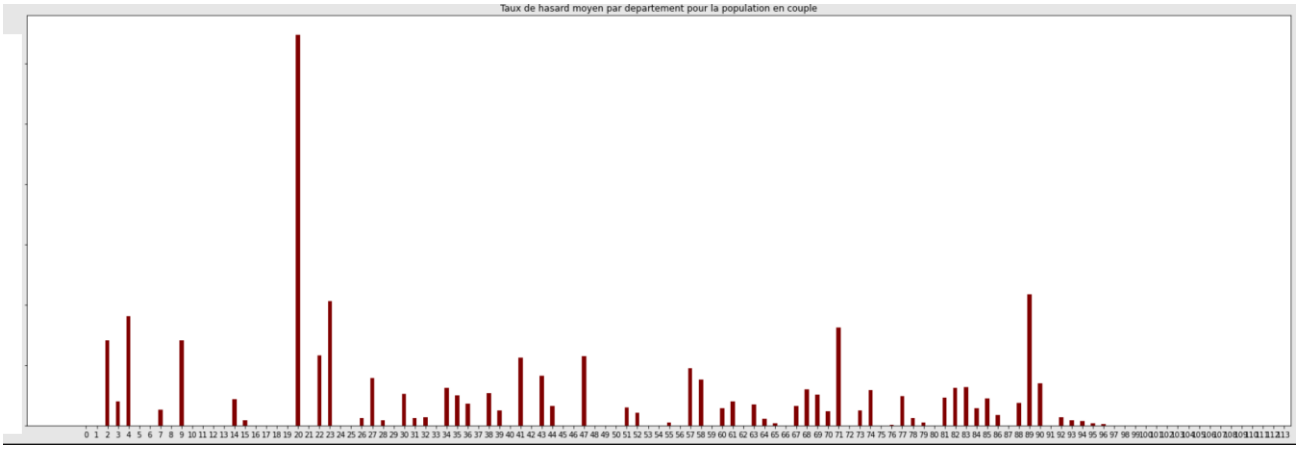
Au préalable, si nous regardons les variables, nous pouvons penser que l'ordre est très différent, cependant il faut garder en tête le fait que dans le second modèle, les variables sont dédoublées.

Nous pouvons remarquer que pour les deux modèles, la variable la plus pertinente est la catégorie socio-professionnelle. Ceci est cohérent car les travaux manuels engendrent beaucoup plus d'accidents physiques qu'un travail intellectuel par exemple.

En seconde variable explicative, nous avons la variable département pour les couples tandis que pour la population générale, c'est le sexe qui prévaut. Si nous observons le taux de hasard par département, nous remarquons que la distribution n'est pas du tout la même.



En effet, dans la distribution de la population totale, le maximum des taux de hasard est atteint en Lozère.



Dans la distribution du taux de hasard par département, le taux maximum est atteint pour la corse et il est 4 fois plus élevé que le maximum atteint par la population totale. De plus, la distribution au niveau des départements est très différente en fonction des deux échantillons. En effet, le tableau ci-dessous confirme nos études.

Ordre des valeurs	Departement Couple	Departement Individuel	Differenc
1	La Corse	Lozère	75%
2	Yonne	Oise	54%
3	Creuse	Alples de Haute Provenance	63%
4	Alples de Haute Provenance	Moselle	11%

Figure 166 - Distribution au niveau des départements

Nous en concluons donc que la disparité en fonction du département est beaucoup plus importante au sein de la population couple qu'au niveau de la population totale.

En troisième position, les deux modèles prennent en compte le capital assuré. Cette variable est importante car elle est entièrement corrélée aux formalités médicales que l'assureur exige à partir d'un certain seuil S de capital assuré. En effet, si nous étudions le taux de hasard en fonction du capital assuré, nous observons que celui-ci est très faible voir nul à partir d'un certain seuil.

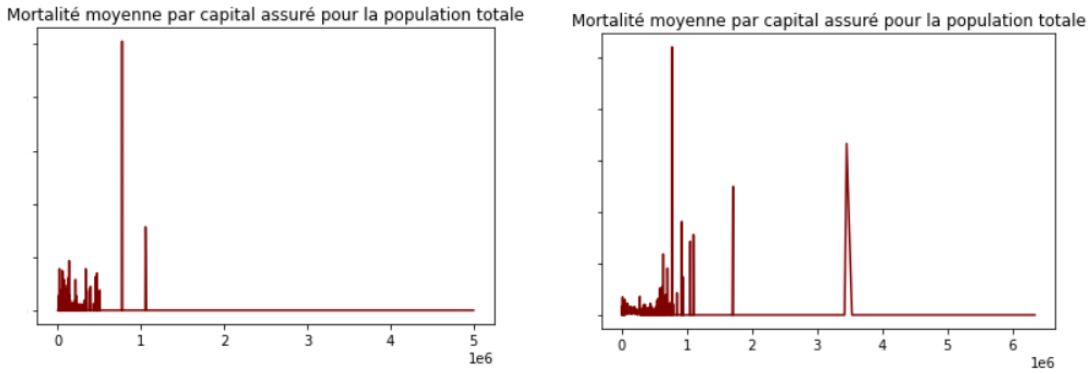


Figure 167 - Mortalité moyenne par capital assuré

Même si le taux de hasard maximum est deux fois plus important chez les couples que dans la population globale, ce taux reste uniformément distribué dans les deux échantillons.

En quatrième position, nous observons de nouveau une différence. Si le sexe est moins significatif dans la population couple, dans la population globale il occupe la seconde position. Pourtant les ratios entre les taux d'incapacité par sexe sont les mêmes parmi les deux populations.

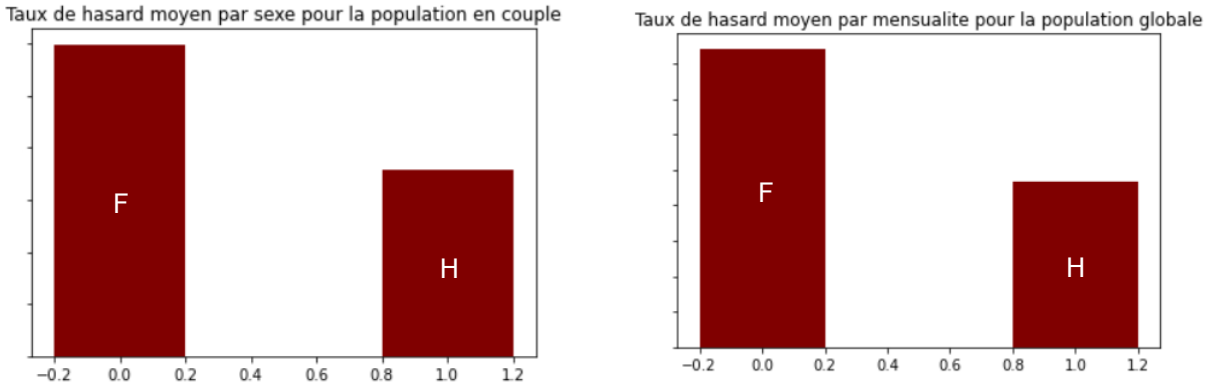


Figure 168 - Taux de hasard moyen par sexe

Nous observons également que les individus en couples sont beaucoup plus malades que les individus appartenant à la population globale. Nous pouvons subodorer que ceci est dû soit :

- Au changement de structure du portefeuille
- Aux contaminations ou au mêmes faits générateurs étant donné que les individus en couple partagent les mêmes conditions de vie.

En cinquième position, nous observons l'influence de la catégorie socio-professionnelle du conjoint de la tête étudiée. Ceci permet de nous conforter dans notre hypothèse de probabilité de contagion. Nous pouvons également supputer que cette relation puisse provenir de l'entraide au sein du couple. En effet, lorsque le conjoint détient un commerce par exemple nous avons tendance à offrir notre aide afin de le soulager dans certaines tâches. Cela peut faire subir à la tête dont le travail l'a protégé jusqu'alors, des risques de maladie plus élevé qu'un autre individu ayant la même catégorie socio-professionnelle.

Si nous reprenons le graphique du taux de hasard moyen par catégorie socio-professionnelle et que nous le comparons à la même étude tout en prenant en compte la catégorie du conjoint, nous constatons une réelle dépendance entre le taux d'incidence instantané et la fonction du conjoint.

En effet, la categorie socio-professionnelle 0 a tendance à diminuer le taux d'incidence de la cetégorie 0, 1 et 3. Avec un conjoint de la catégorie 0, la catégorie 2 augmente son taux d'incidence.

Avec un conjoint de la catégorie 1 les catégories 0, 1 et 3 augmentent légèrement par rapport à un conjoint de la catégorie 0. Les taux restent égaux ou superieurs à ceux de la population totale.

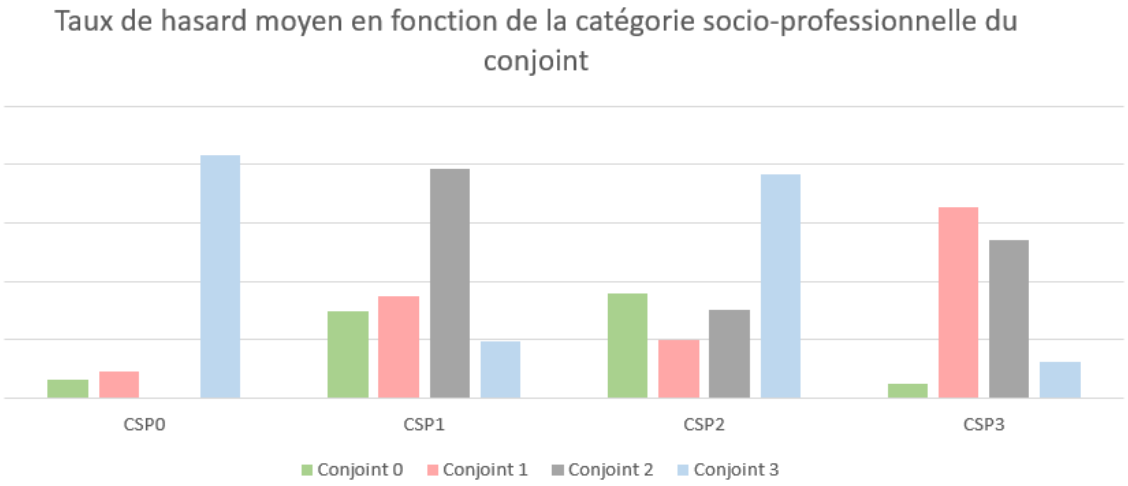


Figure 169 - Taux de hasard moyen en fonction de la CSP du conjoint

Si nous nous focalisons sur les catégories socio-professionnelles 0 et 1, nous remarquons une tendance haussière en fonction de la catégorie socio-professionnelle du conjoint. En effet, le conjoint de la csp_1 a une tendance à aggraver la santé des personnes de la csp_0. Il en est de même pour le conjoint de la csp_3.

Les individus de la csp_1 réagissent globalement de la même manière face aux professions de leur conjoint.

Nous observons que les assurés des csp_2 et csp_3 ne réagissent pas de manière aussi tranchée que les deux premières. Cela est dû à leur hétérogénéité. En effet la csp_2 peut contenir des agriculteurs mais également des salariés non-cadres. Or, nous savons que leurs conditions de vie peuvent énormément différer. Quant à la catégorie 3, celle-ci contient les personnes qui n'ont pas pu être classée jusqu'alors. Ce qui nous conduit à conclure que les catégories 0 et 1 sont les catégories les plus pertinentes dans cette étude. Et ces deux dernières nous permettent de souligner une corrélation entre l'état de santé de l'assuré et le métier de son conjoint.

En sixième variable, nous avons l'ensemble des encours du conjoint de la tête assurée pour le modèle couple tandis que pour le modèle individuel, cette variable n'apparait qu'en 11ième position.

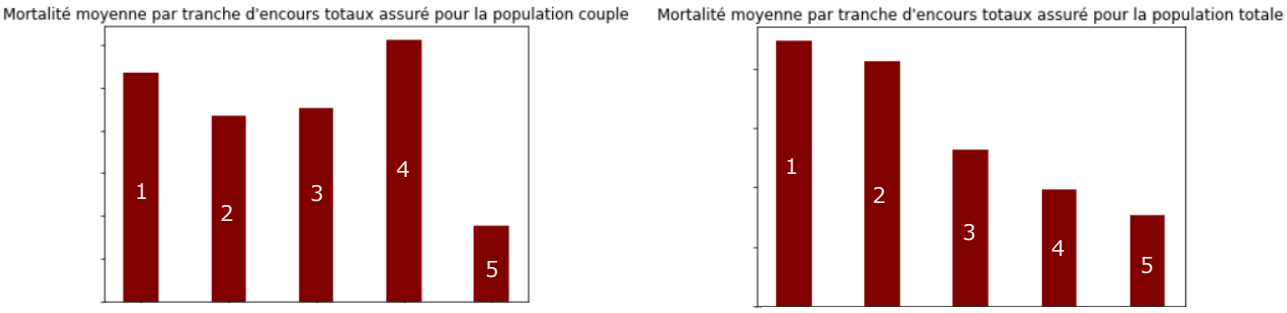


Figure 170 - Mortalité moyenne par tranche d'encours totaux assuré

Lorsque nous observons le taux de hasard en fonction des encours assurés pour la population des couples, nous observons une plus grande différence entre la classe 0 et la casse 1. La différence entre la classe 3 et la 4 est encore plus brutale. C'est pour cette raison que le modèle est beaucoup plus sensible à cette variable dans le modèle avec l'échantillon des couples qu'avec l'échantillon individuel.

Nous pouvons par la suite remarquer que les variables suivantes apparaissent dans le même ordre :

- Age_step
- Mensualité
- Durée du prêt

Ces trois variables contribuent sensiblement de la même manière dans les deux modélisations.

Cependant dans le modèle couple, nous observons une inversion de la contribution de l'âge au sein d'un couple : en effet, si nous zoomons sur ces deux variables nous remarquons le schéma ci-dessous :

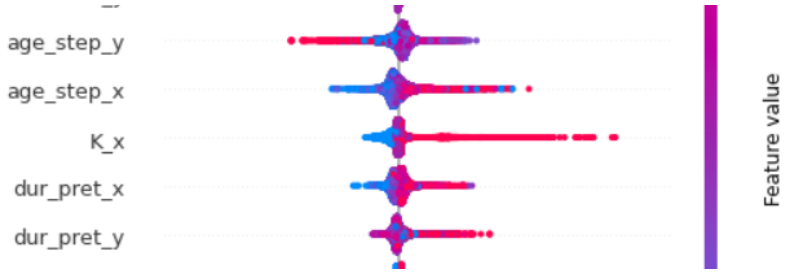


Figure 171 - Contribution de l'âge au sein d'un couple

Le taux d'incidence en arrêt de travail est croissant avec l'âge. Néanmoins, l'âge du conjoint a plutôt l'effet inverse. En effet, plus le conjoint est âgé, plus il contribue négativement à l'entrée en arrêt de travail de la tête modélisée. Les conjoints les plus jeunes quant à eux ont une contribution légèrement positive.

Comme nous pouvons le constater sur ce graphique, les personnes entrant en arrêt maladie n'ont pas une grande différence d'âge avec leur conjoint.

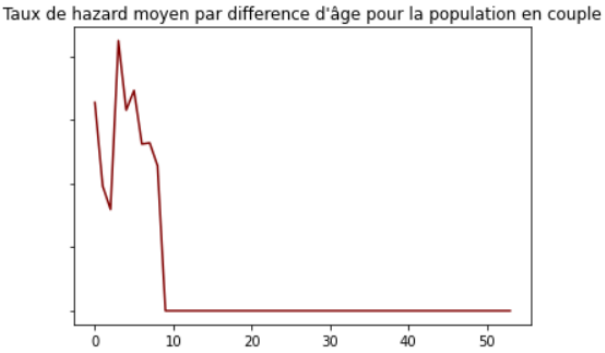


Figure 172 - Taux de hasard moyen par différence d'âge

Nous remarquons par la suite que l'ordre de deux autres variables a été inversé :

13	Form_Med	Top_fumeur_x
14		Top_fumeur_y
15	Top_fumeur	Form_med

Figure 173 - Inversement de l'ordre des variables

En effet, dans le modèle couple le statut de fumeur de la tête modélisée ainsi que celui de son conjoint paraissent plus importantes que la variable « Form_Med » qui indique le type d'analyses médicales auxquelles le couple a été soumis.

Si nous regardons la mortalité moyenne par type de formalités médicales auxquelles les têtes modélisées ont été soumises, nous remarquons une différence notable entre la base globale et la base restreinte aux couples.

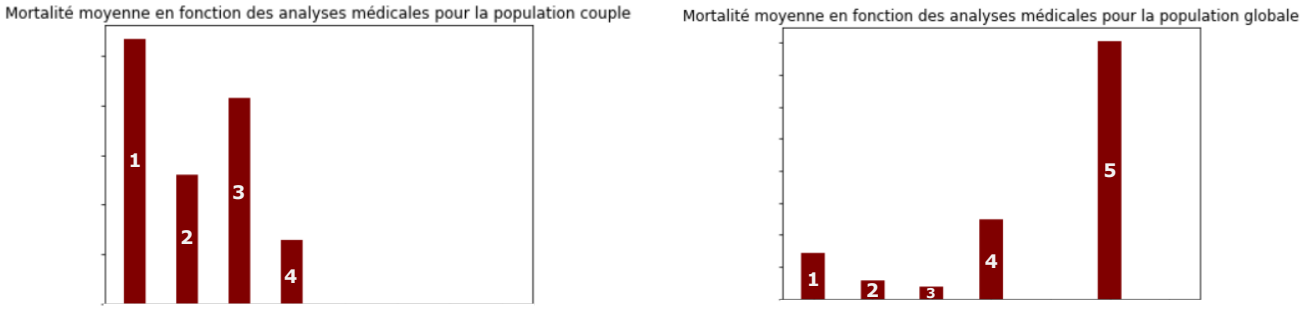


Figure 174 - Mortalité moyenne en fonction des analyses médicales

En effet, en moyenne, les entrées en arrêt de travail concernant la population en couple sont beaucoup moins dispersées qu'au sein de la population globale, lorsque nous regardons ce phénomène en fonction du type d'analyses médicales effectuées par la tête assurée. De plus, les maximas sont d'autant plus écartés en fonction des classes de formalités médicales.

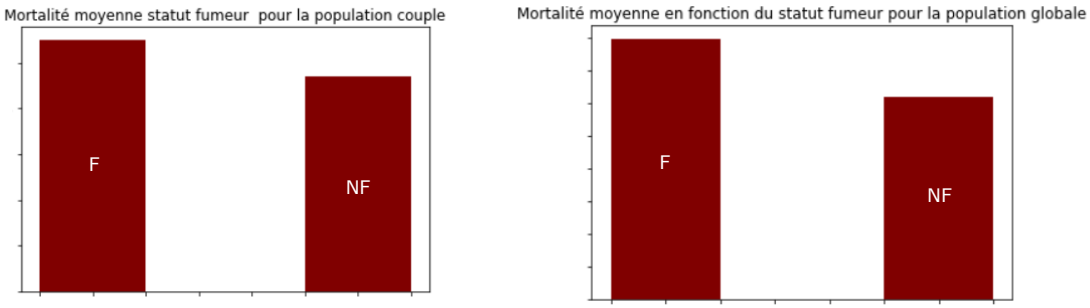


Figure 175 - Mortalité moyenne en fonction du statut fumeur

La distribution du taux de hasard moyen en fonction du sexe est globalement la même pour les fumeurs et non-fumeurs. C'est pour cela que la perturbation de la mortalité en fonction des formalités médicales de la tête assurée fait passer cette variable devant la variable d'indication du statut de fumeur dans le modèle global.

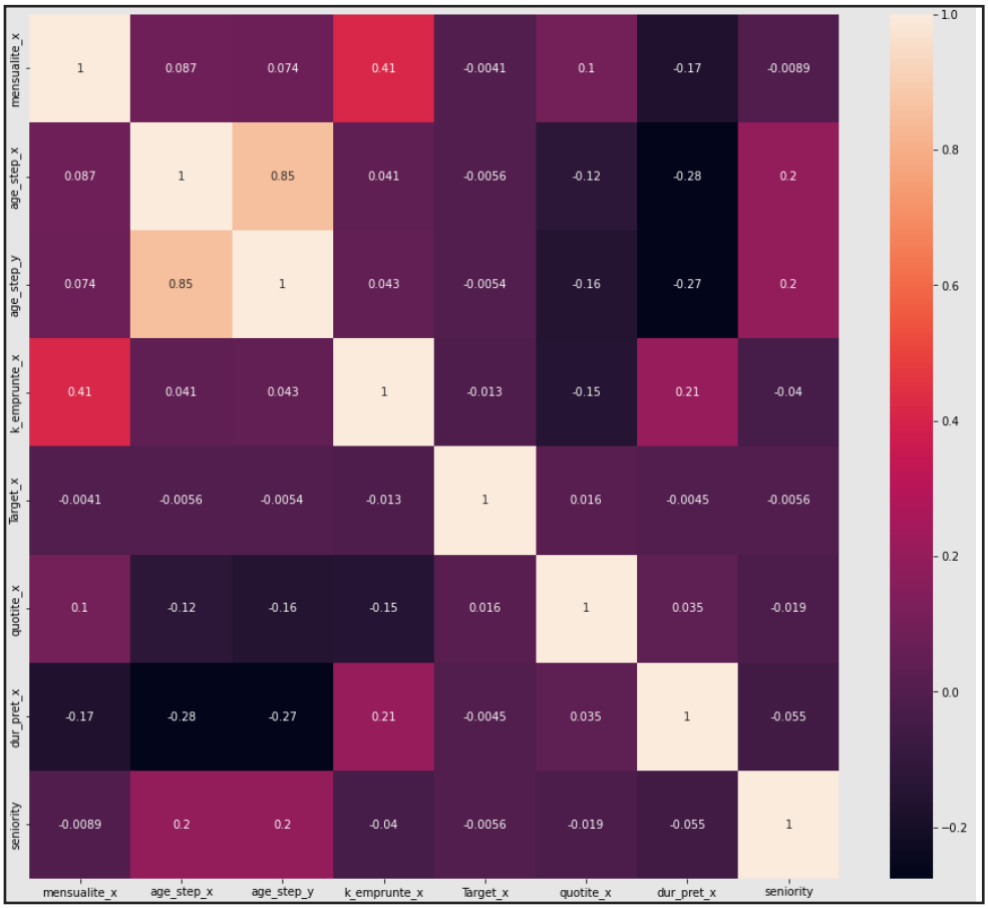


Figure 176 - Distribution du taux de hasard en fonction du sexe

Si nous regardons le graphique des corrélations linéaires, et que nous nous concentrons sur la variable Target_x qui représente le taux de hasard modélisé, nous pouvons confirmer la cohérence des valeurs de Shap contenues dans le graphique plus haut. En effet, les corrélations linéaires respectives des variables explicatives sont dans le même ordre que le niveau de pertinence de ces variables selon le graphique de Shap.

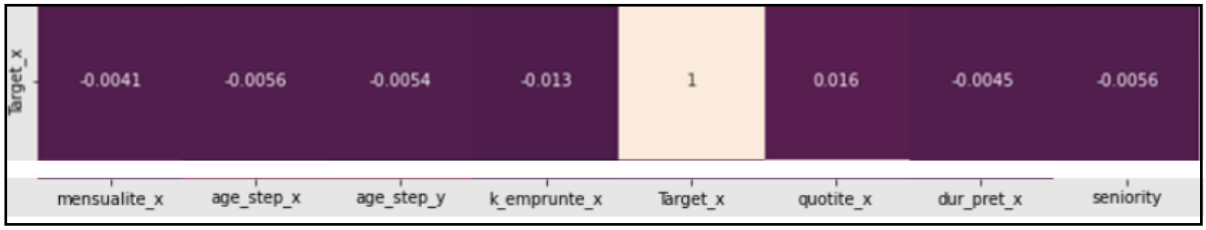


Figure 177 - Corrélation linéaire de la variable de prédiction

Si nous regardons les statistiques descriptives ci-dessous, nous pouvons comprendre d'où vient cette différence entre la population limitée aux couples et la population globale.

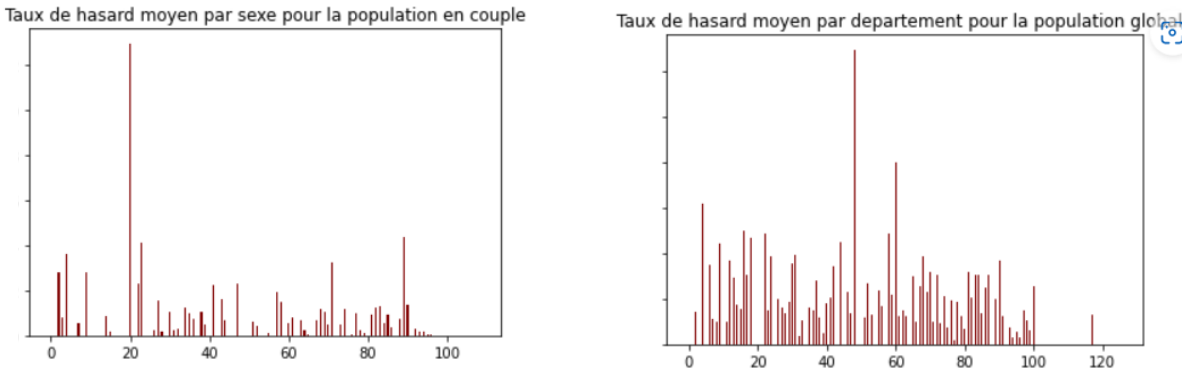


Figure 178 - Taux de hasard moyen par sexe

Dans ces deux graphiques, il apparait clairement un comportement différent du taux de hasard en fonction du statut matrimonial de l'assuré. En effet, le maximum est 30% plus élevé pour les couples que pour la population totale.

D'autre part, si nous regardons la différence des deux taux, nous pouvons voir que le comportement des deux échantillons de population n'est pas du tout le même.

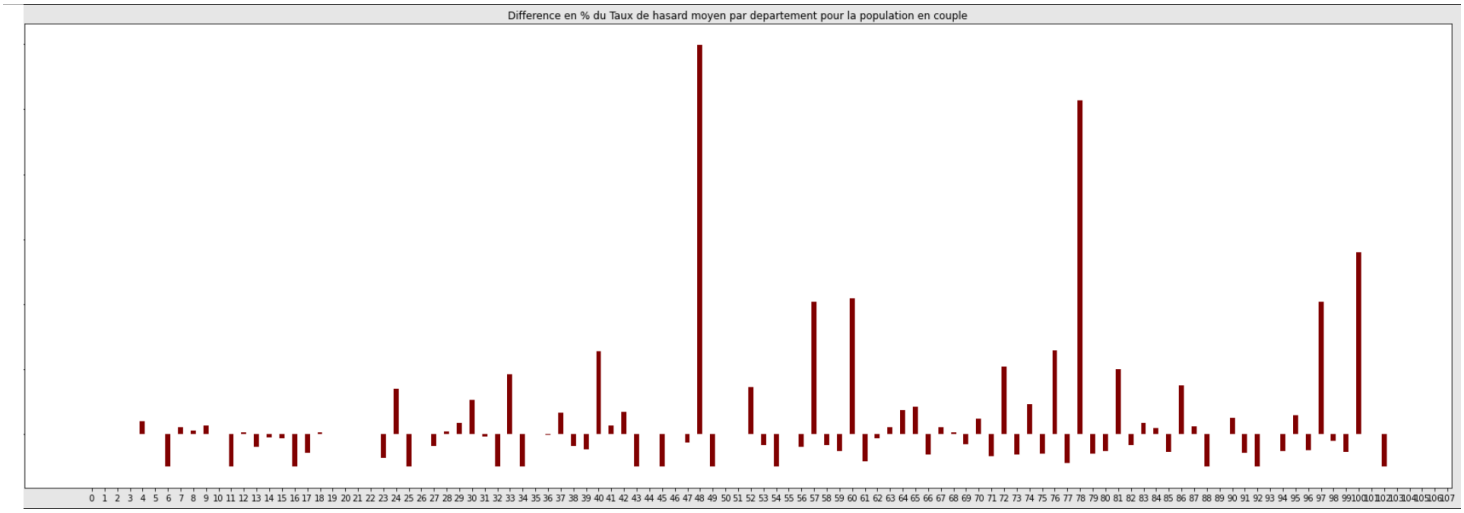


Figure 179 - Comportement des échantillons de population

Nous remarquons ainsi que pour la population en couple, le taux de hasard pour l'entrée en arrêt de travail est beaucoup plus sévère que pour la population totale du portefeuille en Lozère, dans la Moselle,

l'Oise, les Yvelines, ou en Guadeloupe par exemple. Ces départements abritent une population qui travaillent dans des métiers beaucoup plus manuels que le reste des départements.

Les deux modèles s'accordent sur le capital assuré. En effet, cette variable est d'autant plus pertinente que les individus sont soumis à partir d'un certain seuil S de capital emprunté à des formalités médicales de plus en plus poussées. D'autre part, plus le capital assuré est élevé et plus la personne doit montrer qu'elle est en bonne santé. Ce filtre permet de limiter les risques au sein des individus qui souscrivent à l'assurance emprunteur. Ceci est démontré par les graphique suivants :

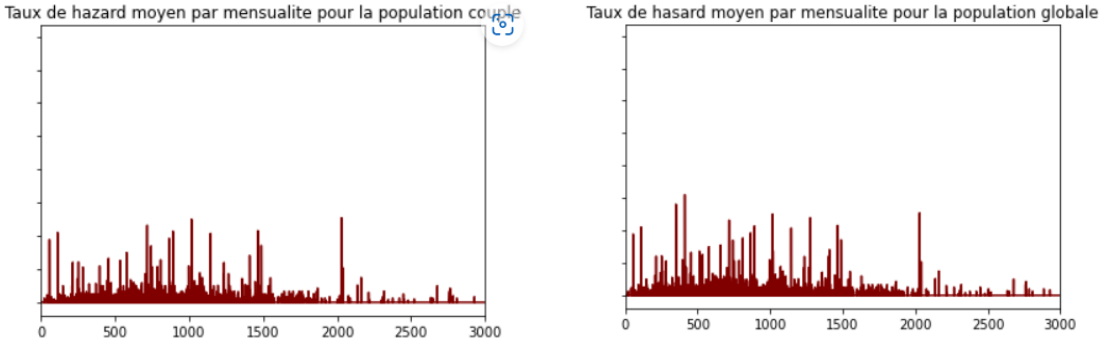


Figure 180 - Taux de hasard moyen par mensualité

Ces graphiques montrent globalement que plus la mensualité est élevée et plus le taux de hasard diminue.

Dans le cas des couples comme pour la population générale (à droite), le taux de hasard est plus important pour les mensualités faibles. Cependant le taux de hasard s'étale sur une zone plus importante dans la population générale, et le pic vers les 18000 qui se trouve dans la population générale n'existe pas parmi les couples. Cette variable ne se présente donc pas de la même manière dans les deux bases.

Ceci est confirmé par les statistiques obtenues à partir de l'étude du capital assuré :

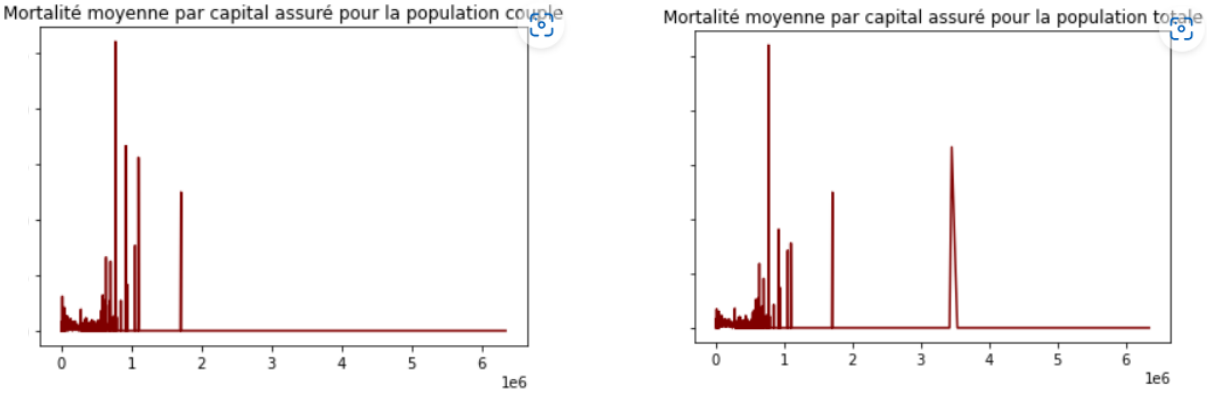


Figure 181 - Mortalité moyenne par capital assuré

Cela dit, le modèle couple nous propose la variable sexe de la tête modélisé comme seconde variable :

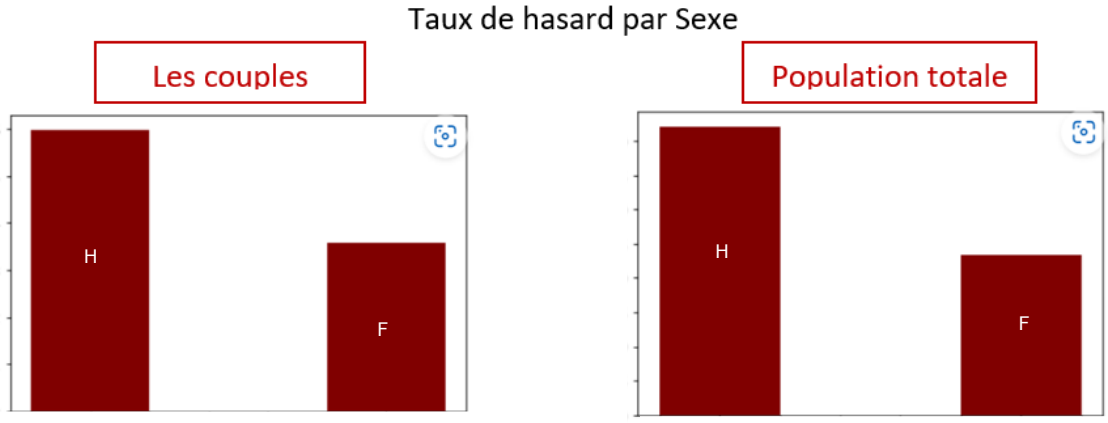


Figure 182 - Taux de hasard par sexe

Dans le cas des couples comme dans le cas de la population générale (à gauche), le taux de hasard moyen se comporte de la même manière. Il est toutefois un peu plus important chez les couples que dans le reste.

En troisième position, nous retrouvons la mensualité. Tout d'abord le calcul de la mensualité est fait par l'établissement de prêt. Celui-ci estime en amont le risque de l'emprunteur non seulement en fonction du capital sous risque, de son âge, de sa catégorie socio-professionnelle mais également du taux d'emprunt. En effet, en fonction de ce taux, toutes choses égales par ailleurs, les emprunteurs n'ont pas toujours la même capacité d'emprunt. De plus, nous sommes sur un risque long, donc le modèle est forcément sensible à cette information.

5.4 Modélisation individuelle par LightGBM et application de la copule

Le but de cette étude est de modéliser le taux de hasard des assurés du portefeuille avec le modèle existant. En effet, il s'agit de modéliser les sinistres tête par tête, puis d'appliquer une couche supplémentaire afin de lier les résultats obtenus par la copule choisie.

Nous modéliserons ainsi deux variables aléatoires indépendantes $\lambda_1$et λ_2 par l'algorithme de machine learning LightGBM qui seront respectivement les taux de décès instantanés des assurés principaux et secondaires au sein de notre portefeuille d'assurés. Ces variables aléatoires correspondront comme expliqué dans la section 3.7, au taux de hasard $\lambda_i(t) = \frac{f(t)}{S(t)} = = \lim_{\Delta \to 0} \frac{\mathbb{P}[t \leq T_i \leq t+\Delta | T > t]}{\Delta}$

A partir de ce taux de hasard, nous calculerons ensuite les fonctions de survies relatives :

$$S_i(t) = \exp\left\{-\int_0^t \lambda_1(x)dx\right\}$$

Nous déduirons par la suite les fonctions de répartitions :

$$F_i(t) = 1 - S_i(t)$$

Après avoir modifié le modèle actuel afin de l'adapter à une base de données contenant les informations du second co-emprunteur, nous avons calculé le taux de hasard empirique pour chacune des personnes présentes dans le portefeuille. Ce taux est calculé par couple et pour chaque période d'observation.

Nous rappellerons ici que l'intervalle total d'observation est $[t_0, t_d]$ avec t_d l'année de la dernière observation des individus constituant le couple. t_d Correspondra donc à l'année :

- De fin de contrat si l'assuré arrive à l'échéance de son prêt.
- De décès si un tel sinistre arrivait.
- De résiliation du contrat d'assurance si l'assuré quittait le portefeuille pour une telle raison.

Nous obtenons ainsi la base suivante.

Clef_couple	ID_Ind_1	ID_Ind_2	Age_Ind_1	Age_Ind_2	CSP_Ind_1	CSP_Ind_2	IMC_Ind1	IMC_Ind2	Taille_Ind_1	Taille_Ind_2
1	1	13	45	43	CSP1	CSP1	18	19	173	167
2	2	14	51	43	CSP2	CSP2	18,5	20	165	178
3	3	15	28	27	CSP3	CSP3	19	21	162	189
4	4	16	19	22	CSP4	CSP3	26	24	178	165
5	5	17	21	45	CSP3	CSP4	27	22	189	149
6	6	18	35	22	CSP1	CSP1	18,9	19	155	198
7	7	19	54	52	CSP2	CSP3	17,2	20	190	185
8	8	20	19	21	CSP1	CSP1	24	23	176	176
9	9	21	33	31	CSP1	CSP4	25	29	169	170
10	10	22	38	36	CSP3	CSP2	23,4	22	156	190
11	11	23	26	20	CSP2	CSP2	25	21	167	185
12	12	24	34	34	CSP1	CSP1	22	19	186	162

Figure 183 - Base de données

5.4.1 Robustesse des données

Tandis que de nombreuses méthodes statistiques reposent sur l'hypothèse de normalité, nous observons en pratique des données de sinistralité qui ne se conforment pas toujours à cette hypothèse. Soulignons que dans la réalité, les comportements étudiés s'éloignent bien souvent de la normalité. Cela peut s'expliquer par l'épaisseur des queues de distributions des phénomènes que nous tentons de décrire donc de leurs comportements dans les valeurs extrêmes. Face à cette non-normalité, l'asymétrie des sinistres, l'aplatissement de la distribution ou encore des corrélations segmentées par zone, nous devons trouver une méthode alternative aux tests statistiques traditionnels. Les copules permettent de s'accommoder à la non-normalité des variables. Afin de s'en convaincre, nous chercherons dans un premier temps à tester la normalité des taux de hasards empiriques. Nous effectuerons ensuite une étude de la dépendance des sinistralités au sein des couples, en commençant par des premiers tests de corrélation linéaire ou de concordance. La notion de dépendance est bien plus large que la relation linéaire. En effet, si deux variables sont indépendantes, cela implique qu'elles sont non-corrélées. Tandis que la non-corrélation linéaire de deux variables aléatoires autres que Gaussiennes ne signifie pas forcément une absence de dépendance.

5.4.2 Le cas du décès

5.4.2.1 Test d'indépendance

Tout d'abord, nous appliquons un test d'indépendance des taux de décès instantanés au sein des couples.

Le résultat ci-dessous permet de rejeter l'hypothèse d'indépendance au sein d'un duo de co-emprunteurs.

Variable	Valeur
Statistique	2.020419
P-valeur	0.04333993

Figure 184 - Test d'indépendance des taux de décès

Calculons maintenant la corrélation entre les taux de hasard au sein des couples du portefeuille. Le résultat n'est pas pertinent, car la corrélation de Spearman ainsi que le taux de Kendall sont très faibles. Nous obtenons en effet respectivement $\begin{pmatrix} \varphi = 0{,}091 \\ p\_valeur = 0 \end{pmatrix}$ et $\begin{pmatrix} tau = 0{,}091 \\ p\_valeur = 0 \end{pmatrix}$. Nous concluons qu'il existe au moins un lien linéaire au sein des couples étudiés. Cela semble en outre confirmer l'intérêt de faire une réduction aux couples.

5.4.2.2 Recherche de la copule optimale

Taux de hasard

Ce résultat, bien que concluant, ne suffit pas pour considérer l'ensemble des liens qui puissent exister au sein d'un dossier sur deux têtes. Nous cherchons alors à trouver la copule optimale qui lierait ces taux de hasard afin de savoir s'il existe un autre type de lien entre ces derniers. Ceci pourrait nous permettre de simuler nos résultats à partir d'une copule, ce qui rendrait les choses plus aisées.

Le test d'adéquation suggère donc la copule de Student avec $\begin{pmatrix} \varphi = 1 \\ Df = 2 \\ p\_value = 0{,}01 \end{pmatrix}$ avec un tau de Kendall de 0,13. Ces résultats nous permettent d'accepter l'hypothèse d'adéquation du modèle. Nous observons d'autre part une déviance de $D_{\hat{C}} = 0{,}00039$ associé à $\varphi = 1$.

Nous décidons d'appliquer l'adéquation aux données, mais le résultat est peu satisfaisant. En effet, lorsque nous comparons le graphe des simulations des taux de hasard par rapport aux taux de hasard empiriques, nous remarquons que la copule de Student s'adapte très bien au niveau des petites valeurs, mais surestime clairement les valeurs au-delà d'un certain seuil. En effet, nos données étant très déséquilibrés (nous avons un taux de décès moyen de 0,7% pour les hommes et 0,6% pour les femmes), la copule s'adapte bien pour prédire la survie mais sous-estime les décès.

Taux de hasard cumulé

Afin de réduire le déséquilibre, nous avons tenté par les mêmes étapes de modéliser le taux de hasard cumulé. Cela réduit non seulement la taille de la base, mais réduit également le déséquilibre des données.

Le test de sélection de la copule optimale nous propose exactement le même résultat que précédemment, et l'adéquation est légèrement meilleure. Cependant, ceci ne suffit pas pour prédire les rares taux de décès positifs qui restent malgré tout à la marge.

Lors de la recherche de la copule optimale sur les taux de décès instantané, le logiciel R a proposé la copule de Student de paramètres $\begin{matrix} \varphi = -0{,}005 \\ Df = 2 \end{matrix}$. Nous nous sommes de ce fait intéressé aux taux de décès cumulés. En effet, ce taux permet de comparer les observations sur l'exposition totale de chaque paire d'individus. Cette modification s'est révélée infructueuse étant donné que la copule proposée par la suite est la copule bivarié de Student.

Variable	Valeur
Famille	2
Nom	Student
Paramètre 1	1
Paramètre 2	2
Taux de Kendal copule	0.99
Taux de Kendal empirique	0.99
P-valeur	<0.02
AIC	-290160.6
BIC	-29143.79

Figure 185 – Paramètres de la copule de Student

Nous pouvons lire son taux de Kendal (qui est de 0.99) alors que le taux de Kendal de la copule empirique est de 0.01. La valeur de l'adéquation est de 0.04, une valeur qui est inférieure à 0.05 donc nous pouvons accepter l'hypothèse nulle de l'adéquation de la copule de Student aux données empiriques.

De plus, si nous regardons les rangs de nos valeurs, nous pouvons constater une dépendance de nos données au niveau des valeurs extrêmes.

La visualisation de la copule de Student nous permet également comparer ces informations de manière graphiques.

5.4.2.3 Visualisation de la copule

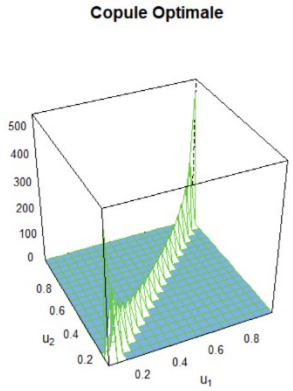


Figure 186 - Visualisation de la copule

Le graphe de la copule ressemble à priori à la copule indépendante, cependant si l'on se concentre sur les quatre coins du graphique, nous pouvons observer un lien qui se dessine discrètement dans ces zones.

Pour la plupart des taux de hasard cumulés, la copule de Student prédit de bonnes valeurs. Cependant, elle ne prédit que les petits taux de hasard. L'adéquation globale ne convient pas.

5.4.3 Le cas de l'incapacité de travail

5.4.3.1 Test d'indépendance

Tout d'abord, nous calculons la corrélation entre les taux de hasard au sein d'un couple. Le résultat n'est pas efficient, car la corrélation de Spearman est très faible.

Cela dit, ce résultat ne suffit pas pour considérer ces deux variables comme indépendantes. Nous faisons alors un test d'indépendance.

Les résultats sont pertinents. Ces taux de hasards cumulés ne sont pas indépendants.

Variable	Valeur
Statistic	20.57
P-valeur	0

Figure 187 - Taux de hasards cumulés

Avec une value plus petite que 5.10^{-2}, nous pouvons rejeter l'hypothèse nulle d'indépendance.

5.4.3.2 Choix de la copule optimale

Dans cette partie, le test sera seulement effectué sur les taux de hasard cumulés.

Lors de la recherche de la copule sur les taux de décès cumulés, nous avons obtenu les résultats suivants :

Variable	Valeur
Famille	234
Nom	Rotated Tawn de type 2 tournée à 270 degrès
Paramètre 1	-20
Paramètre 2	0.87
Taux de Kendal copule	-0.83
Taux de Kendal empirique	-0.86
P-valeur	<0.01
AIC	-11114.43
BIC	-1107.31

Figure 188 - Paramètre de la copule optimale

Ceci signifie que la copule optimale est la copule de tawn de type 2, tournée à 270 degrés. Son taux de Kendal est de 0,83 contre un taux empirique de 0,87 ce qui approche assez bien la réalité. La valeur de la valeur est de 0.01 qui est inférieure à 0.05 donc ce test est significatif. Le choix de la copule a été basé sur les indices AIC et BIC comme expliqué précédemment. Cette sortie R estime le paramètre param1 = -20 et le param2 = 0.87. Ces paramètres seront utilisés dans la suite de l'étude cependant l'adéquation globale ne permet pas de choisir cette copule.

Pour rappel, le hasard cumulé ou hasard intégré (cumulative hazard, integrated hazard) est défini par :

$$\Lambda(t) = \int_0^t \lambda(x)dx = -\ln S(t)$$

Dans cette partie, nous avons tenté de trouver un lien au sein des taux de hasards cumulés des conjoints.

L'adéquation pour le taux de hasard cumulé ne prédit pas de lien au sein des copules.

5.4.4 Résultats

5.4.4.1 La sinistralité décès

Nous remarquons que la copule optimale est la copule de Student. Ce résultat suggère ainsi la copule de Student de paramètres ($\nu = 4{,}99$ Df= -0,0013) comme la copule idéale afin d'approcher la modélisation de la loi jointe des deux risques. Les paramètres de la copule sont estimés par la méthode d'itau.

La méthode "itau" est une méthode d'optimisation utilisée pour calibrer une copule de Student. Elle consiste à minimiser l'information de Kullback-Leibler entre la copule calibrée et la copule de référence (ici, la copule de Student) en utilisant une méthode d'optimisation itérative basée sur l'algorithme de Newton-Raphson.

La copule de Student est une copule spéciale qui utilise une distribution de Student à deux paramètres pour décrire la dépendance entre deux variables aléatoires. Les deux paramètres de la copule de Student sont :

- ν : Le nombre de degrés de liberté, qui contrôle la forme de la distribution de Student. Plus le nombre de degrés de liberté est élevé, plus la distribution de Student se rapproche de la distribution normale.
- ϱ : Le coefficient de corrélation qui mesure l'intensité de la dépendance entre les deux variables aléatoires. Lorsque ϱ est proche de 1, cela suggère une forte dépendance positive entre les deux variables, tandis que lorsque ϱ est proche de -1, on notera alors une forte dépendance négative entre les deux variables observées. Enfin, un ϱ proche de 0, suppose que la dépendance n'est pas significative entre les deux variables considérées.

5.4.4.2 La sinistralité incapacité de travail

Nous remarquons que la copule optimale est la copule de tawn de type 2, tournée à 270 degrés comme la loi jointe optimale qui puisse représenter les deux risques.

Comme l'échantillon est parcimonieux et très hétérogène, les résultats obtenus ne sont pas pertinents. De plus, les individus de la base n'ont pas tous la même exposition ce qui entraine un biais dans l'analyse de la corrélation. D'autre part, nous ne pouvons pas améliorer la modélisation. L'idéal aurait été de

mettre en place l'adéquation des copules sur plusieurs groupes d'individus afin d'utiliser une base d'observation plus homogène et de prendre les mêmes expositions au sein de chaque population étudiée. Cependant, cela réduirait fortement le nombre de sinistres observés, au point où la loi des grands nombres ne serait plus applicable. Il est ainsi plus prudent de changer de modèle.

5.4.5 Conclusion

L'analyse par les copules nous montres des signaux de liens au sein des couples. Cela dit les contraintes de la structure du portefeuille ne nous permettent pas d'arriver à des résultats satisfaisants. En effet :

- Les données sont très déséquilibrées du fait que le risque de mortalité soit très faible, et qui est encore plus faible au sein de notre portefeuille dû à la sélection des assurés effectuée à la souscription.
- La notion de corrélation n'est pas vraiment adaptée à un modèle de survie car elle n'est pas stable dans le temps. Nous pouvons observer une corrélation parfaite entre deux vecteurs de personnes ayant survécu jusqu'à t_{n-1}, mais le décès de l'une des deux à un instant t_n pourrait ainsi tout remettre en cause.
  Une des solutions serait donc de calculer la corrélation par classe (en prenant idéalement des cohorte couples observés sur la même période) mais cette méthode est très difficile à mettre en place du fait du manque de données, de la complexité computationnelle et du niveau de technicité demandé.
- Il y a un bais qui est provoqué par la structure du contrat.

D'autre part, nous savons que tous les couples d'assurés qui sont totalement couverts au premier décès voient le conjoint survivant quitter le portefeuille au premier sinistre. Ceci inhibe assurément l'observation du second décès. De plus, étant donné que la variable cible (le décès du second conjoint) est dépendante de la variable « de censure », nous ne pouvons pas appliquer les modèles de Cox avec censure à droite par exemple. Ici, il serait plus approprié d'utiliser une modélisation avec la méthode des risques compétitifs.

5.5 Modélisation du risque compétitif

Dans cette partie, nous utiliserons le modèle LightGBM, explicité dans la section 3.7.4, afin de simuler les taux de hasard des assurés en tenant compte des données explicatives de leurs conjoints. La simulation se fera jusqu'à la fin du contrat de chaque individu. Cette méthode surestime la fidélité des assurés au sein du portefeuille. On peut la considérer « toutes choses égales par ailleurs ».

Nous appliquerons par la suite les formules du risques compétitifs définis dans la section 3.11 afin de calculer la fonction de survie et, par la même occasion, la probabilité de décès de chaque individu tout en tenant compte du risque probable encouru par son conjoint. Ce risque porté par le conjoint représente un risque compétitif total lorsque la tête décédée est couverte à 100%. En effet, lors de l'indemnisation du sinistre, le contrat est clôturé car le capital restant dû est remboursé dans son intégralité. A partir de cet évènement, le décès du conjoint survivant peut intervenir à tout moment, cependant cette information ne sera jamais observée par l'assureur. Même s'il existait un lien plus fort que celui que nous avons détecté au sein des taux de hasards des conjoints, cette information est cachée par la survenance du risque compétitif total. Cela biaise donc l'estimation de la fréquence de mortalité au sein du portefeuille.

Dans le cas où le taux de couverture d'un couple est dans l'intervalle $[100\%, 200\%[$, mais que la quotité de l'assuré décédée est inférieure à 100%, alors nous continuerons à observer ce dernier, donc nous n'avons pas de biais sur la fréquence. Cela dit le risque encouru au décès du second assuré est proportionnel à $1 - q_d$.

Avec q_d le ratio de couverture affecté à l'individu ayant subi le sinistre.

Or, le risque de l'individu survivant a été estimé proportionnellement à son taux de couverture q_s, donc nous nous retrouvons à estimer un risque $R_s = q_s.CRD$ tandis que le risque réel est de $R_s = (1 - q_d)CRD$

Avec $(1 - q_d) < q_s$

Pour résumer le problème, lors de la souscription nous appliquons une prime à chaque individu contractant une assurance emprunteur. Cette tarification est faite de manière individuelle et chaque

individu peut choisir son taux de couverture. Mais lorsque nous appliquons ce principe à un couple d'assurés, cela pose un problème dans l'estimation de leur risque exact. Lorsque le taux de couverture est supérieur à 100%, nous devons tenir compte du prorata joint qui parfois confère une mauvaise estimation du risque lors de la souscription.

Rappelons que la construction de la base de données présente les informations liées aux couples par période d'observation sur un an et par individu. Chaque ligne représente un individu vivant, pour une année d'observation donnée. Ce qui signifie que nous observons une ligne par âge de l'assuré. Ce modèle fait ainsi l'hypothèse que le taux de décès instantané est constant par unité de période donc constant par morceau. Enfin, l'unité de segmentation correspond à une année.

5.5.1 Résultats

Après la modélisation par LightGBM du risque considéré (décès, ITT) tête par tête en tenant compte des variables explicatives du conjoint, nous appliquerons une surcouche de correction pour le calcul de la prime par tête.

Il s'agit de la formule tenant compte des risques compétitifs au sein d'un couple. En effet, dans le cas où l'un des partenaires est assuré à 100%, son décès entrainera inéluctablement la clôture du contrat et le remboursement du prêt. Dans ce cas, nous pouvons dire que le risque décès du premier conjoint est en compétition avec le risque du second.

La tarification proposée ci-dessous tient donc compte de ce type de relation au sein d'un couple.

Tout d'abord rappelons les formules des fonctions de survie par le modèle de Poisson Equivalent décrit dans la section 3.7

$$S_t^{(1)} = e^{-\int_0^t \lambda_u^1 du}$$

$$S_t^{(2)} = e^{-\int_0^t \lambda_u^2 du}$$

Dans le cas d'une assurance sur une tête, le calcul de la prime est :

$$PP^1 = \int_0^T S_t \lambda_t \times CRD_t \, dt$$

Dans le cas d'un couple, le calcul de la prime devient :

$PP^1 = \int_0^T S_t^{(1)} \times \lambda_u^1 \times \mathbb{E}(Severitet_t) \, dt$, avec T la durée du contrat.

Avec $Severite_t = \begin{cases} q^1 \times CRD_t \; si \; le \; second \; conjoint \; est \; vivant \\ \min(1-q^2, q^1) \times CRD_t \; si \; le \; secont \; conjoint \; décède \; avant \; l'instant \; t \end{cases}$

Cela nous conduit au calcul final de la prime pour chaque tête qui est :

$$PP^1 = \int_0^T S_t^{(1)} \times \lambda_t^1 \times (1 - S_t^{(2)}) \times min(1-q^2, q^1) \times CRD_t \, dt + \int_0^T S_t^{(1)} \times \lambda_t^1 \times S_t^{(2)} \times q^1 \times CRD_t \, dt$$

Pour calculer PP^2, il faudra appliquer la même formule que celle de son conjoint.

5.5.1.1 Les primes décès

Nous pouvons ainsi observer le résultat graphique de la dispersion du chiffre d'affaires hors taxe et hors marge de sécurité en fonction des classes d'âge et de l'exposition.

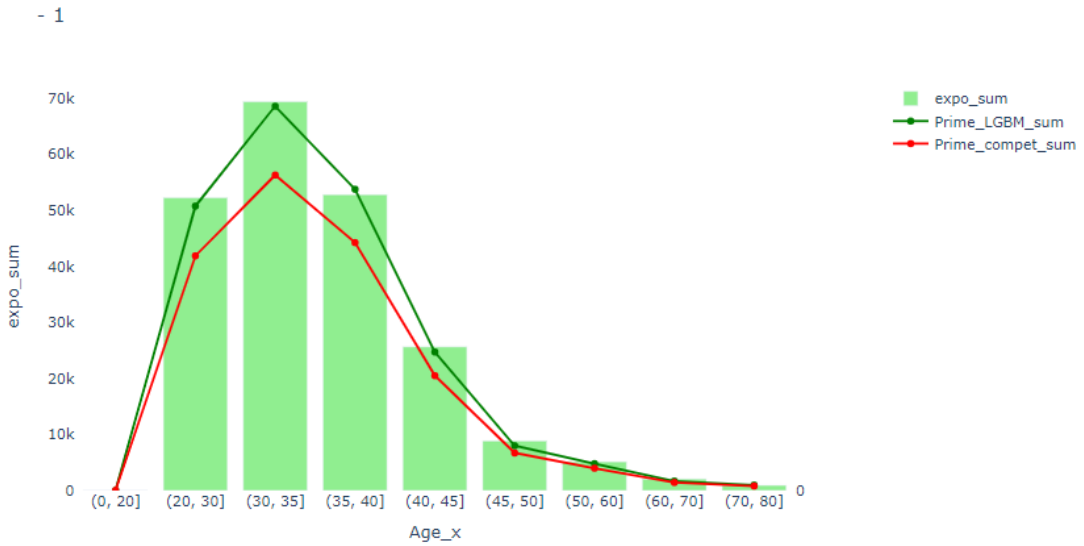


Figure 189 - Résultat graphique de la dispersion du CA hors taxe et hors marge pour le décès

Nous remarquons que globalement les primes corrigées par la formule des risques compétitifs sont légèrement moins fortes que les primes obtenues par le modèle LightGBM.

La quantification de la différence globale est de 12%. Ceci est proche de la réduction de 10% appliquée par la compagnie lors de la souscription d'une assurance emprunteur en couple. Cependant, la réduction se fait linéairement sur tout le portefeuille, tandis que la formule des risques compétitifs permet d'amener une correction uniquement sur l'estimation des risques des assurés concernés. Cette application permet une meilleure équité au détriment de la mutualisation.

5.5.1.2 Les primes pour l'arrêt de travail

Pour l'arrêt de travail, le graphe observé montre que la correction est faite au même endroit que le décès.

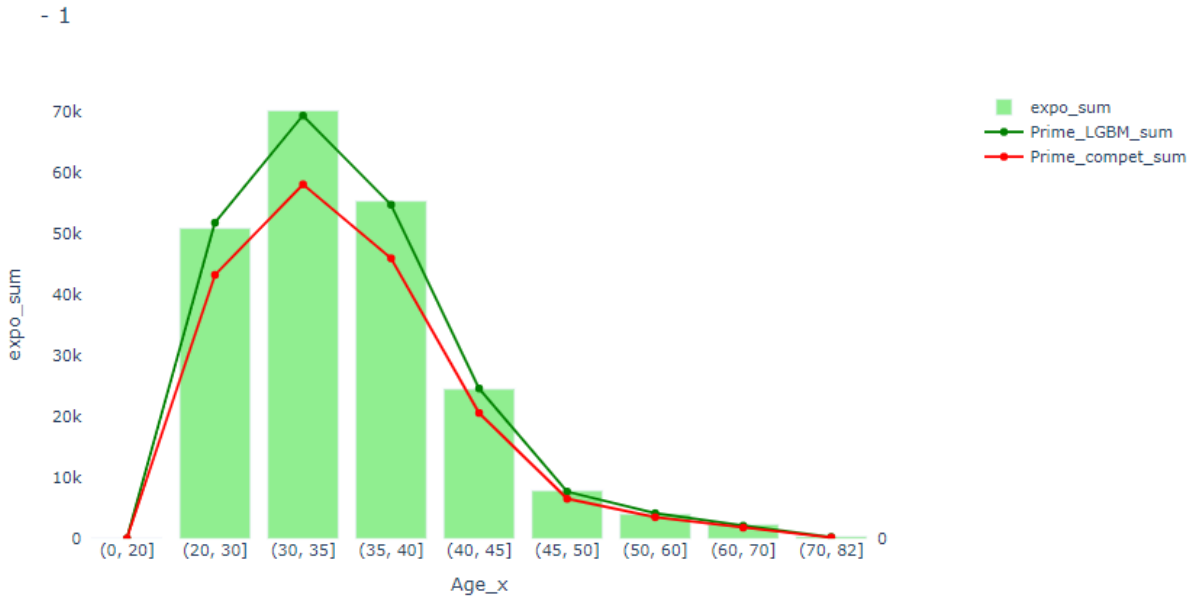


Figure 190 - Résultat graphique de la dispersion du CA hors taxe et hors marge pour l'arrêt de travail

Cette information signifie que le risque compétitif agit sur les mêmes individus et de la même manière. Ceci corrobore le fait que les quotités soient appliquées sur le capital assuré, donc souligne la cohérence de la modélisation.

Lorsque nous calculons la différence entre les primes pures modélisées par le modèle LightGBM et celles qui disposent d'une surcouche avec la formule des risques compétitifs, nous obtenons une différence de 13%. Nous constatons de nouveau que cette différence est proche de la réduction appliquée sur le prime global lors de la souscription à l'assurance des emprunteurs en couple pour le produit étudié.

Nous rappellerons tout comme dans le cas du décès que l'application de cette méthode rend l'estimation de la prime pure plus équitable qu'une simple réduction globale au sein de la population du portefeuille.

5.5.2 Conclusion

Lors de la souscription en couple au produit étudié, les assurés ont le choix du taux de couverture appliqué sur chaque tête. En d'autres termes, ils peuvent choisir de s'assurer ensemble au-delà de 100% du capital emprunté. Ces décisions sont souvent prises lorsqu'un individu essaie de protéger son conjoint au cas où celui-ci deviendrait insolvable, ou défaillant (décès, ITT …). Dans ce cas, la compagnie d'assurance devra prendre en compte le risque augmenté de ce contrat de manière équitable. L'usage de la méthode des risques compétitifs permet ainsi de mieux estimer les engagements des assurés afin qu'ils ne puissent jamais dépasser celui de l'assureur.

6 POUR ALLER PLUS LOIN

6.1 Les limites des modèles

Cette étude a été effectuée sur un échantillon donné arrêté dans le temps. Si le début de notre étude concorde à la mise en place du produit étudié, les dates d'arrêt ont provoqué une censure à droite.

Ce portefeuille continue d'exister tandis que la date d'extraction des données est nettement dépassée.

Cette étude est fondée sur les informations dont nous disposions au moment de l'extraction. Elle ne prend donc pas en compte la probabilité d'augmentation de l'espérance de vie dans le futur, ni l'hypothèse d'une surmortalité liée à une pandémie étant donné que nous avons supprimé les données relatives à ce type de risque de notre base de données.

D'autre part, nous faisons l'hypothèse de non-évolution des variables médicales au cours du temps (et donc au fur et à mesure de la progression de l'âge des individus). C'est pour cela que nous pouvons observer un bais dans la modélisation.

Une suggestion sur l'accès à des données plus complètes peut s'engager sur cette base : si la possibilité de produire des tables de mortalité efficientes sur la base d'informations non temporelles, l'accès à des données suivies au cours du temps serait un progrès.

Du point de vue statistique et informatique, le modèle LightGBM peut justement permettre de décrire le risque de mortalité par les valeurs explicatives mais aussi par l'évolution de celles-ci dans le temps. Au moment de l'écriture de ce Mémoire et de l'avènement du Big Data dans de nombreux domaines, une nouvelle famille de logiciels commence à faire son apparition : des logiciels spécialement mis en œuvre pour traiter de très grandes quantités de données. Par exemple, sur une version améliorée du logiciel R, pour un GLM et sur le même nombre d'observations, le temps de calcul est réduit de 100 fois. En outre, la segmentation plus ou moins poussée des variables en catégories peut éventuellement mener à saturation dans les modèles utilisés. Un choix pertinent du nombre total de catégories utilisées doit être effectué pendant la modélisation. Ce dilemme n'est pas nouveau pour l'actuaire, mais restera présent avec la mise en place de Solvabilité II, qui exige aux actuaires de démontrer que les décisions de modélisation (comme la granularité de la segmentation) n'influent pas sur les best estimate, ou bien de chiffrer l'influence. Une contribution récente à ce sujet dans le cas du « *risque de mortalité* » se trouve dans Alho90 sur la création de taux prospectifs.

6.2 Type de lien au sein du modèle LightGBM

Lorsque nous estimons la corrélation linéaire entre les prédictions des taux de hasard au sein des couples, nous quantifions une corrélation linéaire de 26% pour le décès. Les prédictions du taux d'incidence en arrêt de travail, présentent quant à elle, une corrélation de 79%.

De plus si nous traçons les Q-Qplot, nous obtenons les graphiques suivants :

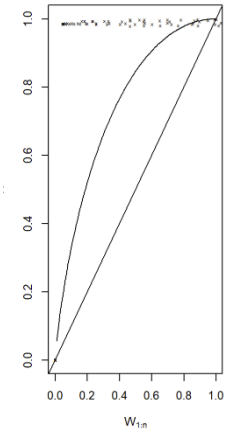


Figure 191 - Lien entre les taux de décès au sein d'un couple

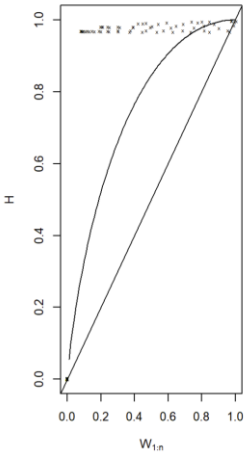


Figure 192 - Lien entre les taux de décès au sein d'un couple

Nous pouvons voir que les lignes sont courbées ce qui signifie que les taux de hasards prédits au sein d'un couple sont dépendants. De plus, la courbe est au-dessus de la diagonale, ce qui implique que la dépendance est positive pour les deux types de risques.

Ces résultats démontrent que l'ajout des variables du conjoint pour la prédiction du taux de hasard constitue une réelle valeur ajoutée dans la modélisation des risques décès et arrêt de travail au sein d'un couple.

Si nous cherchons d'autres types de lien en utilisant les copules, nous obtenons les résultats suivants :

6.3 Le cas du décès

6.3.1 Sélection de la copule optimale

Lors de l'estimation de la copule à partir des prédictions obtenues, la meilleure copule sélectionnée est Survival BB1.

Variable	Valeur
Famille	17
Nom	Survival BB1
Paramètre 1	0.05
Paramètre 2	2.07
Taux de Kendal copule	0.53
Taux de Kendal empirique	0.54
P-valeur	<0.01
AIC	-23006.2
BIC	-22989.9

Ces résultats d'adéquations sont confirmés par le test de Vuong-Clark qui présente la 17ième copule comme étant la meilleure. En effet, cette copule correspond bien à BB1 Copula dans la liste des copules de R.

Famille		4	5	6	7	8	14	16	17	18
Test	Vuong	-4	-2	-8	2	-6	6	0	8	4
	Clarke	-4	2	-8	0	-6	6	-2	8	4

Figure 193 - résultat du test de Vuong-Clark

La copule BB1 appartient à la classe des copules bivariées de Clayton. Elle est définie par une fonction de répartition bidimensionnelle $C(u,v)$ qui relie les fonctions de répartition marginales $F_1(u)$ et $F_2(v)$ des deux variables aléatoires de survie.

$$C(u,v) = [1 + \frac{(\alpha * u * v)}{(1-u)*(1-v)}]^{\frac{-1}{\beta}} \text{ pour } \beta \neq 0$$

Où

- u et v sont respectivement les fonctions de répartition $F_1(u)$ et $F_2(v)$.
- α, également appelé « paramètre de forme », contrôle la forme globale de la copule de survie BB1. Il varie généralement dans l'intervalle $]-\infty,+\infty[$. Les valeurs de α déterminent si la dépendance entre les temps de survie des deux événements est positive ($\alpha > 0$) ou négative ($\alpha < 0$). Dans notre cas $\alpha = 2.07$ ce qui induis une dépendance positive pour les taux de décès.
- β, également appelé « paramètre de queue », contrôle la décroissance des queues de la copule de survie BB1. Il est généralement positif. Les valeurs de β influencent la force de la dépendance entre les valeurs extrêmes des temps de survie des deux événements. Dans notre cas, $\beta = 0.07$ donc la dépendance au niveau des queues n'est pas très forte.

6.3.2 Visualisation de la copule Survival BB1

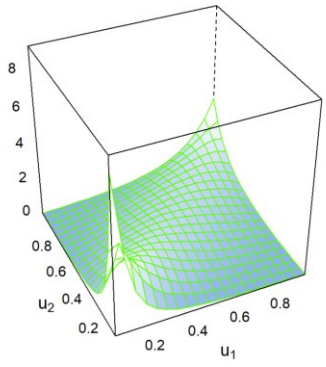
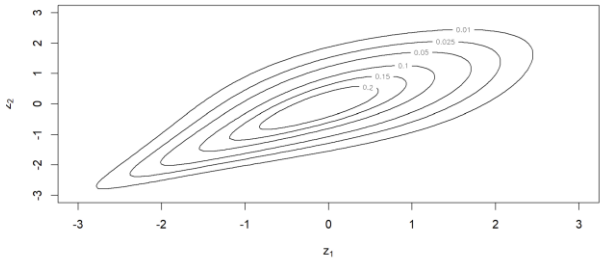


Figure 194- Copule BB1 en 3D Figure



195- Section de la copule survival BB1

Si nous regardons la section de la copule BB1, nous remarquons qu'il y a un lien plus fort vers les petites valeurs. En effet, plus les lignes de la section sont éloignées moins il y a de lien. En effet, les lignes s'éloignent un peu plus vers les valeurs extrêmes.

La section indique également des contours incurvés, ce qui fait référence à une dépendance non linéaire.

6.3.3 Simulations

Si nous comparons les données historiques et simulées, nous pouvons voir le niveau d'adéquation de la copule Survival BB1 par rapport à la sinistralité empirique.

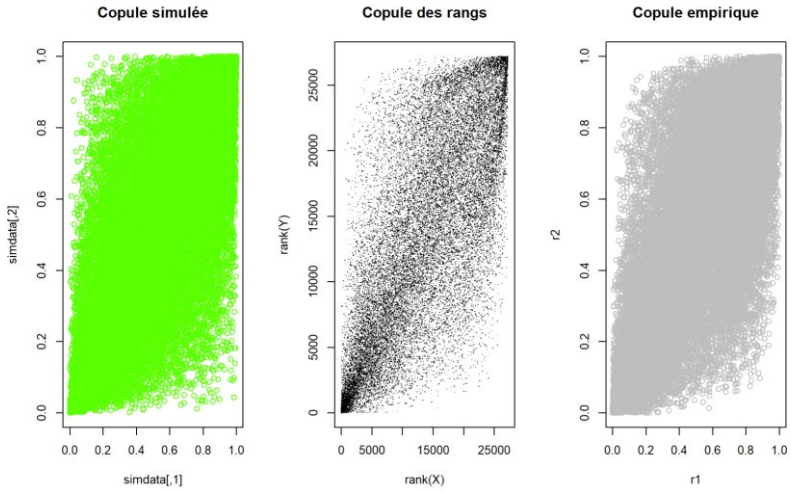


Figure 196 - Comparaisons des copules

Si nous regardons la forme des copules, nous pouvons voir que celles-ci sont proches.

Si, de plus, nous regardons le score des simulations, nous obtenons une adéquation de 93% (en termes de variance expliquée).

6.4 Le cas de l'arrêt de travail

6.4.1 Sélection de la copule optimale

Les prédictions sur les sinistres arrêt de travail sont corrélés par la copule de Gauss. En effet, lors de l'adéquation, la copule optimale est donnée par les résultats suivants :

Variable	Valeur
Famille	1
Nom	Gaussian
Paramètre 1	0.78
Taux de Kendal copule	0.57
Taux de Kendal empirique	0.56
P-valeur	<0.01
AIC	-15889.77
BIC	-15881.79

Figure 197 - Sélection de la copule optimale

Comme nous pouvons le lire, la recherche de la copule optimale nous propose la copule gaussienne. Ce résultat est confirmé par le test de Vuong-Clark ci-dessous :

Famille		0	1	2	3	4	5	6	13	14	16
Test	Vuong	-9	9	7	3	-3	1	-7	-5	5	-1
	Clarke	-9	9	7	-1	1	3	-7	-5	5	-3

Figure 198 - Résultat du test de Vuong-Clarke

La copule de Gauss, ou copule normale est quant à elle associée à une relation linéaire.

La formule de la copule de Gauss est : $C(u,v) = \Phi_2[\Phi^{-1}(u)\Phi^{-1}(v)]$, où :

- u et v sont respectivement les fonctions de répartition $\mathrm{F}_1(\mathrm{u})$ et $\mathrm{F}_2(\mathrm{v})$.
- Φ_2 est la fonction de répartition bivariée de la distribution normale standard bivariée.
- Φ^{-1} est la fonction quantile inverse de la distribution normale standard.
- ρ est le paramètre de corrélation de la copule gaussienne qui détermine la force de la corrélation. Dans notre cas $\rho = 0.78$ ce qui indique une corrélation forte. Cette correlation donnée par la copule rejoint la corrélation de Pearson calculée plus haut.

Ces résultats d'adéquations sont confirmés par le test de Vuong-Clark qui présente la 1ère copule comme étant la meilleure. En effet, cette copule correspond bien à la copule Gaussienne dans la liste des copules de R.

6.4.2 Visualisation de la copule Gaussienne

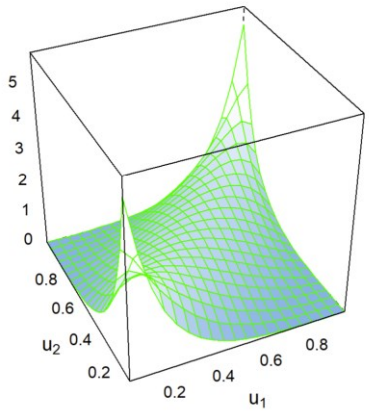


Figure 199- Copule normale en 3D

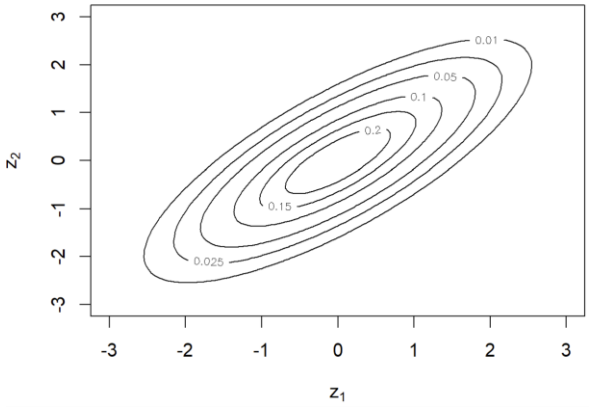


Figure 200- Section de la copule sélectionnée

La section de la copule présente une forme elliptique, ce qui sous-entend que la corrélation est non nulle. Les contours elliptiques sont dessinés sur le graphique. La forme et l'orientation de ces contours fournissent des informations sur la structure de dépendance. En l'occurrence, dans ce cas, les contours sont orientés vers le haut, ce qui présente une dépendance positive. Ces résultats rejoignent le coefficient de corrélation calculé plus haut (0.78) qui indique une forte corrélation positive dans la sinistralité arrêt de travail au sein des couples.

6.4.3 Simulations

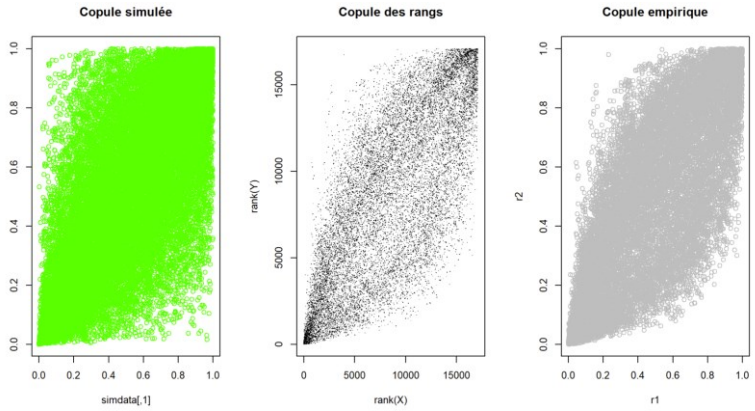


Figure 201- Simulations

Les graphes montrent que l'adéquation est d'un bon niveau. En effet, lorsque nous comparons la copule empirique ou celle obtenue à partir des rangs des taux d'incidence à la copule simulée, nous observons que les simulations sont proches de la sinistralité historique.

Si nous regardons l'adéquation, nous obtenons un score beaucoup trop petit, donc la copule Gaussienne ne permet pas de décrire le lien de manière homogène, mais il y a des zones dans lesquelles celle-ci peut expliquer le type de lien.

6.4.4 Autres modèles

6.4.4.1 Introduction

L'objectif de cette partie est de challenger le modèle individuel pour vérifier si l'adéquation globale peut être aussi bien mesurée voir améliorée en utilisant des modèles plus simples tels que les modèles linéaires généralisés (GLM). Etant entendu que les GLM sont capables d'ajuster des jeux de données aussi réduits que ceux utilisés dans ce mémoire.

La modélisation prédictive d'une variable cible se formalise comme suit :

$$y_i = g(\mu_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$$

Où :

- $g(\mu_i)$ est la fonction de lien.
- μ_i est la moyenne de la distribution de la famille pour l'observation i.
- $x_1, x_2, \ldots, x_n$ sont les variables explicatives pour l'observation i.
- $\beta_0, \beta_1, \ldots, \beta_n$ sont les coefficients associés aux variables explicatives $x_1, x_2, \ldots, x_n$.

En supposant disposer d'un échantillon de taille n, le but d'un GLM est de quantifier les paramètres $\beta_0, \beta_1, \ldots, \beta_n$ en maximisant le logarithme de la vraisemblance de la façon suivante :

$$\ell(\beta_0, \beta, y_i | x_i)_{1 \le i \le n} = \sum_{i=1}^{n} \log\left(f(y_i, \beta_0 + x_i'\beta, \phi)\right)$$

Où :

- β est le vecteur des paramètres du modèle.
- $g(\mu_i)$ est la valeur observée de la variable de réponse.
- f est la densité de probabilité pour la distribution de la famille associée.

6.4.4.2 Le modèle GLM

Ici, la modélisation consiste à prédire la variable aléatoire Y qui indique si l'individu est vivant ou décédé, tout en tenant compte de l'exposition de chaque observation. Cette variable binaire est modélisée par la fonction de lien Logit utilisée dans un GLM.

En utilisant les métriques décrites dans la section 3.11.6 et 5.2.5.1, et après apprentissage, l'algorithme GLM affiche une déviance expliquée de 12%, ce qui est moins performant que l'algorithme LightGBM. De plus, l'adéquation globale présente seulement 48% contre 95% pour l'algorithme LightGBM.

Modèle	$M_{i(GLM)}$	$M_{i(LightGBM)}$	Gain
Déviance expliquée	12%	18%	33%
RMSE	3,80%	4,30%	12%
MAE	0,08%	0,0095%	689%
Tot_pred/Tot_Obs	48%	95%	49%
MSE	0,15%	0,19%	23%

Figure 202 - Comparaison des métriques entre GLM et LightGBM

6.4.4.3 Les modèles de pénalisation

Afin d'améliorer l'ajustement, une pénalisation a été introduite pour optimiser l'adéquation.

En effet, lors des processus d'apprentissage, nous sommes souvent confrontés à un problème de sur-apprentissage.

Pour y remédier, nous pouvons utiliser des modèles de pénalisation comme les régressions Ridge, Lasso ou Elastic Net.

Ces modèles limitent l'impact de certaines variables en introduisant une pénalisation, réduisant ainsi l'erreur commise et le sur-apprentissage.

La pénalisation est une technique de régularisation qui agit en réduisant les coefficients de variables explicatives simplifiant de fait le modèle et diminue l'influence des variables non pertinentes.

Le sous-apprentissage quant à lui, peut être évité grâce à une sélection judicieuse du paramètre λ qui permet de prendre en compte toutes les variables explicatives impactantes.

- **Régression Ridge**

Cette méthode utilise la pénalité L2 pour améliorer l'adéquation du modèle en réduisant sa variance.

La fonction de coût à minimiser devient : $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}{\beta_j}^2, \lambda \geq 0$ avec β_j les coefficients du modèle linéaire classique.

Le but de la pénalisation est donc de trouver les coefficients β_j sous la contrainte de la minimisation de la fonction coût.

- **Régression Lasso**

Pour équilibrer avec Ridge qui utilise la pénalisation L2, la régression Lasso régularise en utilisant une pénalisation L1 qui contraint certains coefficients à converger vers 0.

La fonction de coût à minimiser devient : $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|$.

La régression Lasso encourage la sparsité ce qui permet la sélection automatique des variables améliore l'interprétabilité du modèle.

- **Elastic Net**

Il s'agit d'une combinaison entre la régression Lasso et la régression Ridge.

La fonction de coût à minimiser devient :$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}{\beta_j}^2$, avec λ_1 et λ_2 les hyperparamètres qui contrôlent respectivement les pénalités L1 et L2.

Cette approche réduit la variance du modèle tout en sélectionnant les variables le plus pertinentes en combinant L1 et L2. Cette méthode tire partie des avantages de la régularisation Lasso et celle de Ridge en offrant une meilleure performance lorsque les variables sont corrélées.

6.4.4.4 Mise en pratique de la pénalisation

En pratique, les valeurs optimales de λ_1 et λ_2 sont déterminées par validation croisée. Ces dernières permettant d'identifier la régularisation la plus appropriée.

Après détermination de λ_1 et λ_2, le résultat indique que la pénalisation la plus adaptée est Elastic Net. L'application de cette correction améliore l'ajustement de 12% à 14%.

Modèle	$M_{i(GLM)}$	$M_{i(Elastic\_net)}$	Gain
Déviance expliquée	12%	14%	14%
RMSE	4%	8,00%	53%
MAE	0,075%	0,29%	74%
Tot_pred/Tot_Obs	48%	52%	8%
MSE	0%	1,00%	85%

Figure 203 - Comparaison des métriques entre GLM et Elastic Net

Toutefois, l'ajustement globale en utilisant Elastic Net reste inférieur à celui obtenu avec LightGBM qui atteint un score de 95% contre 48% avec GLM et 52% avec Elastic Net.

6.4.4.5 Conclusion

Pour ce portefeuille avec ses propres spécificités, la méthode GLM même améliorée par Elastic Net n'atteint pas les performances de l'algorithme LightGBM utilisé dans ce mémoire.

6.5 Conclusion

La prise en compte de la dépendance lors de la modélisation de sinistres est un paramètre très souvent négligé, mais qui semble crucial dans la compréhension de certains phénomènes. L'ajout simple de cette information, nous a permis d'améliorer notre modélisation de 30% au sein des couples.

Cela semble d'autant plus pertinent que la proportion de couple de ce portefeuille est très importante.

Il s'agit d'autre part d'une population à moindre risque car elle représente des profils moins risqués et permet grâce à l'usage des quotités (hors assurance totale sur chaque tête) de diversifier le risque.

La copule de survie BB1 est une copule utilisée pour modéliser la dépendance entre deux variables aléatoires dans le contexte de la survie. Cette copule est couramment utilisée en statistiques de survie pour modéliser la dépendance entre les temps de survie de deux événements ou individus.

7 CONCLUSION GENERALE

Cette étude a été mise en place dans le cadre de l'estimation du risque décès et incapacité de travail au sein d'un couple d'emprunteur. Le but de ces travaux était de confronter le modèle de tarification actuel basé sur une tête, afin de savoir si l'on pouvait proposer des modèles plus performants en prenant en compte le lien au sein d'un couple. Nous avons tout d'abord pensé à capter le lien au sein de deux assurés liés par les mêmes conditions de vie en utilisant les variables explicatives du conjoint de la tête modélisée. Il en résulte un signal non négligeable du fait de l'amélioration de l'adéquation du modèle.

Nous avons observé d'autre part, un changement de la pertinence des variables explicatives à travers le graphique de SHAP. Cette dernière place en effet certaines variables explicatives du conjoint avant certaines variables de l'individu étudié. C'est ainsi que nous pouvons estimer le taux d'entrée en incapacité de travail en fonction de la catégorie socio-professionnelle du conjoint par exemple.

Ces résultats nous ont conduit à l'usage des copules. Il s'agit en effet d'un puissant outil mathématique, très pratique et facile à mettre en place. Après plusieurs essais afin de trouver un moyen d'assurer la meilleure adéquation possible, nous nous sommes rendu compte que les copules présentaient des limites quant à la nature du risque étudié. En effet, le très fort déséquilibre de la variable cible (taux de hasard), la structure des contrats, ainsi que les ratios de couvertures des personnes empruntant par paire censurent beaucoup d'informations dans le but de capter ce lien tant recherché avec les copules. D'autre part, utiliser un modèle de corrélation pour prédire un phénomène qui dépend du temps ne constitue pas un outil stable dû à nature de la variable cible.

C'est donc après cette prise de conscience que nous avons préféré nous tourner vers l'utilisation des risques compétitifs qui semblent traduire une partie du lien qu'il pourrait y avoir au sein de la sinistralité d'un couple d'assurés. En effet, cet outil nous a permis de corriger l'estimation du risque décès et incapacité de travail au sein d'un couple en considérant les quotités assurées pour chaque tête de la manière la plus juste.

Ces modélisations nous ont permis de voir l'intérêt d'étudier le risque non seulement au niveau individuel mais également au niveau couple. Elles permettront de mieux évaluer la structure de risque liée aux pairs d'assurés. Elles permettent en outre de capter l'évolution potentielle de cette population dans le portefeuille. Il restera à étudier quelle tarification proposer aux clients en tenant compte de ses résultats.

8 BIBLIOGRAPHIE

1. Christophe Clech, Modèle à risques compétitifs et analyse de propension appliqués à l'atteinte rénale aiguë en réanimation, HAL Id: tel-00566132, 2011.
2. Myriam Chabot, Concepts de dépendance et copules, CaMUS 4, 48 – 71, Université de Sherbrooke.
3. Jean Louis Cassablian, Mathématiques et calculs de l'assurance, 2020.
4. Michel Denuit, Arthur Charpentier, Mathématiques de l'assurance non-vie, Tome I, Principe fondamentaux de la théorie du risque, 2004.
5. Michel Denuit, Arthur Charpentier, Mathématiques de l'assurance non-vie, Tome II, Tarification et provisionnement, 2004.
6. Alvin E. Roth, The Shapley value, Cambridge University press, ISBN0-521-36177-X, 1988.
7. Cécile Cornudet, Les Echos, le rapport du risqué décès selon le nombre d'enfants, 2007.
8. Biostatistics: The Good, the Bad and the Ugly Réflexions méthodologiques et statistique, Modèles à risques compétitifs, 2022.
9. Grégoire Mercier, Modélisation des lois multidimensionnelles par la théorie des copules, 2006.
10. Camille CHAPUIS, Spécificités et enjeux de l'assurance emprunteur, 2013.
11. Tianqi Chen, Carlos Guestrin, XGBoost : a Scalable Tree Boosting System, University of Whashington, ISBN 978-1-4503-4232-2/16/08, 2016.
12. Boris Dilane Noumedem Djieuzem, De la communauté à l'individu : Quantifier pour comprendre l'impact des pratiques actuarielles sur la mutualisation, 2022.
13. Christian-Yann Robert : Théorie du risque, Notes de cours ENSAE IP Paris M2,2021.
14. Léonor Fasse, Le deuil des conjoints après un cancer : entre évaluation et expérience subjective, Num. national de thèse : 2013PA05H125, 2013.
15. Insee, L'espérance de vie s'accroît, les inégalités sociales face à la mort demeurent, 2011.
16. Statista, Taux de mortalité en France de 2004 à 2022, 2023.
17. WikiSrat, Agrégation de modèles, université de Toulouse, st-m-app-agreg (univ-toulouse.fr).
18. Salim Amoukou, Tangi Salaûn, Nicolas J.B Brunel, Accurate Shapeley Values fot explaining tree_based models, 2023.
19. German Rodriguez, cours Generalizes linear model, univerité de Prinston, chapitre 7 Survival model, 2010.
20. Mélanie Vonderschelden, Economie et statistiques, N°398-399, Homogamie socioprofessionnelle et ressemblance en termes de niveau d'étude : constat et évolution au fil des cohortes d'union.
21. Arthur Charpentier, Chapitre 6 Copules et risques multiples.
22. Marius Hofert, Ivan Kojadinovic, Martin Mâchler, Jun Yan, Elements of Copula modelinf with R,2017.
23. Deborah Seror, David Nkihouabonga Yengue, modélisation des comportements de rachat dans un contexte de risques compétitifs, 2013.
24. Assurance maladie, Le Tabagisme passif, tabac-info-service.fr.
25. Nayhalie Graffeo, Méthode d'analyse de la survie nette : utilisation des tables de mortalité, test de comparaison et détection d'agrégats spatiaux, DOI : 10.1002/sim.5493, 2014.
26. Georges Bresson, Modèles Econométriques de Durée, Master 2, Université Paris 2, 2021.
27. Spoorenberg, Brieuc ; Nguyen, Alisson. Corrélation entre nombres et coûts des sinistres. Faculté des sciences, Université catholique de Louvain, 2020.
28. V.Vaillon, Régression pénalisée : le Lasso, pbil.univ-lyon1, 2020
29. NGUYEN Alisson SPOORENBERG Brieuc, Corrélation entre nombres et coûts des sinistres, 2020.
30. Magalie Fromont, Apprentissage Statistique : Partie III, 2015.
31. Nicolas Baradel, Actuariat de l'assurance non-vie, 2024.

9 ANNEXES

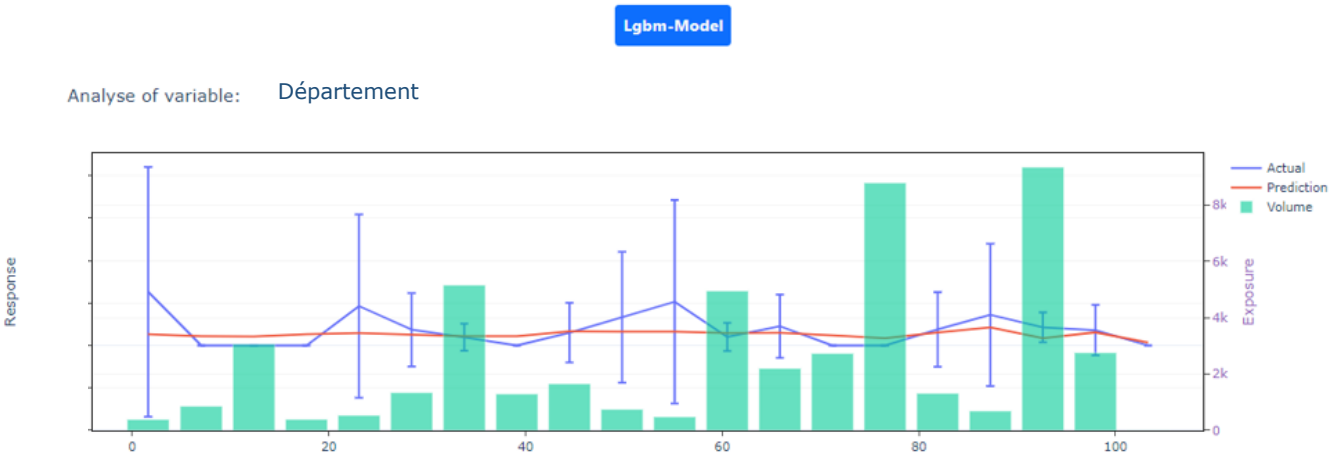


Figure 204 - Analyse de sinistralité en fonction du département du modèle couple

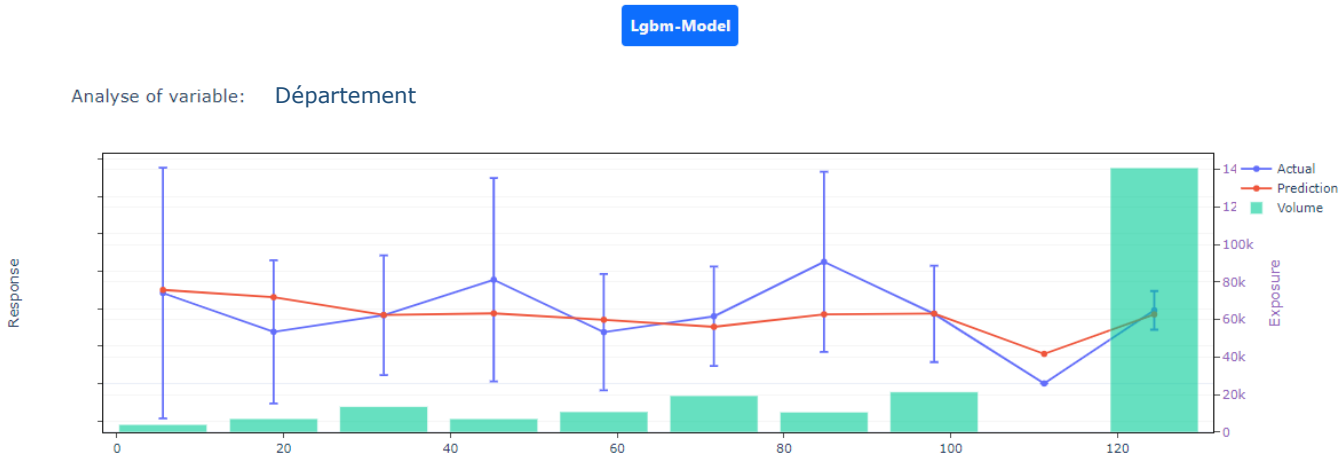


Figure 205 - Analyse de la variable département en fonction de l'exposition du modèle individuel

Lgbm-Model

Feature	Permutation_Importance
filter data...	
age_step	0.00052
K	0.00023
cd_sexe	0.00012
K_ass	0.00012
departement	0.00008
seniority	0.00005
dur_pret	0.00003
csp	0.00002
period	0.00002
nb_co_emprunteur	0.00001

Figure 206- Les valeurs des Permutations importance pour lé décès avec Mi

Shap Feature importance

Lgbm-Model

Feature	Avg Shap Effect
filter data...	
seniority	0.21279
age_step_x	0.14010
mensualite_x	0.11081
cd_sexe_x	0.10679
K_ass_y	0.10517
age_step_y	0.08804
K_ass_x	0.05863
dur_pret_x	0.03603
csp_y	0.03545
dep_katia	0.03443

Figure 207- Les valeurs des Permutations importance pour le décès avec Mc(x,y)

Lgbm-Model

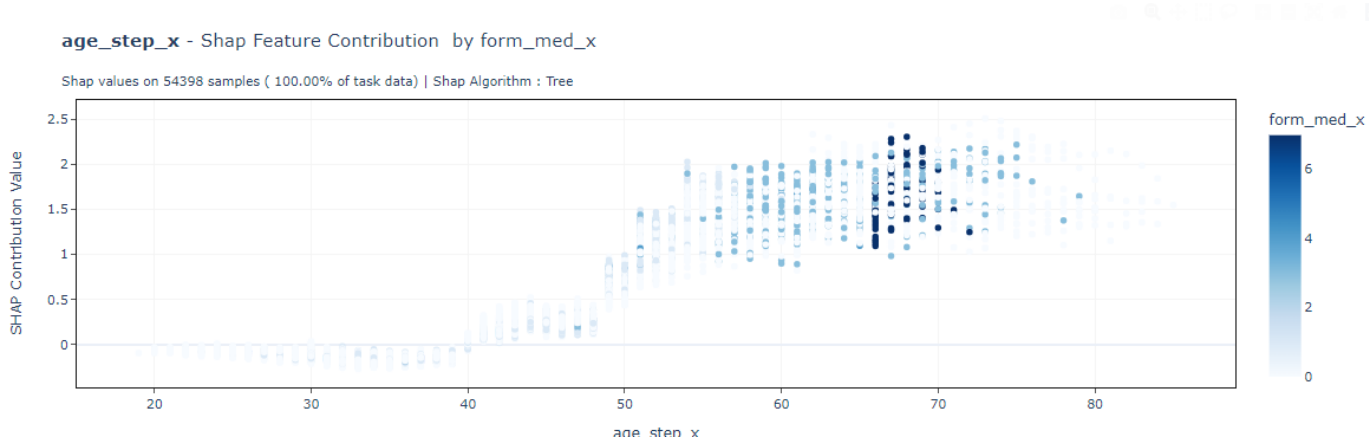


Figure 208- Contribution des formalités médicales en fonction de l'âge dans le modèle en couple

Lgbm-Model

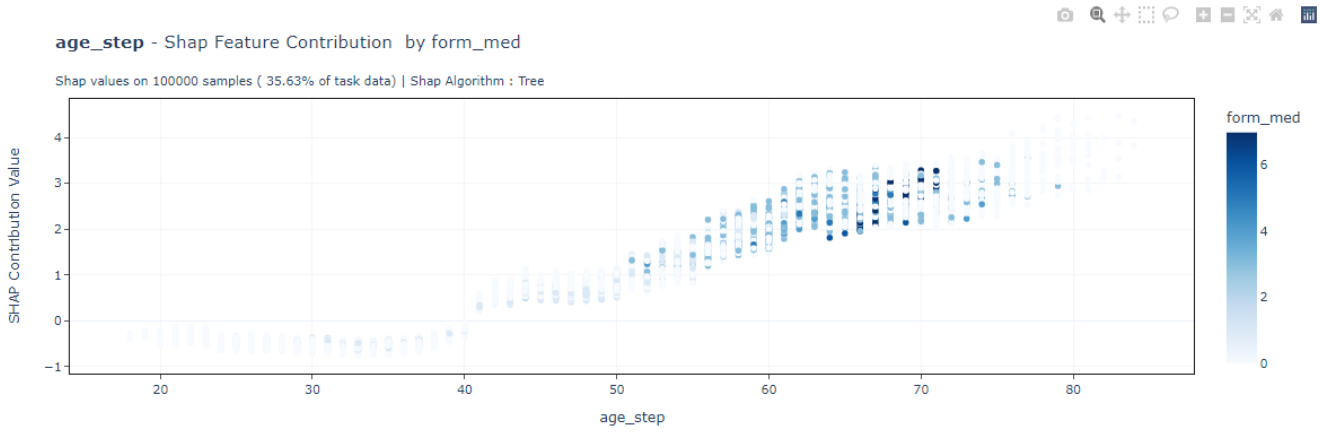


Figure 209- Contribution des formalités médicales en fonction de l'âge dans le modèle individuel

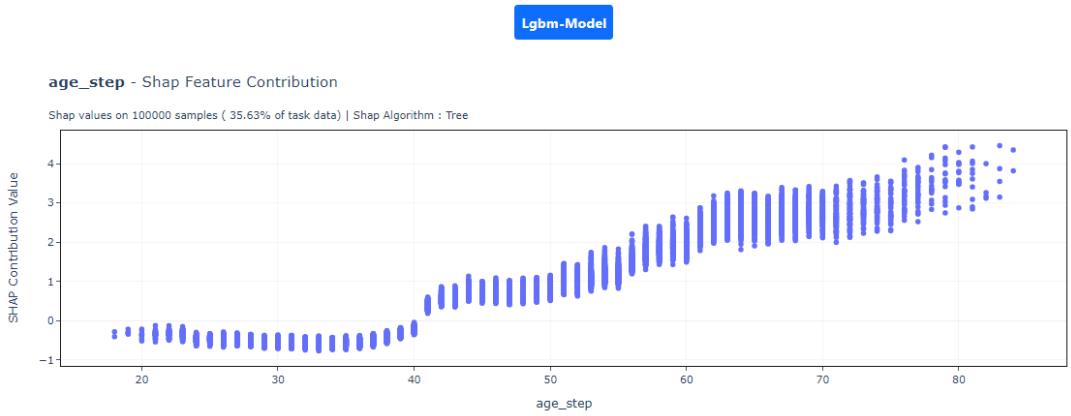


Figure 210- Contribution des formalités médicales en fonction de l'âge dans le modèle en couple sur la mortalité

Lgbm-Model

Feature	Permutation_Importance
filter data...	
seniority	0.00022
dep_katia	0.00016
K_x	0.00006
cd_sexe_x	0.00005
csp_y	0.00005
csp_x	0.00005
age_step_y	0.00005
cd_sexe_y	0.00003
age_step_x	0.00002
top_fumeur_y	0.00002

Get help on get_permutation_importance

Figure 211- Modèle en couple

Lgbm-Model

Feature	Permutation_Importance
filter data...	
age_step	0.00054
csp	0.00048
departement	0.00037
cd_sexe	0.00037
dur_pret	0.00026
mensualite	0.00023
K_ass	0.00014
seniority	0.00009
top_fumeur	0.00004
Dept_Indic	0.00003

Figure 212- Modèle en individuel

Lgbm-Model

Feature	Avg Shap Effect
filter data...	
csp_x	0.22789
dep_katia	0.13077
K_ass_y	0.09108
seniority	0.08806
K_ass_x	0.08058
csp_y	0.06411
cd_sexe_x	0.04940
K_x	0.04910
mensualite_y	0.04784
age_step_x	0.03418

Figure 213- Modèle en couple

Lgbm-Model

Feature	Avg Shap Effect
filter data...	
csp	0.22058
cd_sexe	0.21927
K_ass	0.12989
age_step	0.08452
departement	0.08016
dur_pret	0.05778
seniority	0.05233
mensualite	0.04602
K	0.03188
form_med	0.03172

Figure 214- Modèle individuel

Lgbm-Model

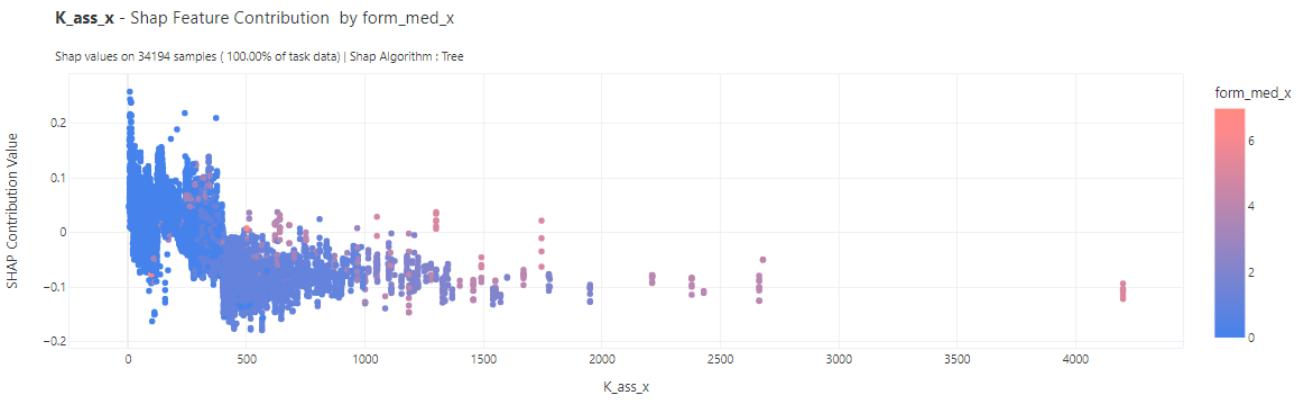


Figure 215- Formalités médicales en fonction du capital assuré pour Mc(x,y)

GLM Model

metrics	On Task data (100%)
filter data...	
Deviance Explained	0.12637
RMSE	0.03838
MAE	0.00075
TotaPred/TotalObs	0.48173
Gini Index (Normalized)	0.72149
AveragePred	0.00024
MSE	0.00147
Average Deviance	0.00730

Figure 216 - Métriques de GLM

GLM Model

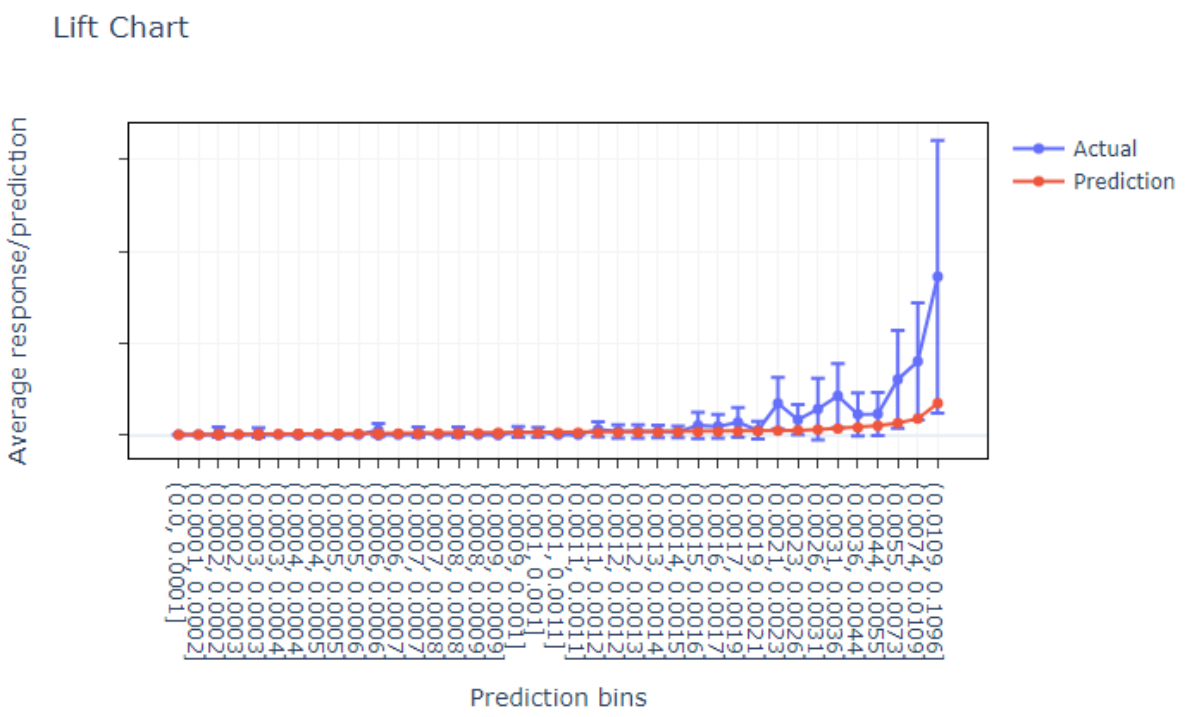


Figure 217 - Graphique des écarts entre les prédictions GLM et les observations

Elastic Net Model

metrics	On Task data (100%)
filter data...	
Deviance Explained	0.13902
RMSE	0.07635
MAE	0.00288
TotaPred/TotalObs	0.51918
Gini Index (Normalized)	0.77658
Average Deviance	0.02326
AveragePred	0.00099
MSE	0.00583

Figure 218 - Métriques de Elastic Net

Elastic Net Model

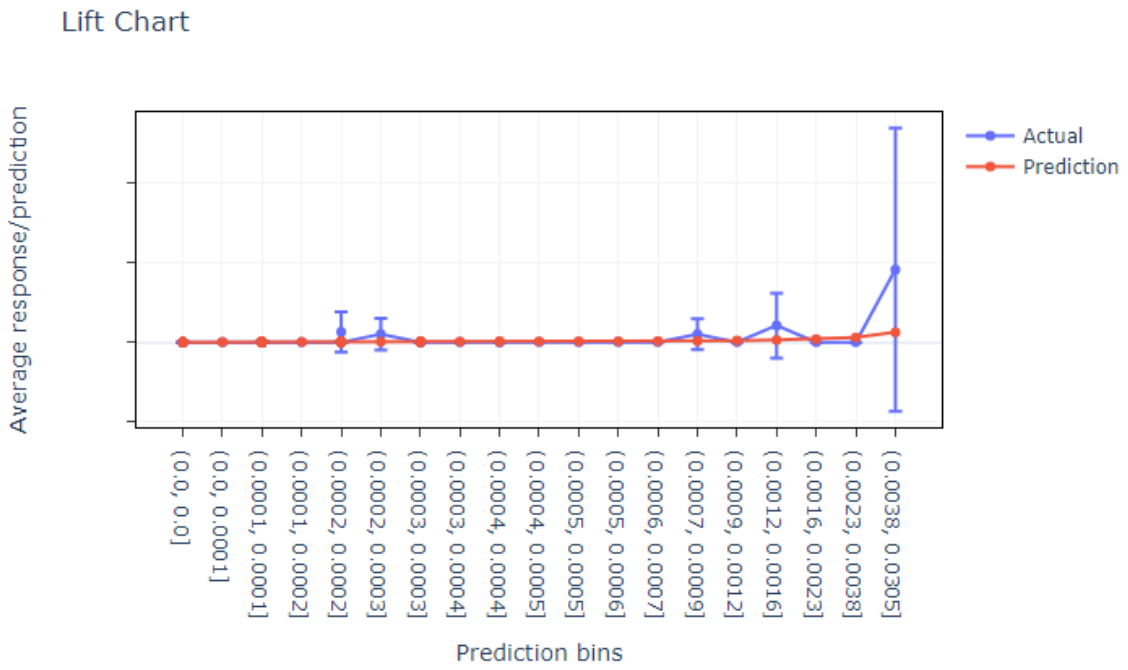


Figure 219 - Graphique des écarts entre les prédictions Elastic Net et les observations