**Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires**

**le**

Par : Matthias GALLON
Titre : Mise à jour des hypothèses de fréquence en risque incapacité au Portugal : l'utilisation d'une modélisation GLM est-elle encore optimale ?

Confidentialité : ☐ Non ☑ Oui (Durée : ☐ 1 an ☑ 2 ans)

---

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

*Membres présents du jury de l'Institut des Actuaires :*

*Entreprise :* **AXA Partners Credit & Lifestyle Protection**
Nom :
Signature :

*Directeur de Mémoire en entreprise :*
Nom : **Four Benoît**
Signature :

*Membres présents du Jury du Certificat d'Actuaire de Paris-Dauphine :*

---

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

*Secrétariat* :

*Signature du responsable entreprise*

*Bibliothèque* :

*Signature du candidat*

**Résumé**

---

Ce mémoire porte sur l'étude de la qualité de la modélisation de la fréquence en risque incapacité via un GLM (Modèle Linéaire Généralisé) des données TTD (Total and Temporary Disability, en français incapacité) du Portugal de différents produits vendus par AXA Partners.

Ces derniers sont présentés ainsi que les spécificités de la couverture TTD, les données et le travail en amont effectué dessus. Puis un GLM de type "régression de Poisson" est appliqué dont le fonctionnement, l'intérêt actuariel, les intérêts internes de la compagnie, et les contraintes qui encadrent sa calibration sont expliqués.

Les résultats obtenus sont comparés avec les données observées ainsi que les données actuellement validées et utilisées par AXA Partners. La robustesse des variables explicatives sélectionnées est analysée via entre autre une validation croisée, ainsi que leur limite.

Finalement une nouvelle modélisation GLM est proposée en se libérant des contraintes du précédent GLM via l'ajout de nouvelles variables explicatives inédites ainsi qu'une modélisation non linéaire de type *machine learning* (apprentissage automatique) : un *boosting*. Les avantages et inconvénients de cette méthode sont également mis en avant afin de conclure sur la meilleure méthodologie à adopter pour modéliser ce portefeuille particulier.

Dans la suite du mémoire, les termes TTD et incapacité sont gardés pour désigner la même chose.

---

*Mots-clés* : *Incapacité*, *GLM*, *Validation croisée*, *Machine learning*, *Boosting*.

**Abstract**

This thesis deals with the study of the quality of the modeling of the frequency of disability risk via a GLM (Generalized Linear Model) of TTD (Total and Temporary Disability) data from Portugal for different products sold by AXA Partners.

These are presented as well as the specificities of the TTD coverage, the data and the upstream work done on them. Then, a GLM of the "Poisson regression" type is applied, whose operation, actuarial interest, internal interests of the company, and the constraints which frame its calibration are explained.

The results obtained are compared with the observed data as well as the data currently validated and used by AXA Partners. The robustness of the selected explanatory variables is analyzed via a cross-validation, as well as their limitations.

Finally, a new GLM modeling is proposed by freeing itself from the constraints of the previous GLM via the addition of new unpublished explanatory variables as well as a non-linear modeling of machine learning type : a boosting. The advantages and disadvantages of this method are also highlighted in order to conclude on the best methodology to adopt for modeling this particular portfolio.

*Mots-clés*: *Disability*, *GLM*, *Cross validation*, *Machine learning*, *Boosting*.

# Note de Synthèse

Le modèle linéaire généralisé (GLM) est couramment utilisé pour sa facilité d'interprétation et son compromis entre biais et variance. En assurance incapacité, AXA Partners l'utilise avec une distribution Poisson pour des calculs de fréquence. Cependant, les hypothèses de fréquences de la couverture incapacité au Portugal sont actuellement limitées (hypothèses datant de 2018). Pour cela, une nouvelle modélisation GLM calibrée différemment est faite, mais des variations de fréquence non prises en compte par celle-ci sont également observées. Cette modélisation est faite sous des contraintes internes (seulement l'âge, le genre et le produit sont des variables explicatives du GLM). C'est ainsi que la question de la pertinence du choix d'un GLM s'est posée pour modéliser ce risque. Est-ce le manque de variables explicatives qui est responsable de la mauvaise prédiction ou bien est-ce le modèle GLM en lui-même qui est limité ? En outre, les bases de données aujourd'hui à disposition sont d'une grande richesse et leur exploitation est rendue possible grâce au développement d'outils de *machine learning*, notamment le *gradient boosting* qui peut être utilisé pour modéliser la fréquence incapacité.

### Cadre de l'étude

La fréquence incapacité doit être modélisée pour différents produits vendus. Il s'agit de produits de protection financière et de protection du niveau de vie : Prêt à la consommation (Personal Loan), Prêt immobilier (Mortgage Loan), Prêt automobile (Car Loan), Carte de crédit (Credit Card), Protection du revenu (Income Protection) et Exonération de prime d'assurance (Waiver of Premium).

Les données sur l'âge, le genre, la date du contrat, la prime d'assurance, etc sont récupérées de bases internes pour chaque produit. Mais seuls le Personal Loan, Mortgage Loan et Car Loan ont des informations sur le genre et l'âge pour chaque assuré. Ce n'est pas le cas du Credit Card, Income Protection et Waiver of Premium, ce qui rend l'étude de leur fréquence via un GLM calibré sur l'âge et le genre impossible. En conséquent ces trois produits sont retirés des modélisations.

Concernant le marché Portugais, celui-ci présente deux spécificités qui vont influencer les modèles. La première est la différence importante dans les fréquences d'accidents entre les hommes et les femmes avec par exemple pour le produit Personal Loan, une fréquence deux fois plus élevée pour les hommes que pour les femmes. Cela s'explique par les professions exercées par les hommes, qui sont plus risquées que celles des femmes. C'est davantage le cas au Portugal où l'économie est encore majoritairement basée sur les secteurs primaire et secondaire. La deuxième spécificité est la crise économique portugaise de 2013, liée à la crise de la dette de la zone euro. Cela a eu pour effet d'augmenter la conscience des Portugais pour les couvertures d'assurance dont ils disposent. Ils ont commencé à déclarer davantage leurs sinistres à partir de 2013, entraînant une hausse globale de la fréquence de l'incapacité tous produits confondus jusqu'en 2019, avant une baisse liée à la crise du covid.

Il est important de vérifier si la hausse de la fréquence observée entre 2013 et 2019 liée à la crise économique au Portugal est prise en compte dans la modélisation GLM sous les contraintes de l'entreprise. Dans le cas échéant, il est intéressant de voir si l'ajout de variables explicatives supplémentaires liées à cette crise permet d'améliorer la modélisation de la hausse de la fréquence dans le GLM. Enfin ces variables sont conservées pour implémenter le modèle de *gradient boosting*.

Concernant le premier modèle GLM sous contraintes, il ne s'agit pas d'un mais de deux GLMs, un pour la maladie et un pour l'accident (les deux couvertures qui composent l'assurance incapacité). Il a en effet été observé suffisamment de différences entre les couvertures :

- pour la maladie, il y a une croissance exponentielle de la fréquence avec l'âge, avec des fréquences très proches entre les hommes et les femmes ;
- pour l'accident, il y a une séparation nette entre les fréquences des hommes et des femmes, l'âge n'influençant pas la fréquence d'accident.

Les variables explicatives retenues pour le GLM maladie sont donc l'âge et le produit, et pour le GLM accident les variables explicatives retenues sont le genre et le produit. Le produit est directement intégré aux GLMs en tant que variable explicative plutôt que de faire un GLM par produit car les tendances observées par âge et genre sont communes pour chaque produit. Pour obtenir la fréquence incapacité globale, il suffit de sommer les résultats des deux GLMs. La variable réponse des GLMs est le nombre de sinistres, et l'exposition est *offsetée* de manière à calculer la fréquence.

Le choix du GLM Poisson est maintenant discuté, les GLMs sont calibrés et les résultats analysés.

**Le premier modèle GLM sous contraintes**

Les données sont séparées en deux jeux de données par couverture, un pour le GLM maladie et un pour le GLM accident. Les éléments à vérifier lors du choix d'une distribution Poisson sont l'indépendance des variables réponses, la distribution Poisson des variables réponses, l'absence de surdispersion et différentes hypothèses sur les résidus, à savoir leur distribution Poisson et si les résidus de déviance sont indépendants, linéaires et normalement distribués. Une fois le modèle calibré, la performance et la qualité de prédiction du modèle sont validées via des tests de déviance et une validation croisée.

Il se trouve que les données ont un grand nombre de valeurs zéros (99.65% de valeurs nulles pour les données maladie et 99.76% de valeurs nulles pour les données accident). Cela induit des moyennes de sinistres très faibles pour les deux jeux de données. Il s'en conclut après des tests d'ajustement du $\chi^2$ que les variables réponses des deux couvertures ne suivent pas une distribution de Poisson. Cependant si la distribution du nombre de sinistres est comparée à la distribution théorique d'une loi de Poisson de paramètre la moyenne du nombre de sinistres, les valeurs obtenues pour zéro et un sinistre (soit 99.99% des données accident et maladie) sont très proches. C'est ainsi qu'il est considéré que le nombre de sinistres dans les deux modèles suit bien une distribution de Poisson.

Une fois les GLM calibrés, il faut vérifier la présence de surdispersion. Via le calcul de la déviance résiduelle et du nombre de degrés de libertés des deux GLMs, il s'obtient un ratio $\phi = \frac{\text{déviance résiduelle}}{\text{nombre de degrés de libertés}}$ inférieur à 1 pour les deux GLMs d'où l'absence de surdispersion. Il est donc inutile de choisir une autre distribution de l'erreur comme une distribution binomiale négative. Cependant, en ce qui concerne les

résidus, différents tests statistiques indiquent que les résidus ne sont pas distribués selon une loi de Poisson, et que les résidus de déviance ne sont ni indépendants, ni linéaire et ni normalement distribués. Il est supposé ici que ces mauvais résultats viennent encore une fois du très grand nombre de valeurs zéros pour les sinistres dans les données. Le modèle GLM avec une distribution Poisson est donc gardé dans la suite de l'étude. Concernant les variables explicatives, le test de déviance anova, la validation croisée et le test de significativité statistique des variables explicatives indiquent qu'elles sont robustes et utiles pour expliquer le nombre de sinistres. Les résultats de la prédiction des deux GLMs sont maintenant comparés à la fréquence observée pour le produit Personal Loan qui représente 65% du portefeuille.

La fréquence prédite (courbe verte) est comparée à la fréquence observée (courbe bleue) et à la fréquence modélisée de 2018 (courbe noire) en fonction de l'année d'incident (figure 1).

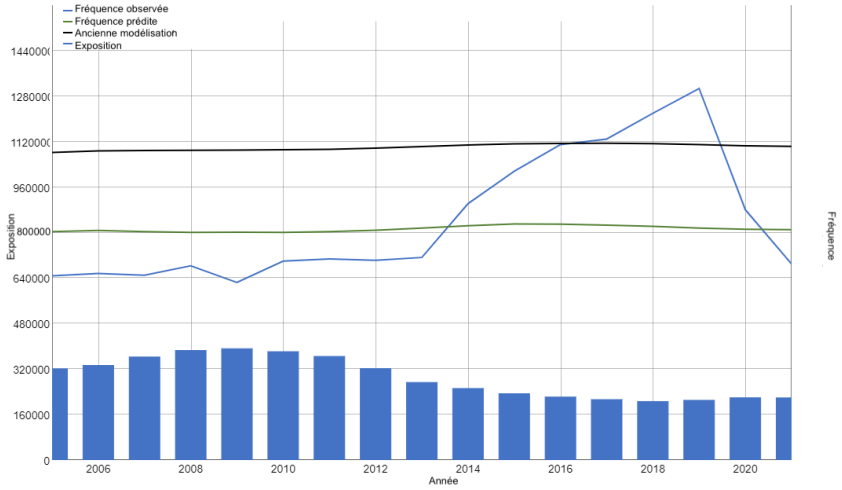



FIGURE 1 : Fréquence observée, fréquence prédite et ancienne fréquence modélisée tous genres et toutes couvertures confondus pour le produit Personal Loan

Il y a une croissance globale de la fréquence avec les années atteignant un pic en 2019. Cette croissance est due à l'augmentation de la conscience que les Portugais ont des couvertures d'assurance dont ils disposent. La baisse après 2019 correspond à la période de covid et n'est pas prise en compte dans la modélisation (les données relatives aux années 2020 et 2021 ont été retirées des modélisations pour ne pas prendre en compte la crise du covid). Le GLM ne modélise pas cette croissance et stagne très en dessous des observations. Ainsi les deux GLMs ne modélisent absolument pas la croissance avec les années de la fréquence pour le produit Personal Loan.

C'est pourquoi il est intéressant de regarder à présent les résultats de la même modélisation GLM, mais cette fois enrichie de variables explicatives en lien avec la situation économique au Portugal.

### Le modèle GLM enrichi de variables explicatives macroéconomiques

Les variables explicatives choisies sont le taux de chômage et l'indice des prix à la consommation (IPC). Le taux de chômage est choisi car il atteint un pic en 2013 lors de la crise économique portugaise, ce qui correspond au début de la conscience des Portugais pour les produits d'assurance qu'ils ont à

disposition. L'IPC est choisi car il mesure l'évolution du niveau moyen des prix des biens et services consommés par les ménages. Il s'agit d'une mesure de l'inflation et augmente avec les années au Portugal. La hausse des prix pousse les Portugais à davantage chercher des moyens de compenser leur perte de pouvoir d'achat, d'où une augmentation des connaissances dans les produits d'assurance auxquels ils sont éligibles. Cela vient compléter l'explication de la crise économique de 2013.

La séparation entre un GLM maladie et un GLM accident est gardée. Comme pour la modélisation GLM sous contraintes, le test de déviance anova, la validation croisée et le test de significativité statistique des variables explicatives indiquent que les variables explicatives sont toutes robustes et utiles pour expliquer le nombre de sinistres. Les résultats de la prédiction des deux nouveaux GLMs sont maintenant comparés à la fréquence observée pour le produit Personal Loan.

La fréquence prédite (courbe verte) est comparée à la fréquence observée (courbe bleue) et à la fréquence modélisée de 2018 (courbe noire) en fonction de l'année d'incident (figure 2).

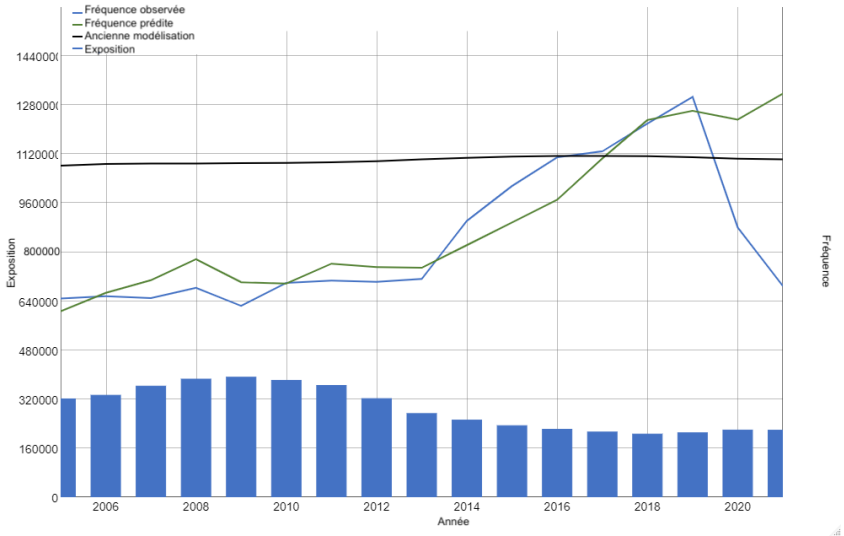



FIGURE 2 : Fréquence observée, fréquence prédite et ancienne fréquence du modèle de 2018 tous genres et toutes couvertures confondus pour le produit Personal Loan

Cette fois, l'ajustement des GLMs s'aligne avec les observations et prédit bien la croissance observée par année d'incident. Cela confirme que l'apport de nouvelles variables explicatives aux GLMs vient enrichir la modélisation et prendre en compte la hausse de la fréquence avec les années d'observations pour le produit Personal Loan.

Il convient maintenant de regarder la modélisation *gradient boosting* de la fréquence incapacité.

### La modèle *gradient boosting*

Le *gradient boosting* est un algorithme d'apprentissage automatique utilisé pour résoudre des problèmes de régression et de classification. Il consiste à combiner de façon adéquate plusieurs modèles de faible complexité pour obtenir un modèle plus performant. Il fonctionne en entraînant des modèles successifs sur les erreurs commises par les modèles précédents pour finalement obtenir un modèle global. Les

modèles de faible complexité sont le plus souvent des arbres de décision.

L'utilisateur doit spécifier trois hyperparamètres qui sont le nombre d'arbres que l'algorithme va construire, la profondeur de chaque arbre et la valeur du taux d'apprentissage. Pour obtenir l'algorithme de *gradient boosting* le plus performant, il faut jouer avec les hyperparamètres et comparer les différents algorithmes avec des critères de déviance et de validation croisée. Dans le cadre de cette étude, l'algorithme gardé est celui ayant 375 arbres de profondeur 4 avec un taux d'apprentissage de 0.01 (noté 375/0.01/4). Les résultats de cet algorithme sont regardés pour le produit Personal Loan.

La fréquence prédite (courbe verte) est comparée à la fréquence observée (courbe bleue) en fonction de l'année d'incident, sans distinction de genre et de couverture (figure 3).

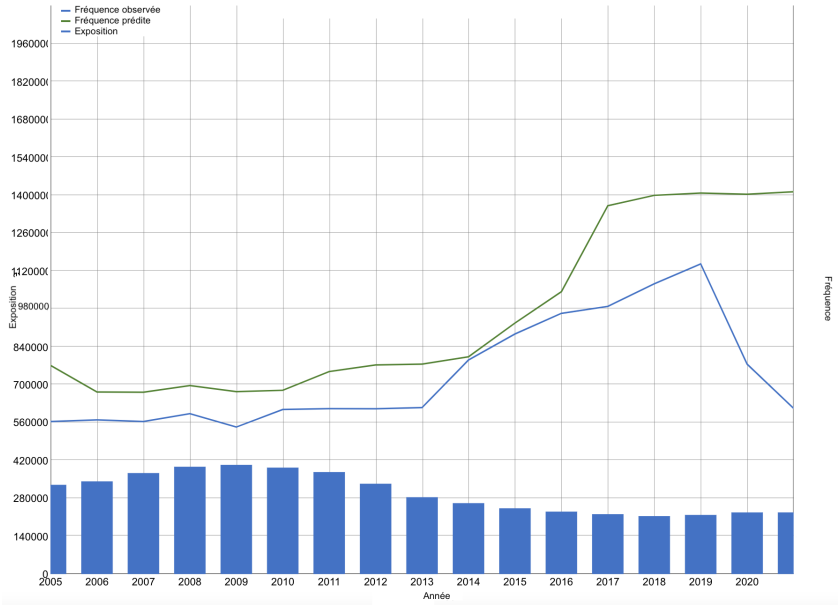



FIGURE 3 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Personal Loan

L'algorithme prédit bien la période stable entre 2005 et 2013 ainsi que la croissance observée à partir de 2013. Cependant, à partir de l'année 2017, il prédit une fréquence plus élevée qui stagne par la suite, là où la fréquence observée continue de croître. La prédiction est moins alignée avec les observations que le GLM augmenté des variables macroéconomiques. Les trois algorithmes de cette étude proposent trois visions différentes du comportement de la fréquence avec les années.

Il convient d'introduire un exemple de calcul de prime pure unique et annuelle pour les comparer d'un point de vue tarification.

### Calcul de la prime pure pour le produit Personal Loan via les trois modèles

Le cadre de l'exemple est le suivant (le nombre de sinistre est confidentiel et change selon les prédictions des trois modèles) (tableau 1).

| Nombre d'assurés | Age moyen | Proportion d'hommes | Montant du prêt | Durée du prêt | Taux d'intérêt annuel | Taux actuariel | Durée d'incapacité |
|---|---|---|---|---|---|---|---|
| 1000 | 43 ans | 75% | 20 000 € | 5 ans 2023-2027 | 5% | 0,75% | 6 mois |

TABLE 1 : Cadre de l'exemple de calcul de prime

Le montant du prêt est de 20000€, à rembourser sur 5 ans avec un taux d'intérêt de 5% soit un remboursement de 374.88€ par mois. Avec une durée moyenne d'incapacité de 6 mois, le coût total $C_i$ que l'assureur doit supporter par année i (i allant de 2023 à 2027) est égal à Nombre de sinistres l'année i $\times$ Remboursement mensuel du prêt $\times$ Durée de l'incapacité. A partir de cela est obtenu le coût par assuré par année i, $c_i = \frac{C_i}{1000}$.

Il s'en déduit la valeur actualisée de ces coûts annuels VA qui correspond à la prime pure unique. Pour obtenir la prime pure annuelle, il faut calculer la valeur actualisée des assurés $VA^{\text{assuré}}$. Par équité actuarielle, il doit y avoir $VA = VA^{\text{assuré}}$. C'est ainsi que la prime pure annuelle est déduite.

Les primes pures obtenues pour les trois modèles sont exprimées ci-dessous (tableau 2).

| | Prime pure unique | Prime pure annuelle |
|---|---|---|
| Modèle 1 : GLM âge/genre/produit | 56,51 € | 11,47 € |
| Modèle 2 : GLM enrichi des variables macroéconomiques | 145,29 € | 29,49 € |
| Modèle 3 : Gradient boosting 375/0.01/4 | 111,81 € | 22,70 € |

TABLE 2 : Récapitulatif des primes pures, unique et annuelle, calculées via les 3 modèles calibrés

La différence importante entre les différentes primes vient du fait que les trois modèles prédisent une évolution distincte de la fréquence. Le premier GLM prédit une fréquence qui stagne bien en-dessous des fréquences observées d'où les primes très faibles obtenues. Le GLM enrichi des variables macroéconomiques prédit une fréquence qui continue de croître dans le futur, là où le *gradient boosting* prédit une fréquence stagnante au dessus des observations à partir de 2017. Le GLM enrichi des variables macroéconomiques est donc plus prudent d'un point de vue tarification, reste à savoir si la fréquence va effectivement se comporter comme il le prédit dans le futur.

Les trois modèles sont maintenant plus largement comparés.

### Comparaison des trois modélisations

Les modèles sont comparés selon des critères de qualité de prédiction, de temps de calibration, d'interprétabilité et de prudence tarifaire.

Le premier GLM sous contraintes modélise très mal la fréquence par année d'incident. En plus d'être largement sous la fréquence observée à partir de 2014, il ne prend absolument pas en compte la hausse de la fréquence observée sur la période 2013-2019. Le GLM augmenté des variables macroéconomiques s'aligne quant à lui complètement sur la fréquence observée, en stagant sur la période 2005-2013 et en augmentant à partir de 2013. Le *gradient boosting* prédit assez bien la fréquence entre 2005 et 2016 avant de se placer largement au-dessus de la fréquence observée. Cependant celui-ci prédit une fréquence stagnante à partir de 2017 là où le GLM augmenté des variables macroéconomiques prédit une croissance continue de la fréquence. Le GLM augmenté des variables macroéconomiques est donc

le modèle qui prédit le mieux la fréquence observée, mais à voir si la fréquence dans le futur suit plutôt sa logique ou celle du *gradient boosting*. En outre, la performance des trois modélisations est analysée via les critères AIC, BIC (pour les deux modèles GLM) et MSE (Erreur quadratique moyenne). Le modèle qui réduit le plus l'AIC, le BIC et le MSE est considéré comme celui qui a une meilleure qualité d'ajustement aux données compte tenu de sa complexité (tableau 3).

| | GLM | | GLM augmenté | | Gradient Boosting | |
|---|---|---|---|---|---|---|
| | Accident | Maladie | Accident | Maladie | Accident | Maladie |
| AIC | 148 999 | 211 450 | 148 636 | 210 497 | | |
| BIC | 149 052 | 211 503 | 148 716 | 210 578 | | |
| MSE | 0,002527 | 0,003729 | 0,002526 | 0,003728 | 0,002543 | 0,003735 |

TABLE 3 : Valeurs obtenues des indicateurs AIC, BIC et MSE pour les trois modélisations

La modélisation GLM enrichie des variables explicatives macroéconomiques est celle qui réduit le plus l'AIC, le BIC et le MSE. C'est donc celle qui est la plus performante selon ces critères.

Le temps de calibration est un élément important à prendre en compte. Alors que le temps de calibration des deux modèles GLM ne prend que quelques minutes, il a fallu des semaines pour implémenter différents *gradients boostings* avant d'obtenir le plus optimal. Le GLM est donc préféré. Concernant l'interprétabilité, le GLM reste très facile à expliquer là où un algorithme de *gradient boosting* demande des connaissances en *machine learning* et en optimisation (descente de gradient). Enfin il a été vu que le GLM augmenté des variables explicatives macroéconomiques est celui qui propose la prime la plus élevée, et donc la prime la plus prudente.

### Conclusion et limites

Le GLM augmenté des variables explicatives macroéconomiques est parmi les trois modèles celui qui présente le meilleur compromis entre interprétabilité, prudence tarifaire, rapidité de calibration et qualité prédictive. Il est donc considéré comme le meilleur de cette étude et cela conclut que le GLM reste adapté pour modéliser la fréquence incapacité, tant que celui-ci est complété de variables explicatives qui prennent en compte les phénomènes observés sur le marché. Cependant le *gradient boosting* présente des résultats encourageants. Avec le développement de bases plus riches en données et d'ordinateurs plus puissants pour les implémenter, les algorithmes de *gradient boosting* risquent dans un futur proche de présenter des résultats plus intéressants que les modèles classiques et peuvent à terme les remplacer.

Il ne faut pas oublier que les conclusions de cette étude sont à prendre avec précaution, notamment du fait de certaines limites observées au cours de son développement. Tout d'abord concernant les variables explicatives, leur nombre est encore très faible, seulement l'âge, le genre et le produit sont des variables explicatives dans la modélisation sous contraintes. Cela pose également problème pour le *gradient boosting* qui n'a que les variables âge, genre, couverture, produit, taux de chômage et IPC alors que son intérêt premier est de gérer un très grand nombre de variables, de trouver seul les liens entre elles et de les trier en fonction de leur importance. Mais pour gérer plus de données, il aurait fallut beaucoup plus de temps de calibration, sans compter le fait de procéder à une *grid-search*, c'est-à-dire tester une multitude d'hyperparamètres pour trouver le modèle le plus optimal. Concernant les GLMs, la très grande quantité de zéros induit de mauvais résultats sur les résidus et la variable à expliquer. Le choix du GLM Poisson dans ce cas peut faire débat.

# Synthesis note

The Generalized Linear Model (GLM) is commonly used for its ease of interpretation and for its compromise between bias and variance. In disability insurance, AXA Partners uses it with a Poisson distribution for frequency calculations. However, the frequency assumptions for disability coverage in Portugal are currently limited (assumptions from 2018). For this, a new GLM modeling calibrated differently is made, but frequency variations not taken into account in this new modeling are also observed. This modeling is done under internal constraints (only age, gender and product are explanatory variables of the GLM). This raises the question of the relevance of the choice of a GLM to model this risk. Is it the lack of explanatory variables that is responsible for the poor prediction or is it the GLM model itself that is limited? Moreover, the databases available today are very rich and their exploitation is made possible thanks to the development of machine learning tools, in particular gradient boosting that can be used to model the frequency of disability.

**Study framework**

The incidence of disability is to be modeled for different products sold. These are financial and lifestyle protection products: Personal Loan, Mortgage Loan, Car Loan, Credit Card, Income Protection and Waiver of Premium.

Data on age, gender, policy date, insurance premium, etc is retrieved from internal databases for each product. Only the Personal Loan, Mortgage Loan and Car Loan products have gender and age information for each insured. This is not the case for the Credit Card, Income Protection and Waiver of Premium products, which makes it impossible to study their frequency via a GLM calibrated on age and gender. Consequently, these three products have been removed from the models.

The Portuguese market presents two specificities that will influence the modeling. The first is the important difference in the frequency of accidents between men and women, with, for example, for the Personal Loan product, a frequency twice as high for men as for women. This is explained by the professions in which men work, which are more risky than those of women. This is more the case in Portugal, where the economy is still mainly based on the primary and secondary sectors. The second specificity is the Portuguese economic crisis of 2013, linked to the Eurozone debt crisis. This had the effect of increasing the awareness of the Portuguese for the insurance coverage they have. They started to report more claims from 2013 onwards, leading to an overall increase in the frequency of disability across all products until 2019, before a decline linked to the covid crisis.

It is important to check whether the observed frequency increase between 2013 and 2019 related to the economic crisis in Portugal is taken into account in the GLM modeling under the firm's constraints.

If so, it is interesting to see if adding additional explanatory variables related to this crisis improves the modeling of the frequency increase in the GLM. Finally, these variables are kept to implement the gradient boosting model.

Concerning the first constrained GLM model, it is not one but two GLMs, one for sickness and one for accident (the two coverages that make up the disability insurance). Indeed, enough differences were observed between the coverages :

- for the sickness, there is an exponential increase in frequency with age, with very similar frequencies between men and women;
- for the accident, there is a clear separation between the frequencies of men and women, with age not influencing the accident frequency.

The explanatory variables retained for the sickness GLM are therefore age and product, and for the accident GLM the explanatory variables retained are gender and product. The product is directly integrated into the GLMs as an explanatory variable rather than making a GLM by product because the trends observed by age and gender are common for each product. To obtain the overall disability frequency, we simply sum the results of the two GLMs. The response variable is the number of claims, and the exposure is offset to calculate the frequency.

The choice of the Poisson GLM is now discussed, GLMs are calibrated and the results are analyzed.

**The first constrained GLM model**

The data are separated into two data sets per coverage, one for the sickness GLM and one for the accident GLM. The elements to check when choosing a Poisson distribution are the independence of the response variables, the Poisson distribution of the response variables, the absence of overdispersion and different assumptions on the residuals, namely their Poisson distribution and whether the deviance residuals are independent, linear and normally distributed. Once the model is calibrated, the performance and predictive quality of the model are validated through deviance tests and cross-validation.

It turns out that the data have a high number of zero values (99.65% of zero values for the sickness data and 99.76% of zero values for the accident data). This leads to very low average claims for both data sets. It can be concluded after $\chi^2$ adjustment tests that the response variables of both coverages do not follow a Poisson distribution. However, if the distribution of the number of claims is compared to the theoretical distribution of a Poisson distribution with as the parameter the average number of claims, the values obtained for zero and one claim (i.e. 99.99% of the accident and sickness data) are very close. Thus, it is considered that the number of claims in both models follows a Poisson distribution.

Once the GLMs are calibrated, we have to check the presence of overdispersion. By calculating the residual deviance and the number of degrees of freedom of the two GLMs, we obtain a ratio $\phi = \frac{\text{residual deviance}}{\text{number of degrees of freedom}}$ lower than 1 for the two GLMs, hence the absence of overdispersion. It is therefore unnecessary to choose another distribution of the error such as a negative binomial distribution. However, with respect to the residuals, various statistical tests indicate that the residuals

are not Poisson distributed, and that the deviance residuals are neither independent, nor linear, nor normally distributed. It is assumed here that these poor results again come from the very large number of zero values for the claims in the data. The GLM model with a Poisson distribution is therefore kept in the rest of the study. With respect to the explanatory variables, the anova deviance test, the cross-validation and the statistical significance test of the explanatory variables indicate that they are robust and useful in explaining the number of claims. The prediction results of the two GLMs are now compared to the observed frequency for the Personal Loan product which represents 65% of the portfolio.

The predicted frequency (green curve) is compared to the observed frequency (blue curve) and the modeled 2018 frequency (black curve) as a function of incident year (figure 4).

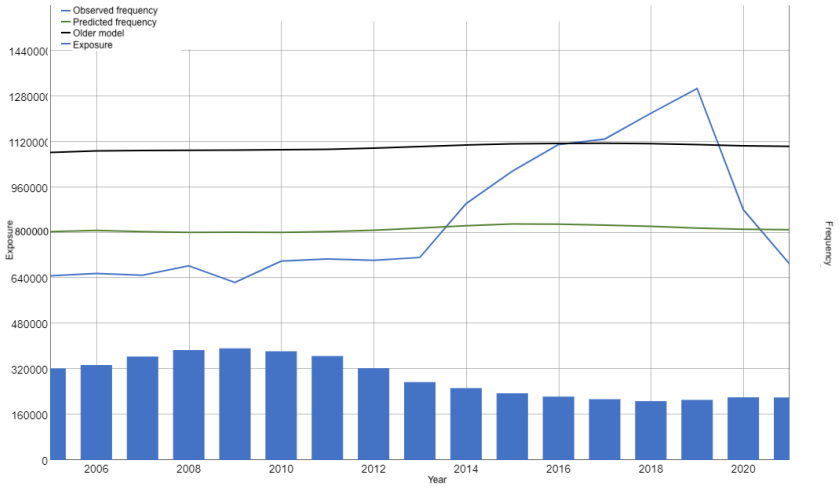



Figure 4: Observed frequency, predicted frequency and former modeled frequency across genders and coverages for the Personal Loan product

There is an overall growth in frequency over the years, reaching a peak in 2019. This growth is due to the increase in the awareness of the Portuguese of the insurance coverage available to them. The decline after 2019 corresponds to the covid period and is not accounted for in the modeling (data for the years 2020 and 2021 have been removed from the models to exclude the covid crisis). The GLM does not model this growth and stagnates far below the observations. Thus, the two GLMs do not model the growth over the years of the frequency for the Personal Loan product at all.

This is why it is interesting to look at the results of the same GLM modeling, but this time enriched with explanatory variables related to the economic situation in Portugal.

### The GLM model enriched with macroeconomic explanatory variables

The explanatory variables chosen are the unemployment rate and the consumer price index (CPI). The unemployment rate is chosen because it peaked in 2013 during the Portuguese economic crisis, which corresponds to the beginning of the Portuguese awareness of the insurance products they have available. The CPI is chosen because it measures the evolution of the average price level of goods and services consumed by households. It is a measure of inflation and increases over the years in

Portugal. Rising prices are causing the Portuguese to look more for ways to compensate for their loss of purchasing power, hence the increase in knowledge of the insurance products for which they are eligible. This complements the explanation of the 2013 economic crisis.

The separation between a sickness GLM and an accident GLM is kept. As with the constrained GLM modeling, the anova deviance test, cross-validation and statistical significance test of the explanatory variables indicate that the explanatory variables are all robust and useful in explaining the number of claims. The prediction results of the two new GLMs are now compared to the observed frequency for the Personal Loan product.

The predicted frequency (green curve) is compared to the observed frequency (blue curve) and the modeled 2018 frequency (black curve) as a function of incident year (figure 5).

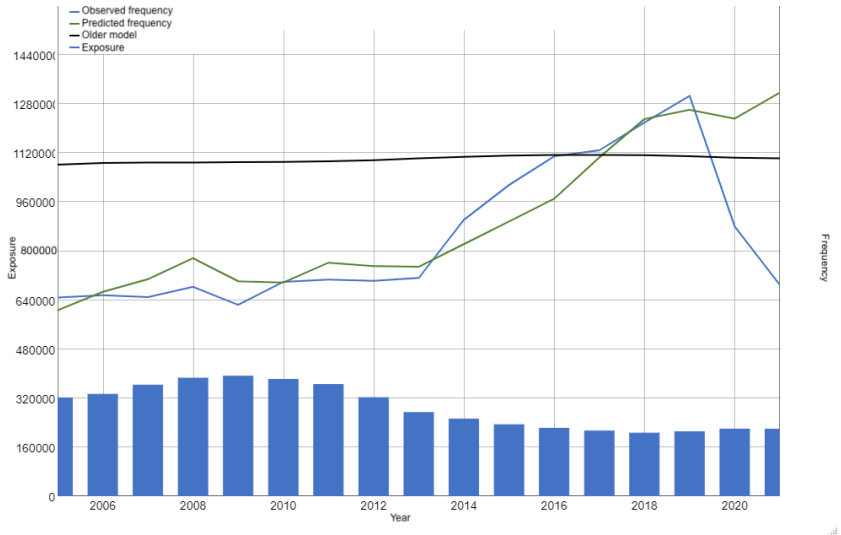



Figure 5: Observed frequency, predicted frequency, and former frequency of the 2018 model across all genders and coverages for the Personal Loan product

This time, the GLMs fit aligns with the observations and predicts well the observed growth by incident year. This confirms that the addition of new explanatory variables to the GLMs enriches the modeling and takes into account the increase in frequency with the years of observations for the Personal Loan product.

Let's have a look now to how the algorithm of gradient boosting models the disability frequency.

### The gradient boosting model

Gradient boosting is a machine learning algorithm used to solve regression and classification problems. It consists in combining several models of low complexity in order to obtain a better performing model. It works by training successive models on the errors made by the previous models to finally obtain a global model. Low complexity models are most often decision trees.

The user must specify three hyperparameters which are the number of trees the algorithm will build,

the depth of each tree and the value of the learning rate. To get the best performing gradient boosting algorithm, we have to play with the hyperparameters and compare the different algorithms with deviance and cross validation criteria. In this study, the algorithm kept is the one with 375 trees of depth 4 with a learning rate of 0.01. The results of this algorithm are looked at for the Personal Loan product.

The predicted frequency (green curve) is compared to the observed frequency (blue curve) as a function of incident year, without distinction of gender and coverage (figure 6).

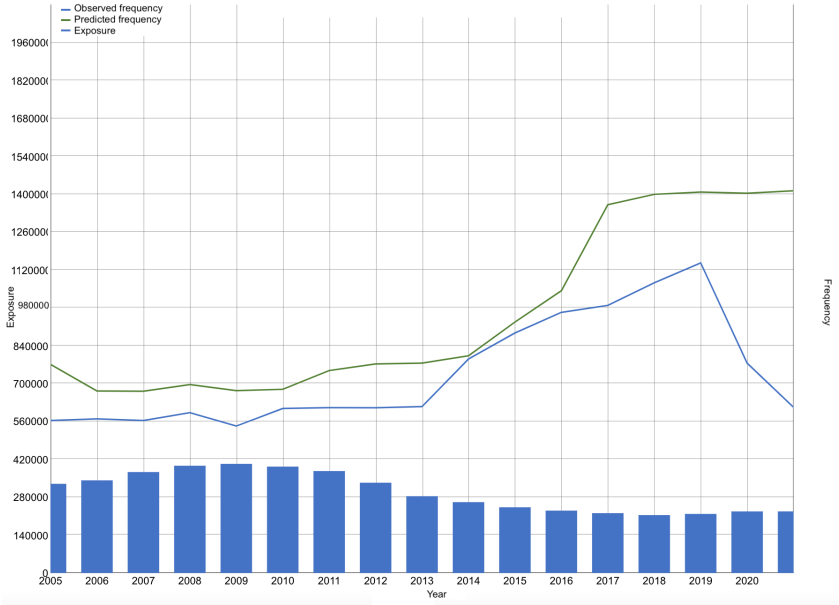



Figure 6: Observed and predicted frequency by year across all genders and coverages for the Personal Loan product

The algorithm predicts well the stable period between 2005 and 2013 as well as the observed growth from 2013 onwards. However, starting in 2017, it predicts a higher frequency that stagnates thereafter, where the observed frequency continues to grow. The prediction is less aligned with observations than the GLM augmented with macroeconomic variables. The three algorithms in this study offer three different views of the behavior of frequency over the years.

An example of a single and annual pure premium calculation should be introduced to compare them from a pricing perspective.

### Calculation of the pure premium for the Personal Loan product via the three models

The framework of the example is as follows (the number of claims is confidential and changes according to the predictions of the three models) (table 4).

The amount of the loan is 20,000€, to be reimbursed over 5 years with an interest rate of 5%, i.e. a reimbursement of 374.88€ per month. With an average duration of incapacity of 6 months, the total cost $C_i$ that the insurer has to bear per year i (i going from 2023 to 2027) is equal to

| Number of insureds | Average age | Proportion of men | Loan amount | Duration of the loan | Annual interest rate | Actuarial rate | Duration of disability |
|---|---|---|---|---|---|---|---|
| 1000 | 43 years old | 75% | 20 000 € | 5 years 2023-2027 | 5% | 0,75% | 6 months |

Table 4: Framework for the premium calculation example

Number of claims in year i × Monthly loan repayment × Duration of disability. From this, the cost per insured per year i, $c_i = \frac{C_i}{1000}$ is obtained.

The present value of these annual costs is VA which corresponds to the single pure premium. To obtain the annual pure premium, we must calculate the present value of the insureds $VA^{\text{insured}}$. By actuarial equity, there must be $VA = VA^{\text{insured}}$. This is how the annual pure premium is deducted.

The pure premiums obtained for the three models are expressed below (table 5).

| | Single pure premium | Annual pure premium |
|---|---|---|
| Model 1: GLM age/gender/product | 56,51 € | 11,47 € |
| Model 2: GLM enriched with macroeconomic variables | 145,29 € | 29,49 € |
| Model 3: Gradient boosting 375/0.01/4 | 111,81 € | 22,70 € |

Table 5: Summary of single and annual pure premiums, calculated via the 3 calibrated models

The important difference between the different premiums comes from the fact that the three models predict a distinct evolution of the frequency. The first GLM predicts a frequency that stagnates well below observed frequencies, hence the very low premiums obtained. The GLM enriched with macroeconomic variables predicts a frequency that continues to grow in the future, while the gradient boosting predicts a stagnant frequency above the observations from 2017 onwards. The GLM enriched with macroeconomic variables is therefore more conservative from a pricing point of view, and it remains to be seen whether the frequency will actually behave as it predicts in the future.

The three models are now more widely compared.

### Comparison of the three models

Models are compared according to criteria of prediction quality, calibration time, interpretability and cautious pricing.

The first constrained GLM models the frequency by incident year very poorly. In addition to being well below the observed frequency from 2014 onwards, it completely fails to account for the increase in frequency observed over the 2013-2019 period. The GLM augmented with macroeconomic variables, on the other hand, aligns completely with the observed frequency, stagnating over the 2005-2013 period and increasing from 2013 onward. The gradient boosting predicts the frequency fairly well between 2005 and 2016 before moving well above the observed frequency. However, it predicts a stagnant frequency from 2017 onwards, whereas the GLM augmented with macroeconomic variables predicts a continuous growth in frequency. The GLM augmented with macroeconomic variables is therefore the model that best predicts the observed frequency, but it remains to be seen whether the frequency in the future follows its logic or that of the gradient boosting. In addition, the performance of the three models is analyzed via the AIC, BIC (for the two GLM models) and MSE (Mean Square Error)

criteria. The model that reduces the most the AIC, BIC and MSE is considered as the one that has a better quality of fit to the data given its complexity (table 6).

| | GLM | | GLM enriched | | Gradient Boosting | |
|---|---|---|---|---|---|---|
| | Accident | Sickness | Accident | Sickness | Accident | Sickness |
| AIC | 148 999 | 211 450 | 148 636 | 210 497 | | |
| BIC | 149 052 | 211 503 | 148 716 | 210 578 | | |
| MSE | 0,002527 | 0,003729 | 0,002526 | 0,003728 | 0,002543 | 0,003735 |

Table 6: Values obtained for the AIC, BIC and MSE criterias for the three models

The GLM model enriched with macroeconomic explanatory variables reduces the AIC, BIC and MSE the most. It is therefore the one that performs best according to these criteria.

An important consideration is the calibration time. While the calibration time of both GLM models takes only a few minutes, it took weeks to implement different gradient boostings before getting the most optimal one. The GLM is therefore preferred. Concerning the interpretability, the GLM remains very easy to explain where a gradient boosting algorithm requires knowledge in machine learning and optimization (gradient descent). Finally, it has been seen that the GLM augmented with macroeconomic explanatory variables is the one that proposes the highest premium, and therefore the most conservative premium.

### Conclusion and limitations

Among the three models, the GLM augmented with macroeconomic explanatory variables is the one that presents the best compromise between interpretability, pricing prudence, calibration speed and predictive quality. It is therefore considered the best in this study and this concludes that the GLM is still suitable for modeling disability frequency, as long as it is supplemented with explanatory variables that take into account the phenomena observed in the market. However, gradient boosting presents encouraging results. With the development of more data-rich databases and more powerful computers to implement them, gradient boosting algorithms may in the near future present more interesting results than classical models and may eventually replace them.

It should be kept in mind that the conclusions of this study should be taken with caution, particularly because of certain limitations observed during its development. First of all, the number of explanatory variables is still very low, only age, gender and product are explanatory variables in the constrained modeling. This also poses a problem for gradient boosting, which only has the variables age, gender, coverage, product, unemployment rate and CPI, whereas its primary interest is to manage a very large number of variables, to find the links between them alone and to sort them according to their importance. But to manage more data, it would have required much more calibration time, not to mention the fact that we would have to proceed to a grid-search, i.e. test a multitude of hyperparameters to find the most optimal model. Concerning GLMs, the very large number of zeros induces bad results on the residuals and the variable to be explained. The choice of the Poisson GLM in this case can be debated.

# Remerciements

Je souhaite remercier toutes les personnes qui m'ont aidé lors de mon stage et de la rédaction de ce mémoire.

Je voudrais dans un premier temps remercier mon tuteur d'entreprise Benoît FOUR, pour ses conseils, sa disponibilité et sa patience qui ont grandement contribué au bon déroulement de mon stage et de la rédaction du mémoire.

Je remercie également ma tutrice pédagogique Anne-Charlotte BONGARD pour ses précieux retours qui m'ont aidé à structurer mon mémoire et adopter la bonne méthodologie de rédaction.

Je tiens à témoigner toute ma reconnaissance aux personnes suivantes, pour leur aide dans la réalisation de ce mémoire :

Monsieur Quentin GUIBERT pour ses retours et ses idées d'ajouts au mémoire.

La cheffe du service actuariat d'AXA Partners Souad MFITIH pour ses conseils et la relecture du mémoire.

Sylvie GUIMARAES ALVES pour son aide au cours du stage et la relecture du mémoire.

Ainsi que toute l'équipe actuariat d'AXA Partners pour son soutien tout au long du stage.

# Table des matières

# Introduction

Le modèle linéaire généralisé (GLM) est l'outil indispensable en tarification IARD (Incendies, Accidents et Risques Divers), son principe étant d'expliquer une variable (appelée variable réponse) à partir de variables explicatives. En effet, dans une volonté de tarification des produits d'assurance non-vie, le calcul de la prime pure via un modèle coût/fréquence s'appuie sur un GLM afin de déterminer les variables explicatives les plus pertinentes et leur impact sur le nombre de sinistres observés. Le GLM est également employé pour sa facilité d'utilisation et d'interprétation par des équipes non techniques et des autorités de contrôle (comme l'ACPR, Autorité de Contrôle Prudentiel et de Résolution). Plus généralement, le GLM présente le meilleur compromis entre biais et variance.

Son utilisation en assurance incapacité est moins courante, mais c'est le cas chez AXA Partners pour ces raisons :

- tous les calculs de fréquence effectués par l'équipe actuariat d'AXA Partners doivent être des GLMs afin de garder une méthodologie de calcul commune au sein de l'entreprise ;
- son interprétation reste très facile et visuelle pour les personnes ne connaissant pas les sciences actuarielles ;
- en assurance incapacité, la variable à analyser est le nombre de sinistres qui permet de déterminer la fréquence d'incident. Il faut donc procéder à un processus de comptage et le GLM de type Poisson reste une approche statistique idéale pour s'en occuper.

Dans un premier temps les hypothèses de fréquences de la couverture TTD (Total and Temporary Disability) au Portugal sont actualisées car des limites dans le modèle actuel (GLM expliquant la fréquence incapacité via les variables explicatives âge et genre par produit) ont été remarquées. Pour cela une nouvelle modélisation GLM calibrée différemment est faite. Des variations de fréquence non prises en compte dans cette nouvelle modélisation sont également observées. Des contraintes internes empêchent l'ajout de variables explicatives supplémentaires à la calibration du GLM pour expliquer ces variations. C'est pourquoi dans un deuxième temps, il est analysé si le fait de se libérer de ces contraintes permet de mieux modéliser la fréquence observée, en ajoutant de nouvelles variables explicatives au GLM, ainsi qu'en procédant à une modélisation non-linéaire de type *machine learning* : un *boosting*.

En effet il se trouve que ces dernières années le monde de l'assurance fait face à un enrichissement des données disponibles, aussi bien en terme de volume que de diversité. Il faut donc que l'actuaire s'arme de nouveaux outils afin de les exploiter et d'en tirer les meilleures interprétations possibles. Bien qu'ils permettent l'utilisation de tests statistiques pour en juger la qualité, les GLMs reposent sur des hypothèses fortes sur la loi de la variable à expliquer ainsi que sur les interactions entre les variables explicatives. Le GLM vient donc limiter les possibilités d'utilisation de la profondeur des données aujourd'hui à disposition.

Le développement de méthodes d'apprentissage statistique et de *machine learning* viennent enrichir l'exploration de ces données et les modélisations qui en découlent. Les méthodes de *machine learning* vont notamment pouvoir capturer les interactions entre les données et proposer des modèles plus affinés pour calculer les fréquences de sinistres.

C'est pourquoi une modélisation de type *boosting* est finalement mise en place, afin d'étudier si le *machine learning* pourrait être une alternative pour évaluer la fréquence du risque incapacité. Pour chaque modélisation sera proposé un exemple de calcul de prime pure pour à terme comparer les trois modèles.

Préalablement, les produits vendus au Portugal et leurs spécificités sont précisés, ainsi que la singularité du marché portugais et les règles relatives aux produits de souscription. Enfin les contraintes qui s'appliquent sur la modélisation sont détaillées. Cela permet de poser le cadre dans lequel la base de données est créée et validée, afin de procéder aux trois modélisations pour répondre à la problématique : le premier GLM sous contraintes, le second GLM libéré des contraintes et augmenté de nouvelles variables explicatives, et enfin l'algorithme de *boosting*.

**Un biais a été ajouté aux données afin de garder confidentiels les résultats obtenus. Cependant, ces changements n'altèrent pas les conclusions de l'étude.**

# Chapitre 1

# L'incapacité, les produits, les règles, le marché et les données

Tout d'abord l'entité "Credit and Lifestyle Protection" (CLP) d'AXA Partners et les catégories de produits qu'elle vend sont présentés, dont ceux couverts en assurance incapacité (en anglais TTD : Total and Temporary Disability). Puis la TTD est détaillée, ainsi que les règles relatives aux produits de souscription. Tout cela permet de bien comprendre les produits et les exigences que l'équipe Risk Management a mis en place pour le compte de la compagnie pour l'ensemble des manipulations et calculs sur les données qui vont être effectués. Enfin les particularités du marché portugais et le premier travail effectué sur les données en amont des modélisations sont précisés. Une fois ce travail effectué, les GLMs et le *boosting* sont calibrés dans les chapitres 2 et 3.

## 1.1 CLP et ses produits

CLP pour "Credit and Lifestyle Protection" est l'une des trois unités principales d'AXA Partners. Cette entité est la spécialiste mondiale d'AXA en matière de produits liés au crédit et à la protection du mode de vie, dont l'objectif est de protéger les assurés et leurs biens tout le long de leur vie.

Pour cela, CLP propose quatre grandes catégories de produits :

- CPI - Credit Protection Insurance, pour protéger les clients et leur famille des conséquences financières d'un événement de la vie sur leurs engagements financiers ;
  - Prêt à la consommation (Personal Loan),
  - Prêt automobile (Car Loan),
  - Carte de crédit (Credit Card),
  - Prêt immobilier (Mortgage Loan),
- Lifestyle Protection, pour aider les clients à faire face à leurs obligations financières et à maintenir leur niveau de vie en cas de perte de revenus ;
  - Protection du revenu (Income Protection),
  - Exonération de prime d'assurance (Waiver of Premium),

- Critical Risks, pour protéger les entreprises et les particuliers en cas d'événements critiques de la vie ;
- Guaranteed Asset Protection, assurance conçue pour compléter l'assurance automobile lorsqu'une voiture est volée ou entièrement endommagée.

**Les produits dont les fréquences de sinistres au Portugal vont être calculées font partie de CPI et de Lifestyle, deux catégories de produits détaillées ci-dessous.** Les catégories Critical Risks et Guaranteed Asset Protection ne sont pas utilisées dans le cadre de cette étude. A partir de la prochaine section, les produits sont appelés par leur nom anglais (par exemple Personal Loan pour le produit Prêt à la consommation).

### 1.1.1 Les produits CPI : Credit Protection Insurance

Les remboursements des cartes de crédit, des contrats de financement, des prêts et des découverts sont désormais monnaie courante. Pour les assurés, un revenu régulier est essentiel pour financer les dettes existantes. Les produits CPI aident à protéger leurs engagements financiers s'ils sont incapables de travailler pour cause de maladie ou de chômage involontaire. Cela réduit les inquiétudes et permet au client de se concentrer sur son rétablissement ou sur sa recherche d'un nouvel emploi.

Voici les produits proposés :

- Personal Loan, couvre les versements mensuels d'un prêt à la consommation en cas de chômage involontaire, d'accident ou de maladie. Il peut également rembourser un prêt en cas de maladie grave ou de décès ;
- Car Loan, couvre les paiements mensuels du prêt automobile en cas de chômage involontaire, d'accident ou de maladie. Il peut également rembourser le prêt en cas de décès, de maladie grave ou d'incapacité ;
- Credit Card, paie le remboursement minimum ou un pourcentage du solde de la carte de crédit à la suite d'un chômage involontaire, d'un accident ou d'une maladie. En outre, en cas de décès, de maladie grave ou d'incapacité, le solde restant est payé ;
- Mortgage Loan, couvre les paiements de prêts immobiliers mensuels en cas de chômage involontaire, d'accident ou de maladie. Il peut également rembourser le solde en cas de décès, de maladie grave ou d'incapacité.

### 1.1.2 Les produits Lifestyle Protection

L'ensemble des produits Lifestyle Protection aide les clients à maintenir leur style de vie en cas d'événements imprévus. Ils aident les clients à faire face à leurs obligations financières et à maintenir leur niveau de vie en cas de perte de revenus.

Les produits proposés :

- Income Protection, l'assuré peut obtenir un montant fixe allant jusqu'à 30% de son salaire pour l'aider à faire face à ses engagements financiers. L'assureur verse à l'assuré une somme forfaitaire ou une prestation mensuelle ;

- Waiver of Premium, couvre la prime des polices d'assurance du client, telles que l'assurance automobile ou l'assurance des frais médicaux majeurs. L'assurance verse un montant forfaitaire ou un versement régulier fixe pour la prime d'assurance, l'épargne ou la cotisation de retraite versée par l'assuré.

Regardons à présent les règles relatives aux produits de souscription afin de mettre en évidence les exigences en terme de Risk Management pour mener à bien l'étude sur les données TTD du Portugal.

## 1.2 Règles relatives aux produits de souscription

Les règles de souscription sont un ensemble de règles et d'exigences qui formalisent le cadre dans lequel se fait le choix des assurés, comment ils sont assurés (sur quelle période, pour combien de sinistres, etc), et impactent directement le remplissage de la base de données. Ces règles servent aussi d'un point de vue interne à protéger l'actuaire lors des modélisations et à les valider avec l'équipe Risk Management.

Les sept règles prises en compte sont la définition de la couverture, l'éligibilité, les prestations, les délais de carence, les franchises, la période de reconstitution de droits et la durée de la couverture.

Elles sont détaillées dans les sections suivantes.

### 1.2.1 Définition des couvertures pour tous les produits

La couverture d'assurance est le type de risque qui est couvert pour un individu ou une entité par le biais de services d'assurance. Les différentes couvertures proposées par AXA Partners sont les suivantes :

- D - Death, décès de l'assuré par n'importe quelle cause ;
- AD - Accidental Death, décès de l'assuré suite à un accident ;
- PTD - Permanent and Total Disability, l'assuré doit être dans l'incapacité totale et permanente d'exercer toute profession ou activité qui pourrait être susceptible de lui procurer un gain financier. L'invalidité doit être reconnue médicalement par l'organisme public compétent en matière de prestations sociales ou suivre la réglementation locale ;
- TTD - Temporary and Total Disability, une blessure corporelle accidentelle ou une maladie qui rend l'assuré totalement et temporairement incapable d'exercer toute occupation susceptible de lui procurer un gain ou un profit. L'incapacité doit être reconnue médicalement ;
- IU - Involuntary Unemployment, chômage involontaire entraînant une interruption totale et continue du travail. Pour être éligible, l'assuré doit être salarié à la date du sinistre et être éligible à une prestation sociale publique ;
- H - Hospitalisation, hospitalisation pour accident ou pour maladie. L'hôpital doit être enregistré/reconnu par un organisme local (sécurité sociale, gouvernement...) ;
- CI - Critical Illness, cancer, crise cardiaque, pontage coronarien, accident vasculaire cérébral, insuffisance rénale (insuffisance rénale terminale), transplantation d'un organe majeur, tumeur

cérébrale bénigne, coma, brûlures au troisième degré couvrant au moins 20% de la surface du corps et nécessitant un débridement chirurgical et/ou une greffe, paralysie des membres, maladie de Parkinson avant 65 ans, sclérose en plaques ;

- LE - Life Events, mariage/union civile (union légalement ou officiellement reconnue de deux personnes en tant que partenaires dans une relation personnelle), naissance ou adoption d'un enfant, diplôme universitaire professionnel (réception ou attribution d'un diplôme académique délivré par un établissement d'enseignement agréé), déménagement professionnel (déménagement permanent à plus de 240 km de l'adresse actuelle).

**La couverture utilisée pour les modélisations est la couverture TTD.**

### 1.2.2 Éligibilités pour tous les produits

Ce sont les conditions que doivent réunir les clients qui veulent se faire assurer pour ces produits.

Le client doit être âgé de 18 à 65 ans (l'âge requis peut varier d'une couverture à l'autre, il faut se référer aux règles spécifiques du produit) lorsqu'il souscrit la police et la couverture cesse lorsqu'il atteint 65 ans (ou l'âge de la retraite dans certains pays). **Dans cette étude au Portugal, l'âge est compris entre 18 et 65 ans, et est une variable explicative importante de la modélisation GLM.**

Pour demander une couverture CLP, il faut être un employé ou travailler activement, avoir été en emploi continu pendant les 6 derniers mois et ne pas avoir de chômage en cours. **C'est le cas des assurés de la base de données de cette étude.**

Les conditions préexistantes sont les informations sur les problèmes de santé passés d'un assuré. Des maladies comme le diabète, le cancer et l'apnée du sommeil peuvent être des exemples de problèmes de santé préexistants. Elles ont tendance à être chroniques ou à long terme. **Pour cette étude, la législation portugaise ne demande aucune condition préexistante**.

Enfin l'assuré doit être un résident permanent dans le pays, **c'est le cas des assurés de la base de données**.

### 1.2.3 Les prestations

Les prestations sont les services couverts par un régime d'assurance.

Il y a le solde impayé remboursé, c'est-à-dire le solde impayé portant intérêt d'un prêt ou d'un portefeuille de prêts, calculé en moyenne sur une période donnée. Les produits concernés sont Death, Accidental Death, Permanent and Total Disability et Critical Illness.

Dans le cas d'une somme forfaitaire, l'assuré recevra un paiement unique effectué à un moment précis. Les produits concernés sont Death, Accidental Death, Permanent and Total Disability, Temporary and Total Disability et Critical Illness.

Il y a également le remboursement mensuel ou d'un montant fixe pendant un nombre déterminé de mois. Les produits concernés sont Temporary and Total Disability, Critical Illness, Involuntary Unemployment et Hospitalisation.

Enfin concernant les petits montants, l'assuré recevra de petits montants fixes afin qu'il puisse profiter

des événements de la vie tels que le mariage, la naissance d'un enfant, l'obtention d'un diplôme universitaire, un déménagement professionnel...

**Dans le cadre de cette étude, les prestations concernées sont les montants forfaitaires et les remboursements mensuels/fixes.**

### 1.2.4 Délai de carence et franchise pour tous les produits

Voici d'abord la vie usuelle d'un contrat d'assurance (figure 1.1).

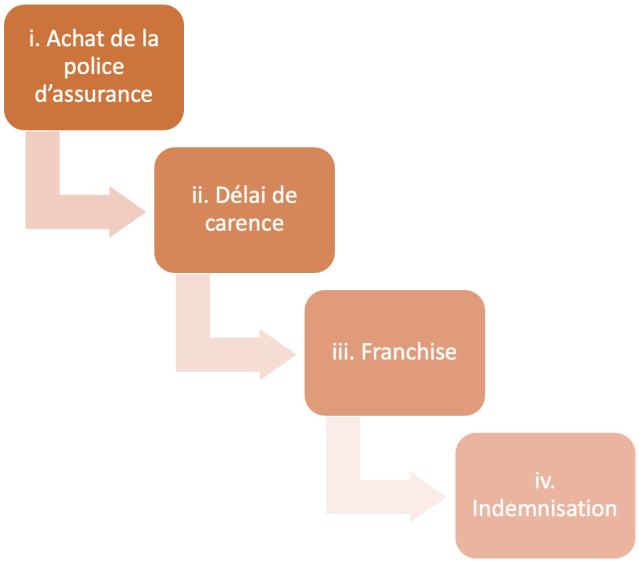



FIGURE 1.1 : Vie usuelle d'un contrat d'assurance

Le délai de carence est la période pendant laquelle la police d'assurance doit être en vigueur avant qu'une couverture soit activée. C'est la période au début d'une police. Tout sinistre potentiel survenu pendant cette période ne peut faire l'objet d'une demande de remboursement. Le délai de carence est normalement différent pour les différentes garanties d'une même police (par exemple, vie vs accident et maladie). **Il faut donc exclure tous les sinistres ayant lieu pendant cette période dans les données.** Le délai de carence doit compenser le nombre de réclamations pendant la période et réduire l'anti-sélection.

La franchise est la période de temps qui doit s'écouler avant qu'une demande puisse être payée. C'est la période que l'assuré doit attendre, à partir du moment où sa demande est introduite jusqu'à ce qu'il puisse recevoir une prestation.

Par exemple, voici un délai de carence initial de 90 jours pour n'importe quel produit au début de la police (figure 1.2).

Et voici une franchise de 30 jours consécutifs pour n'importe quel produit (figure 1.3).

**A chaque produit vendu est associée une franchise sous-jacente, qui impacte le calcul de la fréquence.** En effet prenons le cas d'un individu qui est malade pendant 40 jours (le délai de carence est supposé passé). Il est considéré deux franchises différentes.

Une franchise de 30EM (30 *Excess Monthly*). Cela veut dire que l'assuré ne touche pas d'indemnisation

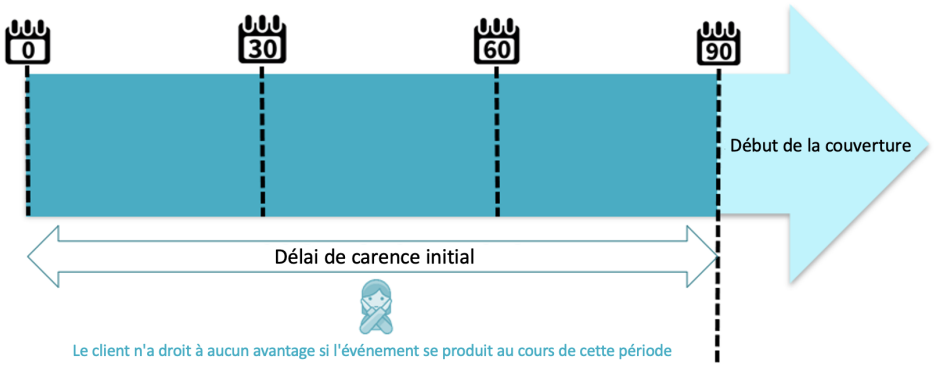



FIGURE 1.2 : Exemple de délai de carence

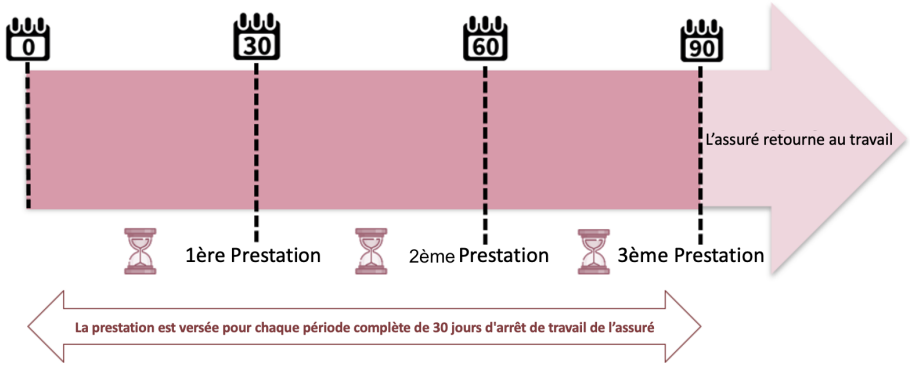



FIGURE 1.3 : Exemple de franchise

pour le premier mois de maladie (principe du 30 *Excess*), et comme il n'est pas malade durant tout le mois suivant (seulement 10 jours), alors il n'est pas indemnisé non plus (principe du *Monthly*). Dans la base de données cela est caractérisé par 0 sinistre (figure 1.4).

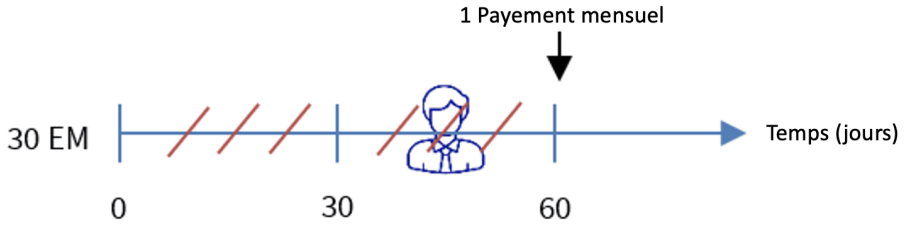



FIGURE 1.4 : Franchise de 30 EM, l'individu ne peut toucher un payement mensuel que s'il est encore malade après le deuxième mois, ici ce n'est pas le cas donc il ne touche rien

Une franchise de 30RM (30 *Retro Monthly*). L'assuré peut toucher l'indemnisation pour le premier mois de maladie (30 *Retro*), mais comme il n'est pas malade durant tout le mois suivant (seulement 10 jours), alors il n'est pas indemnisé pour ces 10 jours (principe du *Monthly*). Dans la base de données

cela est caractérisé par 1 sinistre (figure 1.5).

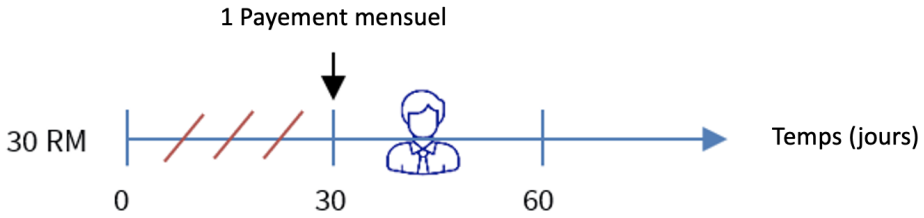



FIGURE 1.5 : Franchise de 30 RM, l'individu touche un payement mensuel car il a été malade pendant plus d'un mois

Il y a donc pour le même scénario, deux valeurs de sinistres différentes selon la franchise appliquée pour le produit. **Les données TTD du Portugal utilisées dans cette étude suivent le même procédé. A chaque produit est associée une franchise sous-jacente qui influence le calcul de la fréquence.**

Tous les produits dans cette étude ont la même franchise de 30RD (30 *Retro Daily*, à savoir que l'assuré peut toucher l'indemnisation pour chaque jour de sinistralité dès le premier mois de déclaration. En reprenant l'exemple de 40 jours de maladie, 1 sinistre est compté).

### 1.2.5 La période de reconstitution des droits

La période de reconstitution des droits est la période pendant laquelle un sinistre est survenu et le paiement correspondant a été effectué, avant qu'un nouveau paiement puisse être effectué à l'occasion d'un nouveau sinistre. En voici un exemple.

L'assuré a une police d'assurance d'une durée de 18 mois, pour différentes couvertures : Involuntary Unemployment, Temporary and Total Disability, Hospitalisation, Critical Illness, Permanent Total Disability et Life Events.

L'assuré ouvre une première demande d'indemnisation pour Involuntary Unemployment (IU) après 6 mois à compter du début de la police et il reçoit l'indemnité pour les 6 mois suivants. Une nouvelle demande peut donc être présentée après 180 jours, à partir du 13e mois de la police. Si le nombre maximal de paiements n'a pas atteint le total, l'assuré peut présenter une nouvelle demande de chômage involontaire après 180 jours ou d'incapacité temporaire après 180 jours ou 30 jours selon le cas (figure 1.6).

Les périodes de reconstitution usuelles :

- TTD et H ;
  - 180 jours pour la même condition,
  - 30 jours pour une condition différente,
- IU ;
  - 180 jours,
- LE, le client ne peut réclamer qu'une seule indemnisation par an.

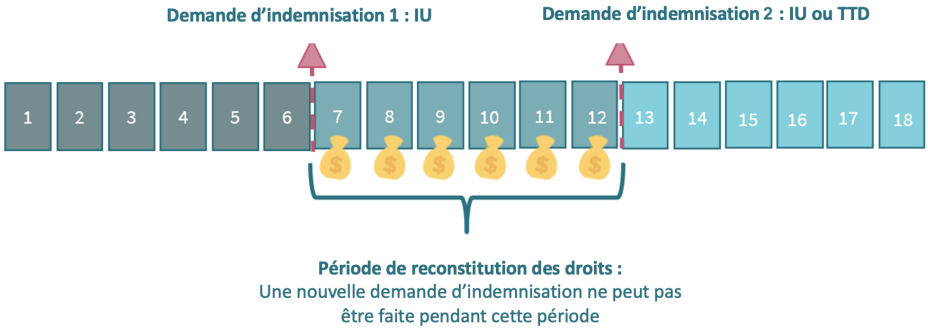



FIGURE 1.6 : Exemple de période de reconstitution des droits

**La base de données utilisées dans cette étude a été remplie en prenant bien en compte la période de reconstitution des droits.**

### 1.2.6 La durée de couverture

La durée de couverture offre une couverture à un taux de paiement fixe pendant une période limitée. Les polices ont une période d'assurance définie, qui est la période pendant laquelle la police est effective. La date de début et la date de fin sont les dates limites pour la documentation, les paiements et la couverture de l'assuré, sauf s'il renouvelle la police. La durée maximale est égale à la durée du prêt, avec un maximum de mois ou d'années. Une police d'assurance sera annulée lorsque le titulaire de la police meurt, part à la retraite ou atteint 65 ans (pour l'IU et la TTD), manque un paiement de prime mensuel, atteint la limite d'indemnisation, annule la couverture.

**La durée de couverture moyenne permet de calculer pour chaque produit le nombre moyen de mois où il faut faire des versements à l'assuré. Quand il s'agit de payer le prêt d'un assuré, le montant total moyen qui doit être versé peut être déterminé à partir de la durée moyenne et du niveau de remboursement du prêt auquel s'est engagé l'assuré.**

Maintenant que les règles de souscription ont été introduites pour permettre de faire la base d'étude et encadrer la modélisation, la TTD est explicitée.

## 1.3 L'assurance TTD - Incapacité

### 1.3.1 Définition

TTD est le sigle de Temporary and Total Disability, donc Incapacité Temporaire et Totale. Cela comprend les blessures corporelles accidentelles et les maladies qui empêchent temporairement et totalement les assurés d'exercer toute activité professionnelle susceptible de leur procurer un gain ou un profit. Il faut que l'incapacité soit médicalement reconnue.

Ce produit d'assurance est conçu pour remplacer le revenu des personnes qui ne peuvent plus travailler. En parallèle, l'assureur octroie une exonération des primes pendant les périodes d'incapacité. Généralement, les prestations sont liées au salaire de l'assuré, mais sont le plus souvent plafonnées à 50-70% pour l'encourager à retourner travailler.

Les prestations fondées sur une incapacité totale exigent que le titulaire de la police d'assurance soit incapable de travailler à son poste habituel. Si celui-ci est en mesure de travailler en partie, il n'aura droit qu'à une indemnité plus faible basée sur l'incapacité partielle. La durée de versement des prestations est choisie par le titulaire de la police d'assurance en début de contrat.

**Les données TTD utilisées dans cette étude sont de deux types, des données accident et des données maladie.** Regardons à présent les particularités du contrat pour l'assurance TTD.

### 1.3.2 Le contrat d'assurance TTD - Incapacité

La définition de l'incapacité totale dans le contrat peut être basée sur l'impossibilité de l'assuré à exercer son propre emploi ou tout autre emploi raisonnable.

L'assurance incapacité est souvent vendue en tant qu'assurance collective et payée par l'employeur pour compenser le coût des prestations d'incapacité de longue durée. L'assurance d'incapacité collective est généralement moins chère car elle présente moins de risque d'anti-sélection, une économie d'échelle et moins de risque de non-paiement des primes.

Il y a anti-sélection lorsque l'assuré dispose d'une information cachée de son propre risque que n'a pas l'assureur. L'assureur ne peut donc pas proposer des primes différentes selon les risques mais une prime moyenne. Les assurés à haut risque auront donc tout intérêt à souscrire à cette assurance à l'inverse des assurés à bas risque qui vont préférer aller voir ailleurs. L'assureur se retrouve à n'assurer que les hauts risques qui considèrent la prime comme inférieure au coût moyen de leurs sinistres. Cela conduit à une perte pour l'assureur. **C'est là l'intérêt de mettre en place un délai de carence qui empêche que des personnes se savant déjà malades souscrivent à une assurance TTD.**

Le produit Personal Loan a un délai de carence de 30 jours, les produits Mortgage Loan, Car Loan et Credit Card un délai de 60 jours et les produits Waiver of Premium et Income Protection un délai de 90 jours.

L'économie d'échelle est la baisse du coût unitaire d'un produit obtenue en accroissant la quantité achetée. Voici l'exemple d'un individu souscrivant à une assurance incapacité.

### 1.3.3 Exemple d'assurance TTD - Incapacité

Il est supposé un contrat d'assurance qui couvre une durée d'indemnisation de deux ans, avec des paiements mensuels de prime, d'une franchise de 1 mois (30EM) et d'une période de reconstitution des droits de 6 mois. Les prestations sont versées le premier jour de chaque mois admissible.

Il est supposé que l'assuré tombe malade le 01/01/N et le reste jusqu'au 30/06/N, puis retourne au travail et retombe malade le 01/09/N. S'il reste malade indéfiniment, quelle sera la date de la dernière prestation versée ?

En raison de la franchise de 30EM, l'assuré touche son premier versement le 01/03/N. Puis touche 4 prestations avant de se rétablir. Il tombe à nouveau malade 2 mois après ce qui est inférieur à la période de reconstitution des droits de 6 mois, par conséquent cette nouvelle période est considérée comme faisant partie de la durée initiale d'indemnisation de 2 ans. Il lui reste $12 \times 2 - 4 = 20$ mois de paiements restants à partir du 01/09/N. Le dernier versement correspond donc à la date du 01/04/N+2.

Maintenant que la TTD a été définie, les spécificités du Portugal sont détaillées. Celles-ci vont influencer les modélisations ultérieures.

## 1.4 Les spécificités du Portugal

Le Portugal présente deux spécificités.

La première concerne la différence importante dans les fréquences d'accident entre les hommes et les femmes. Pour un produit comme le Personal Loan, il y a une fréquence de plus du double pour les hommes par rapport aux femmes (0.18% vs 0.38%) tous âges confondus. Cette différence est moins marquée dans les pays voisins comme l'Espagne, avec une fréquence pour le Personal Loan de 0.18% pour les femmes contre 0.30% pour les hommes. Elle s'explique par les professions exercées selon le genre. Les hommes vont davantage avoir des métiers à risques qui impliquent plus d'accident (exemple : marin-pêcheur, élagueur, agriculteur, ouvrier sidérurgiste, éboueur, chauffeur routier, ouvrier du bâtiment, militaire, manutentionnaire, etc). Au Portugal c'est davantage le cas, puisque l'économie est encore majoritairement basée sur les secteurs primaire et secondaire. Le secteur tertiaire représente 52% de la population active contre 76% en Espagne.

La deuxième spécificité est la crise économique portugaise de 2013 qui est liée à la crise de la dette de la zone euro. Avec la Grèce, le Portugal est le pays le plus touché et voit sa courbe du chômage atteindre un pic en 2013 (figure 1.7). Cette crise a eu pour effet d'augmenter la conscience des Portugais des couvertures d'assurance dont ils disposent. Ces derniers ont donc commencé à partir de 2013 à davantage déclarer leurs sinistres et **s'en est suivie une hausse globale de la fréquence TTD tous produits confondus jusqu'en 2019 avant une chute liée à la crise du covid (figure 1.8 de la section 1.5.3 plus loin).**

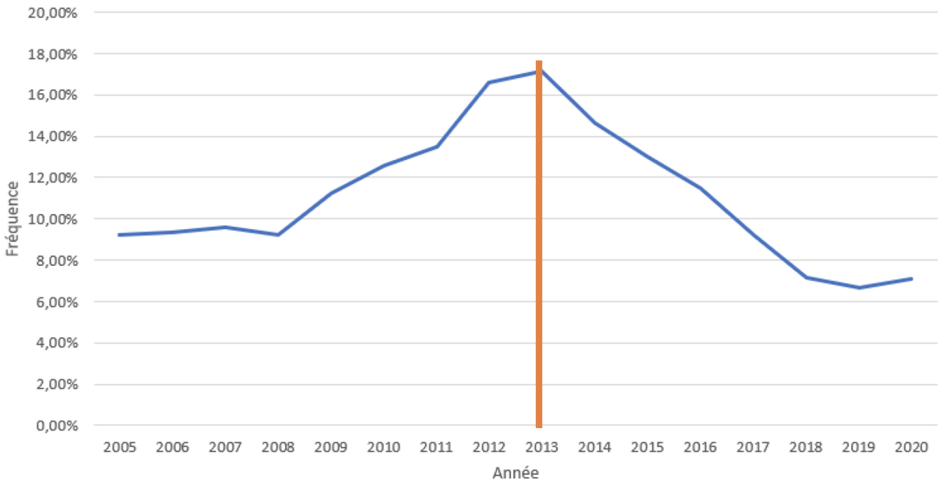



FIGURE 1.7 : Courbe du chômage au Portugal au cours des dernières années, la barre verticale indique la crise de 2013

Il se trouve que cette hausse de la fréquence observée entre l'année 2013 et l'année 2019 liée à la crise économique au Portugal est inédite et n'a pas été prise en compte dans la précédente modélisation GLM. Il est intéressant de voir si la nouvelle modélisation GLM prend en compte cette hausse, malgré l'absence d'ajout de nouvelles variables explicatives.

Le cas échéant, il sera nécessaire en interne de mettre en place une méthodologie pour prendre en compte cette hausse dans la première modélisation GLM pour ne pas proposer une fréquence trop faible. C'est pourquoi dans le chapitre 3, il est regardé si l'ajout de variables explicatives

supplémentaires en lien avec cette crise (des variables explicatives macroéconomiques comme le taux de chômage) permet d'améliorer la modélisation de la hausse de la fréquence dans le GLM. Ces variables seront gardées pour implémenter le modèle de *boosting*.

Mais avant de procéder à ces modélisations, il est nécessaire de produire une base de données pertinente. Pour cela il convient de regarder toutes les données qui sont à disposition. Après les avoir récupérées, nettoyées et réunies, il faut en vérifier la qualité via la phase dite de réconciliation.

## 1.5 Préparation des données et vérification de leur qualité

Ce travail est effectué sur le logiciel SAS (1985), de la récupération des bases initiales à la production de la base utile aux modélisations ultérieures. **Les données utilisées sont confidentielles et un biais a été ajouté.**

### 1.5.1 Les bases de données initiales

Trois bases internes disponibles sont utilisées.

La base *Policies* recense tout l'historique des contrats d'assurance, leur date de début et de fin, le genre de l'assuré, son âge, le montant de la prime versée et d'autres informations techniques. Elle se compose elle-même de quatre bases, MF, MR, Upfront et Bulk :

- la base MF (*Monthly Fixed*) a une vision des primes payées tous les mois avec une date de fin fixée au début de la souscription du contrat ;
- la base MR (*Monthly Renewable*) a une vision des primes payées tous les mois et où le contrat est renouvelé tous les x mois ou années ;
- la base Upfront a une vision où le client paye entièrement sa prime en début de contrat ;
- la base Bulk a une vision agrégée, c'est-à dire qu'une ligne dans la base correspond à plusieurs clients. Pour chaque mois elle donne le nombre de client qui ont payé leur prime.

**Les données de la base Bulk ne sont donc pas exploitables dans la modélisation GLM puisque la vision agrégée limite les informations sur le genre et l'âge des assurés, deux variables explicatives essentielles au GLM de cette étude. Il faut donc programmer le GLM sans ces données, et pour les produits ne présentant que des données Bulk, ils seront inclus dans l'équation du GLM via une méthode différente**.

La base *Claims* recense tout l'historique des réclamations d'indemnisation de sinistres, la date de l'incident, le contrat d'assurance correspondant, les dates de début et de fin de couverture, le type de couverture, la *scheme* (c'est un code qui permet d'unifier un contrat particulier avec un partenaire spécifique, un partenaire étant un client professionnel à qui AXA Partners propose des produits d'assurance à vendre à ses clients), les versements effectués et d'autres informations techniques.

La base *Schemes* recense toutes les informations sur les partenaires avec qui AXA Partners travaille, les *schemes* qui leur sont associées, le produit et la couverture qui correspondent et d'autres informations techniques.

De ces trois bases vont être récupérées les variables les plus intéressantes, qui vont être nettoyées et fusionnées pour obtenir une base de données complète et exploitable pour les modélisations.

### 1.5.2 Travail sur les bases de données

#### Définition du cadre d'étude

La période d'extraction des données est du 01/01/2005 au 31/12/2021 afin d'avoir un large spectre d'informations et étudier l'évolution des fréquences au Portugal au cours des années. Lors de la calibration des trois modèles, les années 2020 et 2021 sont retirées afin de ne pas prendre en compte les changements potentiels de la fréquence liés aux années covid. Il est supposé que ce sont des années exceptionnelles et qu'elles n'influencent pas les fréquences des années futures.

Les couvertures retenues sont maladie et accident (les deux couvertures en TTD), et les 4 types de bases *Policies* à savoir Upfront, MR, MF et Bulk sont gardées.

Les produits dont la fréquence est modélisée sont les produits vendus en TTD définis précédemment (1.1), à savoir Personal Loan, Mortgage Loan, Car Loan, Credit Card, Waiver of Premium et Income Protection.

Avant de regarder les manipulations de la création des bases de données, introduisons la notion d'exposition qui est essentielle au calcul des fréquences d'incident et qui est ajoutée à la base de données finale.

#### L'exposition

L'exposition correspond à la susceptibilité d'un individu de rencontrer un sinistre. Le calcul de l'exposition se fait sur une base annuelle. Un assuré couvert du 01/01/N au 31/12/N présente une exposition de 1. En revanche un assuré couvert du 01/01/N au 30/06/N présente une exposition de 0.5.

En rassemblant l'ensemble des sinistres et l'ensemble des expositions, il est possible de calculer la fréquence d'incident via le calcul $Frequence = \frac{Somme\ du\ nombre\ de\ sinistres}{Somme\ de\ l'exposition}$.

Avec toutes ces informations, il est possible de procéder au choix des variables et la création de la base de données.

#### Les variables clés retenues et les choix de création de base

Tout d'abord, il faut importer les bases *Policies*. Seules les variables les plus utiles sont gardées (tableau 1.1).

Puis pour chaque base *Policies*, selon sa nature, il faut procéder à différents nettoyages et suppression des erreurs. Par exemple sur la base Upfront, il faut supprimer toutes les données avec des erreurs de date, notamment les données où la date de début de contrat est après la date de fin.

Une fois les 4 bases *Policies* nettoyées, elles sont formatées pour pouvoir être fusionnées et n'avoir qu'une seule base *Policies*. Puis il est nécessaire de s'occuper de la base *Claims*. Là aussi seule une liste de variables utiles est gardée (tableau 1.2).

Il faut également procéder à divers nettoyages pour supprimer les erreurs et rendre la base exploitable.

La même logique est répétée pour la base *Schemes*, où ne sont gardées que les variables utiles (tableau 1.3).

| Code Variable | Description |
|---|---|
| country_CD | le code du pays, ici PT pour Portugal |
| policy_number | le numéro de contrat |
| product | le nom du produit |
| scheme | code qui permet d'unifier un contrat particulier avec un partenaire spécifique |
| cover_code | le code de couverture : DA/IA pour accident, DS/IS pour maladie |
| start_date | date de début du contrat |
| end_date | date de fin du contrat |
| cancel_date | si le contrat est annulé avant sa date de fin |
| insured_gender | homme ou femme |
| insured_dob | date de naissance de l'assuré |
| premium | prime d'assurance |

TABLE 1.1 : Variables utiles des bases *Policies*

| Code Variable | Description |
|---|---|
| country | le nom du pays |
| policy_number | le numéro de contrat, permet de faire le lien avec la base Policies |
| cover_start_date | date de début de contrat |
| cover_end_date | date de fin de contrat |
| cover | le code de couverture : DA/IA pour accident, DS/IS pour maladie |
| product | le nom du produit |
| scheme | code qui permet d'unifier un contrat particulier avec un partenaire spécifique |
| incident_date | la date de l'incident |
| notification_date | la date de notification de l'incident |
| first_open_date | date d'ouverture de la gestion du sinistre |
| first_close_date | date de fermeture de la gestion du sinistre |
| status | le statut de la demande d'indemnisation : ouvert/fermé |
| total_payments | le total de payements versés |
| gender | le genre de l'assuré |
| birth_date | la date de naissance de l'assuré |
| occupation | le métier de l'assuré |
| max_no_of_payments | nombre de payements maximum qu'un assuré peut toucher lors de son indemnisation |

TABLE 1.2 : Variables utiles de la base *Claims*

| Code Variable | Description |
|---|---|
| country | le nom du pays |
| scheme | code qui permet d'unifier un contrat particulier avec un partenaire spécifique |
| cover_code | le code de couverture : DA/IA pour accident, DS/IS pour maladie |
| Agent_Name | nom du partenaire associé à la Scheme |
| Agent_Id | l'identifiant du partenaire |
| product_start_date | date de début du produit |
| product_end_date | date de fin du produit |
| max_benefit_month | nombre de mois maximum durant lesquels l'assuré peut toucher son indemnisation |

TABLE 1.3 : Variables utiles de la base *Schemes*

Après avoir été rendue exploitable, cette base est fusionnée avec la base *Policies*. Pour cela les clés *country*, *scheme* et *cover_code* sont utilisées, de manière à unifier la fusion. Les deux bases auront en amont été triées dans l'ordre ci-dessus. Il faut ajouter à cela les franchises sous-jacentes venant d'une autre base, qui sont liées via la *scheme* et le *cover_code*.

Enfin, cette nouvelle base est fusionnée avec la base *Claims* via les clés *country*, *scheme*, *cover_code*, *policy_number*, *age* (calculé à partir de la date de naissance des assurés), *maturity* (nombre correspondant

au combientième mois de payement de l'indemnisation aux assurés), *insured_gender* et *insured_dob*.

Une fois le nettoyage des 3 bases et leur réunion effectués, il est possible de passer à la récupération des expositions et de leur nettoyage. Les expositions des *schemes* pour lesquelles aucune demande d'indemnisation n'a encore été enregistrée sont retirées. Les données d'exposition sont ensuite ajoutées à la précédente base créée.

Finalement, la base *Aggregate* est créée (tableau 1.4).

| Code Variable | Description |
|---|---|
| date | date de l'incident |
| year | année de l'incident |
| qtr | trimestre de l'incident |
| cover_code | le code de couverture : DA/IA pour accident, DS/IS pour maladie |
| insured_gender | homme ou femme |
| age | age de l'assuré |
| type | type de données Policies : Upfront/Bulk/MF/MR |
| sub_product | le nom du produit : Personal Loan/Mortgage Loan/Car Loan/Credit Card/Waiver of Premium/Income Protection |
| agent_name | nom du partenaire |
| WP_ID | la franchise : information sur la franchise, le délai de carence et le nombre de mois de payement maximum |
| maturity | la maturité |
| scheme | code qui permet d'unifier un contrat particulier avec un partenaire spécifique |
| nber_of_claims | le nombre de sinistres |
| revised_exposure | l'exposition |

TABLE 1.4 : Variables de la base finale *Aggregate*

C'est cette base qui va servir à la modélisation via les packages `stats` et `gbm` du logiciel R CORE TEAM (2022). Mais pour s'assurer de la qualité de la base dans son ensemble avant toute modélisation, il faut procéder à la phase dite de réconciliation.

### 1.5.3 La réconciliation

La réconciliation est essentielle pour confirmer l'exploitabilité des données. Son principe est le suivant. Une fois que les données ont été préparées après différentes étapes, il faut les comparer à une source fiable pour s'assurer de leur pertinence. Au sein d'AXA Partners, cette source est l'entité "Portfolio & Monitoring Management" ( sigle PMP par la suite). Sa méthode de récupération des données est différente avec une vision moins segmentée que notre équipe "Actuarial Services" (sigle AS par la suite), pas de vision par âge par exemple.

#### Tous produits confondus

Il convient de commencer avec une vue d'ensemble **tous produits et toutes couvertures confondus** des fréquences par année d'incident de chaque entité. Les courbes correspondent aux fréquences et les barres verticales correspondent à l'exposition des deux sources de données (le bleu pour PMP, l'orange pour AS) (figure 1.8).

**Dans la suite de ce mémoire, les fréquences et l'exposition sont toujours représentées comme sur ce graphique**. Plus l'exposition est élevée (et donc la barre haute), plus la base est riche et le calcul de la fréquence affiné. C'est pourquoi l'exposition est gardée sur les graphiques de fréquence afin d'avoir un aperçu de la quantité d'informations disponibles au calcul de cette dernière.

Il se trouve que les fréquences ont été calculées pour un niveau d'exposition similaire, et qu'elles sont alignées et suivent la même logique (hausse entre 2013 et 2019 entre autre). De ce point de vue général sur les données, la réconciliation est validée. En outre, en considérant l'exposition comme le nombre

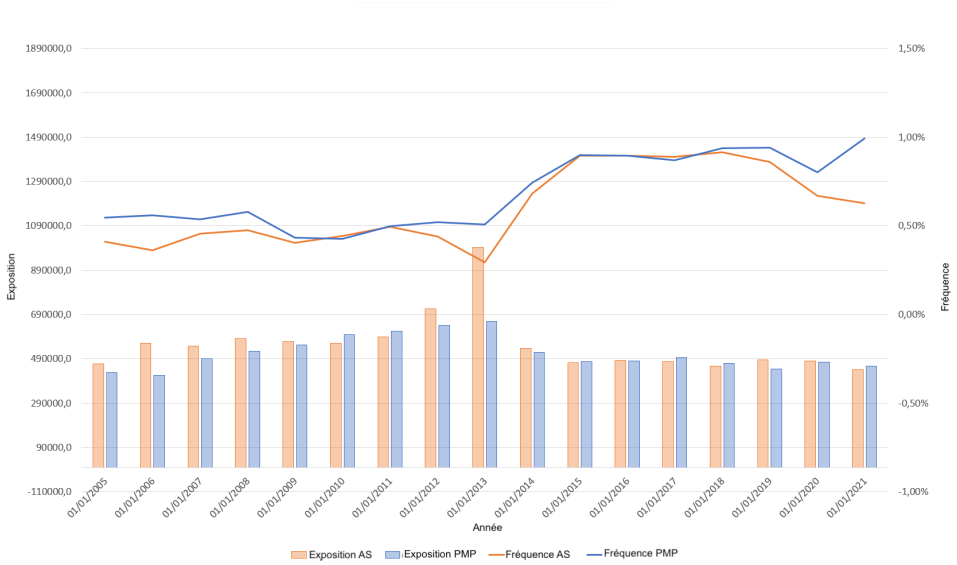



FIGURE 1.8 : Comparaison des fréquences par année obtenues par l'équipe "Actuarial Services" et l'équipe "Portfolio & Monitoring Management" tous produits et toutes couvertures confondus

d'assurés dont les informations sont présentes dans la base, une moyenne de 489 794 d'expositions par année indique un volume de données conséquent pour les modélisations ultérieures.

### Produit par produit

Il convient de regarder maintenant à l'échelle d'un produit si la réconciliation se fait bien, ici le Personal Loan (accident et maladie ensemble) (figure 1.9).

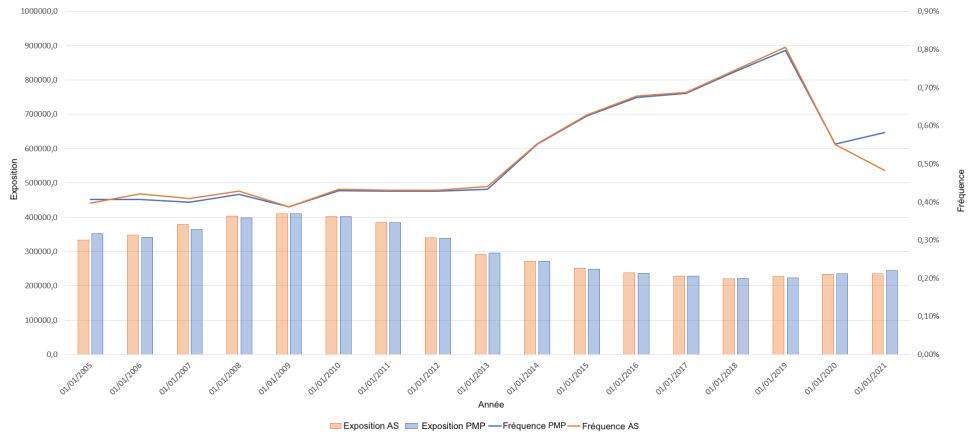



FIGURE 1.9 : Comparaison des fréquences par année obtenues par l'équipe "Actuarial Services" et l'équipe "Portfolio & Monitoring Management" pour le produit Personal Loan toutes couvertures confondues

Pour le Personal Loan, il s'agit ici d'un exemple de réconciliation parfaite. Ce n'est pas toujours le cas comme avec le produit Mortgage Loan (figure 1.10).

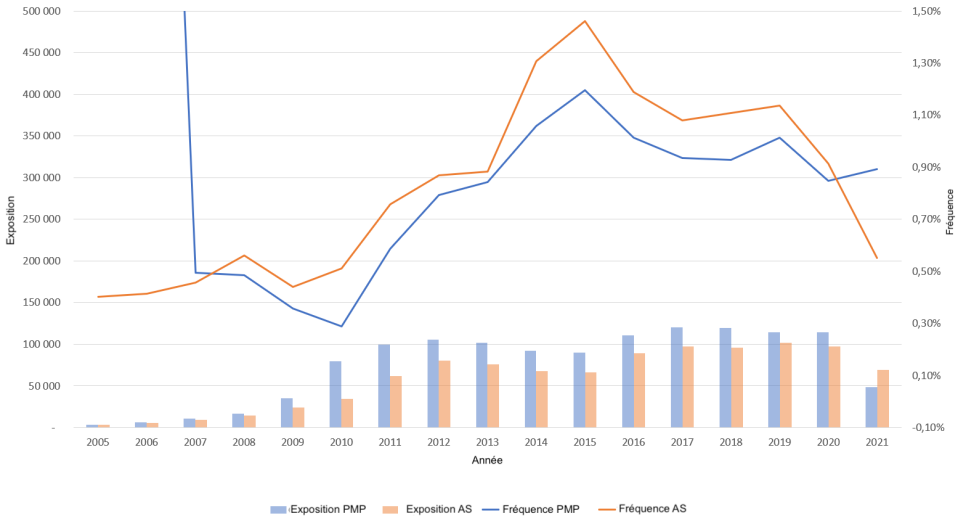



FIGURE 1.10 : Comparaison des fréquences par année obtenues par l'équipe "Actuarial Services" et l'équipe "Portfolio & Monitoring Management" pour le produit Mortgage Loan toutes couvertures confondues

Les deux courbes de fréquences suivent la même logique année par année mais avec un décalage. En effet la fréquence obtenue via la base de données de l'équipe "Actuarial Services" est supérieure à celle de l'autre équipe.

Dans ce cas là, il faut regarder la fréquence partenaire par partenaire et déterminer lequel creuse cet écart. Dans le cas du produit Mortgage Loan, il s'agit du partenaire "F". Il convient alors de chercher à connaître les caractéristiques de ce partenaire. C'est un partenaire qui travaille avec des banques publiques et a un portefeuille atypique avec des fréquences d'accidents plus élevées que les autres partenaires. Les données collectées par AS et PMP sont sensiblement différentes et c'est cette différence qui explique l'écart de fréquence (figure 1.11).

Concernant les autres produits, la réconciliation se fait parfaitement. La qualité des données produites est confirmée, et pour le cas particulier du produit Mortgage Loan, les données relatives au partenaire "F" sont retirées avant de procéder à la modélisation.

La dernière étape avant de procéder au premier GLM est la mise en place de la méthodologie d'étude via quelques analyses descriptives de la base de données.

## 1.6 La méthodologie d'étude

### La différenciation entre données Bulk et non Bulk

Il est nécessaire de regarder dans un premier temps la quantité d'exposition à disposition par produit dans la base de données, ainsi que l'origine de cette exposition (de quelle base *Policies* provient-elle ?) (tableau 1.5).

Le Personal Loan est le produit le plus présent dans la base de données, représentant 65% de l'ex-

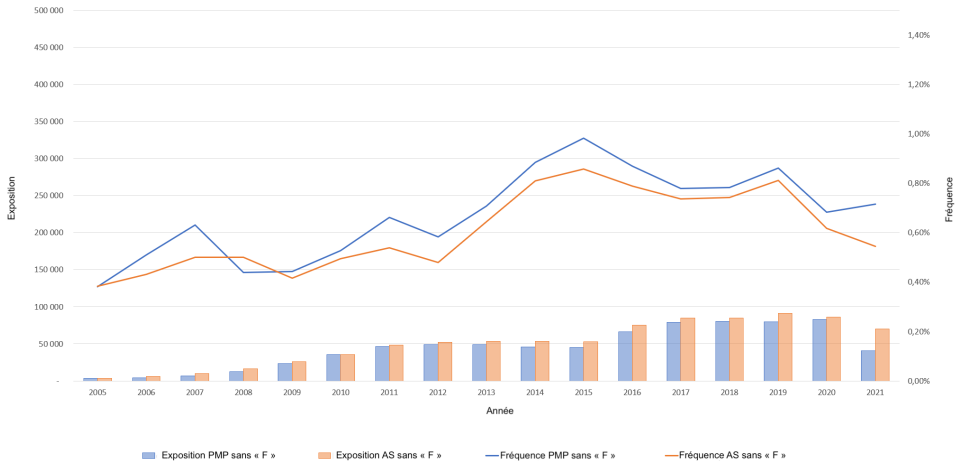



FIGURE 1.11 : Comparaison des fréquences par année obtenues par l'équipe "Actuarial Services" et l'équipe "Portfolio & Monitoring Management" pour le produit Mortgage Loan toutes couvertures confondues sans le partenaire "F"

| Produit | Exposition | % de l'exposition issue de la base Upfront | % de l'exposition issue de la base MR | % de l'exposition issue de la base Bulk |
|---|---|---|---|---|
| Car Loan | 567 730 | 97% | 3% | 0% |
| Credit Card | 660 526 | 0% | 2% | 98% |
| Income Protection | 250 749 | 0% | 2% | 98% |
| Waiver of Premium | 273 250 | 0% | 0% | 100% |
| Mortgage Loan | 846 335 | 39% | 25% | 36% |
| Personal Loan | 4 729 951 | 99% | 0% | 1% |

TABLE 1.5 : Exposition par produit, et proportion de l'exposition par base *Policies*

position totale. Il se trouve que les produits Credit Card, Income Protection et Waiver of Premium ont respectivement 98%, 98% et 100% de leur exposition issue de la base *Policies* Bulk. Comme il a été vu précédemment (1.5.1), les données Bulk ne sont pas exploitables dans la modélisation du GLM compte tenu du manque d'information à disposition pour des variables explicatives telles que l'âge et le genre. **C'est pourquoi ces produits sont retirés du calibrage du GLM, et le peu de données Bulk des autres produits est également retiré**.

**Le calibrage du GLM se fait donc sur les produits Car Loan, Personal Loan et Mortgage Loan, auxquels sont retirés les données Bulk, et les produits Credit Card, Waiver of Premium et Income Protection sont retirés du calibrage du GLM.**

Dans la précédente modélisation, ces trois produits avec des données Bulk avaient également été retirés de la modélisation GLM, ne proposant qu'une fréquence commune à tous les âges et tous les genres pour chaque produit (tableau 1.6).

Dans la nouvelle modélisation, ces trois produits sont ajoutés à l'équation du GLM via une méthodologie interne explicitée ultérieurement, mais ne participent pas à sa calibration.

| Produit | Précédente modélisation |
|---|---|
| Car Loan | GLM par âge/genre |
| Credit Card | 0,45% |
| Income Protection | 1,70% |
| Waiver of Premium | 0,30% |
| Mortgage Loan | GLM par âge/genre |
| Personal Loan | GLM par âge/genre |

TABLE 1.6 : Récapitulatif de la précédente modélisation

**Séparation de la modélisation de l'incapacité en deux GLMs : accident et maladie**

La précédente modélisation GLM expliquait la fréquence de la maladie et de l'accident ensemble, avec comme variables explicatives l'âge et le genre pour chaque produit. Par rapport à l'ancienne modélisation, **la modélisation de la fréquence est séparée en deux GLMs, un pour chaque couverture : accident et maladie.**

Cette séparation est faite car il est observé suffisamment de différences entre les deux couvertures (figure 1.12) :

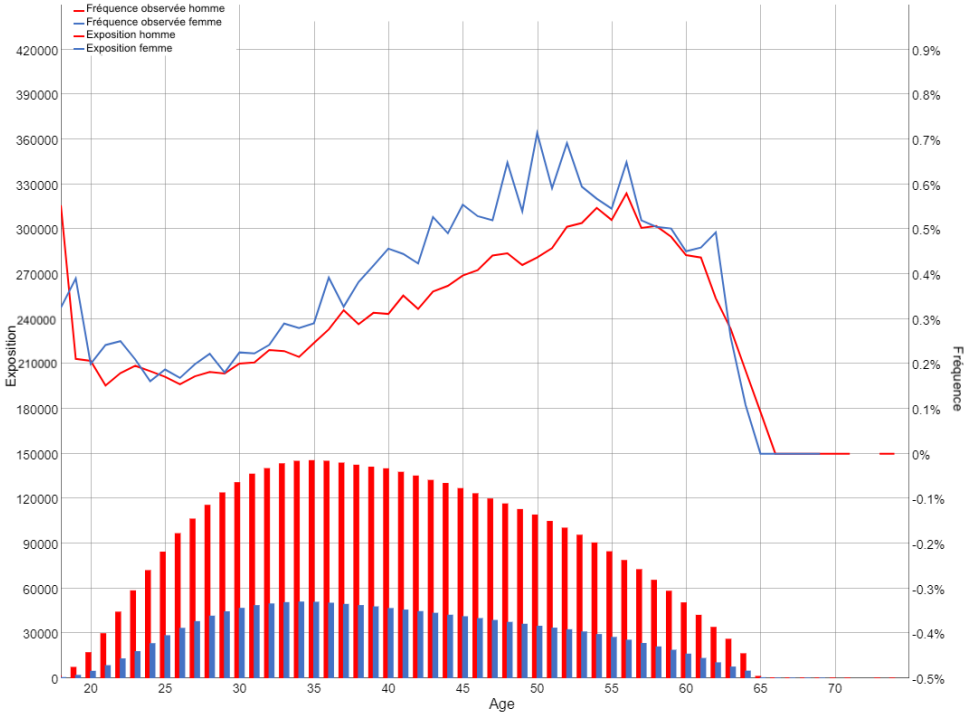
- pour le GLM maladie, il y a une croissance exponentielle de la fréquence avec l'âge, avec des fréquences très proches entre les hommes et les femmes. **Les variables explicatives retenues pour ce GLM sont donc l'âge et le produit ;**
- pour le GLM accident, il y a une séparation nette entre les fréquences des hommes et des femmes, l'âge n'influençant pas la fréquence d'accident (jusqu'à 55 ans). **Les variables explicatives retenues pour ce GLM sont donc le genre et le produit.**

Il est choisi dans cette modélisation de garder une variable âge continue plutôt que de faire des catégories d'âge car la tendance exponentielle de la fréquence avec l'âge est clairement visible. Le produit est également directement intégré aux GLMs en tant que variable explicative plutôt que de faire un GLM par produit car les tendances observées par âge et genre sont communes pour chaque produit. Il faut faire attention au fait que la première modélisation GLM de cette étude n'est pas un GLM, mais deux GLMs dont les résultats sont sommés pour les comparer à la précédente modélisation.

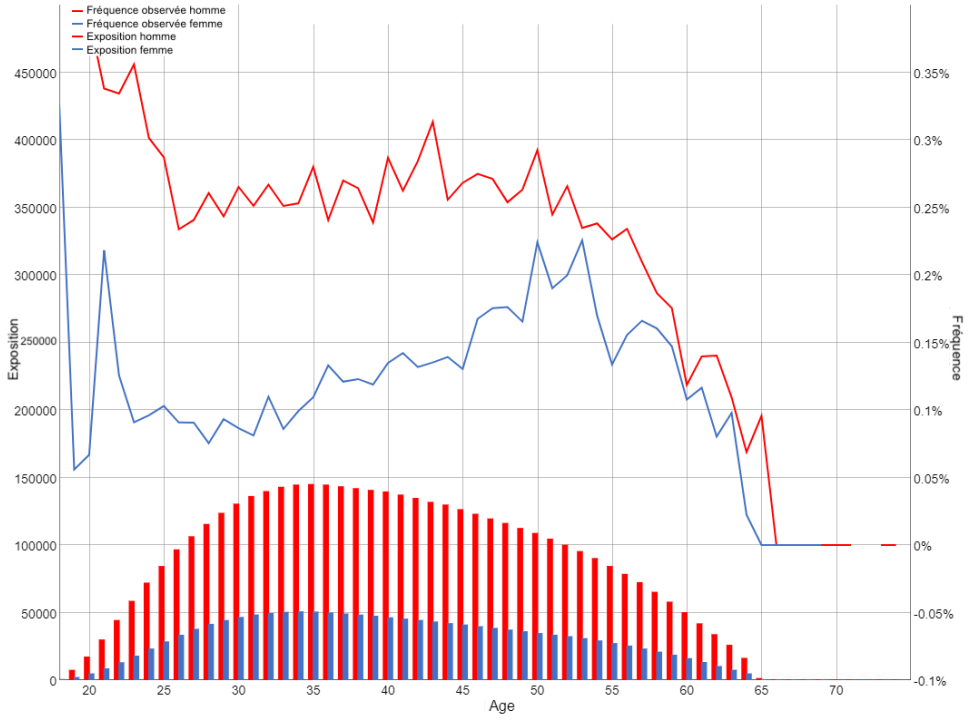
Les GLMs ne sont pas calibrés avec d'autres variables explicatives que l'âge, le genre et le produit à cause des contraintes suivantes :

- **le modèle doit être utilisé dans l'outil de tarification d'AXA Partners qui ne peut accueillir que les variables âge et genre par produit ;**
- le modèle doit répondre aux exigences des souscripteurs (ceux qui utilisent les calculs de fréquences pour calculer les primes) c'est-à-dire une interprétation intuitive des variables et une justification des écarts avec les anciennes hypothèses de fréquences ;
- tous les résultats doivent être validés avec l'équipe Risk Management.

En outre, ces graphiques (figure 1.12) montrent que le portefeuille de clients est composé avant tout d'hommes (l'exposition des hommes représentée par les barres rouges est plus haute que celle des




(a) Fréquences observées par âge et genre pour la couverture maladie




(b) Fréquences observées par âge et genre pour la couverture accident

FIGURE 1.12 : Fréquences observées par âge et genre pour les deux couvertures

femmes représentée par les barres bleues pour chaque âge). Il est important de rappeler que depuis le 21 décembre 2012, les assureurs ont interdiction de proposer des primes différentes selon le genre

du client. C'est pourquoi la différentiation faite dans cette modélisation sur le genre est appuyée par la distribution du genre dans le portefeuille pour proposer un prix commun aux hommes et aux femmes. Par exemple, si le portefeuille d'assuré compte 80% d'hommes, la prime commune est calculée en prenant d'avantage les résultats de fréquence obtenus pour les hommes que pour les femmes, via l'utilisation d'un coefficient de proportionnalité.

### Evolution de la fréquence par année d'incident pour chaque produit TTD

Regardons l'évolution des fréquences par produit en fonction des années d'incidence (couvertures maladie et accident ensemble, figure 1.13). Il se remarque une tendance croissante avec les années d'incidents pour les produits Personal Loan (en rouge), Mortgage Loan (en vert) et Income Protection (en marron). Le produit Income Protection a lui une fréquence observée très volatile liée à une faible exposition dans le jeu de données. Comme vu précédemment (1.4), ces hausses de fréquences sont expliquées par la prise de conscience des Portugais pour les couvertures d'assurance dont ils disposent, mais ne sont pas présentes pour tous les produits du portefeuille.

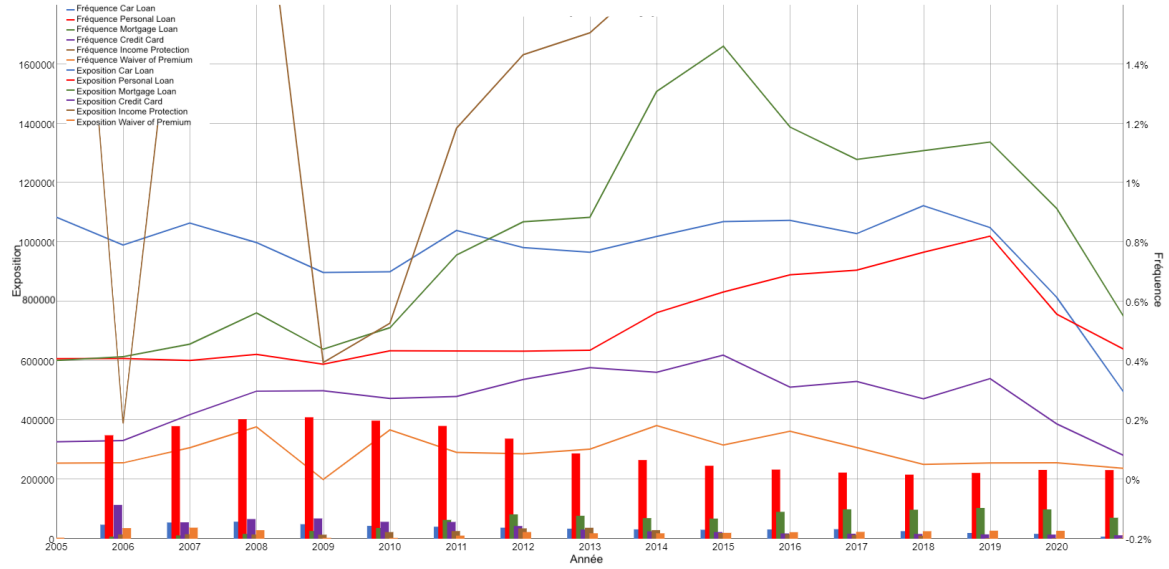



FIGURE 1.13 : Fréquences observées par produit et par année d'incident, maladie et accident ensemble

La capacité à prédire la hausse de la fréquence par année est l'élément central de comparaison des trois modèles implémentés. Il convient de comparer les prédictions pour le produit Personal Loan, qui représente 65% de l'exposition totale (donc 65% du portefeuille).

### Conclusion du chapitre 1

Les informations nécessaires avant de commencer les modélisations sont désormais introduites : au-delà de présenter les produits et les données, leur impact sur la modélisation a été explicité. Le marché portugais et ses caractéristiques ont été présentés, ainsi que la base de données de l'étude et la validation de sa qualité pour son exploitation. Les deux GLMs qui vont être implémentés ont été mis en avant, tout comme le choix des variables explicatives et les contraintes qui s'appliquent dessus.

La première modélisation peut maintenant se faire, celle des deux GLMs sous les contraintes introduites précédemment (1.6). Cela permettra une première vérification de la pertinence ou non du choix d'un modèle linéaire généralisé afin de modéliser le risque incapacité, avant toute comparaison avec d'autres modélisations comme le *gradient boosting* pour répondre à la problématique.

# Chapitre 2

# Première modélisation GLM sous contraintes

Le chapitre 1 a permis de préparer le jeu de données, ainsi que de faire les premiers choix de modélisation liés aux caractéristiques des produits, et du marché dans lequel ils évoluent. Il convient maintenant de calibrer et valider le premier modèle de fréquence. Il est important de rappeler que dans cette partie le modèle doit répondre à certaines contraintes. Il doit pouvoir être utilisé dans l'outil de tarification d'AXA Partners qui ne peut accueillir que les variables âge et genre par produit et répondre aux exigences des souscripteurs.

Dans un premier temps, la pertinence du choix du GLM Poisson pour modéliser la fréquence incapacité est rappelée. Puis, après avoir calibré les équations des GLMs, les résultats obtenus sont validés statistiquement. Enfin les équations sont ajustées pour répondre aux limites précédemment introduites (1.6), et un exemple de calcul de prime pure est implémenté pour être comparé aux deux modélisations suivantes (pour pouvoir comparer les modèles d'un point de vue tarification).

## 2.1 Le choix du GLM comme modèle de fréquence

Le modèle linéaire généralisé est une extension de la régression linéaire dont le principe est rappelé ci-dessous.

### 2.1.1 La régression linéaire

La régression linéaire est un modèle statistique qui permet d'estimer les effets de plusieurs variables prédictives qui peuvent être numériques ou catégorielles sur une unique variable à expliquer (ou variable réponse) qui est de type numérique continue (McCullagh et Nelder (1989)).

Notons $X_{i1}$, $X_{i2}$, ..., $X_{ip}$ les p variables explicatives pour l'observation i (avec n le nombre d'observation soit $i \in \{1, 2, ..., n\}$), $Y_i$ la variable à expliquer pour l'individu i, $\beta_0$, $\beta_1$,...,$\beta_p$ les paramètres du modèle à estimer ($\beta_0$ correspond à l'ordonnée à l'origine) et $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ le terme d'erreur qui correspond à l'écart entre la variable observée et la variable à modéliser.

Cela donne la relation linéaire $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i$.

Par la suite, l'estimation des paramètres $\beta_j$ peut se faire via un maximum de vraisemblance, la vraisemblance étant la densité de probabilité associée aux données. Pour cela introduisons les notations suivantes :

- $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^t$, les paramètres du modèle à estimer ;
- $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^t$, le vecteur erreur ;
- $\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & ... & X_{1p} \\ 1 & X_{21} & X_{22} & ... & X_{2p} \\ ... & ... & ... & ... & \\ 1 & X_{n1} & X_{n2} & ... & X_{np} \end{pmatrix}$, la matrice des variables explicatives pour chaque observation ;
- $\boldsymbol{Y} = (Y_1, ..., Y_n)^t$, le vecteur variable à expliquer.

Le modèle linéaire s'écrit $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$.

En supposant que $y_i$ est la réalisation de $Y_i$ et en notant $\boldsymbol{y} = (y_1, ..., y_n)^t$, le vecteur des réalisations, la vraisemblance associée au modèle s'écrit

$L(\boldsymbol{\beta}, \boldsymbol{y}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)'(\boldsymbol{y} - \boldsymbol{X\beta})\right)$.

Il s'en déduit par maximum de vraisemblance que l'estimateur $\hat{\boldsymbol{\beta}}$ des paramètres du modèle est $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$.

Il faut s'assurer en amont que $\boldsymbol{X}^t\boldsymbol{X}$ est bien inversible.

Le modèle GLM est maintenant introduit, il s'agit d'une extension de la régression linéaire.

### 2.1.2 Le modèle linéaire généralisé

Le principal intérêt du GLM est qu'il est le meilleur compromis entre pouvoir prédictif et capacité d'expliquer le modèle à des personnes qui ne maîtrisent pas forcement les outils statistiques. Le GLM permet de manipuler des données de type binaire et de type comptage (d'où son intérêt dans cette étude puisqu'il s'agit d'un comptage d'un nombre de sinistres). En effet le modèle de régression linéaire classique ne peut pas modéliser ce dernier type de données parce que celui-ci suppose que les variables réponses sont distribuées selon une loi normale, ce qui implique que la variance des résidus est homogène donc constante. Sauf que les données de comptage ne sont pas distribuées selon une loi normale mais selon une loi de Poisson (dans le cas d'un GLM Poisson), et donc la variance des résidus n'est pas constante mais dépend du comptage moyen que prédit le modèle (McCullagh et Nelder (1989), Monnier (2016)).

Plusieurs éléments définissent un Modèle Linéaire Généralisé.

Il y a le prédicteur linéaire qui est la combinaison linéaire des variables explicatives. Il est noté $\eta$ et $\eta = \boldsymbol{X\beta}$ (avec les notations précédemment définies).

Il y a la fonction de lien. Dans une régression classique, les valeurs prédites par le prédicteur linéaire correspondent à la prédiction moyenne d'une observation. Dans un GLM ce n'est pas le cas, elles correspondent à une transformation de la prédiction moyenne d'une observation. Il convient donc d'introduire la fonction de lien $g$ tel que $g(\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]) = \boldsymbol{X\beta} = \eta$, avec $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]$ l'espérance conditionnelle de $\boldsymbol{Y}$ sachant $\boldsymbol{X}$ (tableau 2.1).

| Type de la variable réponse | Distribution de la variable réponse | Fonction de lien | Fonction de la moyenne |
|---|---|---|---|
| Binaire | Binomiale | $(Logit)\ \beta_0+\sum_{j=1}^{p} X_{ij}\beta_j = \log(\frac{\mu}{1-\mu})$ | $\mu = \frac{1}{1+\exp(-\beta_0-\sum_{j=1}^{p} X_{ij}\beta_j)}$ |
| Quantitatif continu | Gaussienne | $(Identité)\ \beta_0+\sum_{j=1}^{p} X_{ij}\beta_j = \mu$ | $\mu = \beta_0+\sum_{j=1}^{p} X_{ij}\beta_j$ |
| Comptage | Poisson | $(Log)\ \beta_0+\sum_{j=1}^{p} X_{ij}\beta_j = \log(\mu)$ | $\mu = \exp(\beta_0+\sum_{j=1}^{p} X_{ij}\beta_j)$ |
| Comptage | Binomiale négative | $(Log)\ \beta_0+\sum_{j=1}^{p} X_{ij}\beta_j = \log(\mu)$ | $\mu = \exp(\beta_0+\sum_{j=1}^{p} X_{ij}\beta_j)$ |

TABLE 2.1 : Récapitulatif des différentes fonctions de lien selon le type de variable à expliquer

Dans le cas des données de comptage de type Poisson, la fonction de lien est la fonction log et il s'obtient $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) = \exp(\boldsymbol{X}\boldsymbol{\beta})$.

La fonction de lien permet de contraindre les valeurs prédites à être dans l'échelle des valeurs observées. Dans le cas d'un GLM de comptage sur les données, la fonction de lien log contraint les valeurs prédites à être positives ou nulles (l'inverse de la fonction log étant la fonction exp).

Enfin il y a la distribution de la variable à expliquer. Dans un GLM de comptage Poisson, la variable réponse est supposée suivre une distribution de Poisson dont la définition est rappelée. Si le nombre moyen d'occurrences dans un intervalle de temps fixé est $\lambda$, alors la probabilité qu'il y ait exactement $k$ occurrences (avec $k \in \mathbb{N}$) est $\mathbb{P}(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$, avec $X$ la variable aléatoire qui suit une loi de Poisson de paramètre $\lambda$. Son espérance et sa variance sont égales à son paramètre $\lambda$, $\mathbb{E}[X] = \mathbb{V}[X] = \lambda$.

Comme pour la régression linéaire, les valeurs des paramètres du GLM sont déterminées par la méthode du maximum de vraisemblance.

Pour valider le bon choix de modélisation, à savoir le choix d'une distribution de Poisson et une fonction de lien log, il faut vérifier que les réponses soient indépendantes, c'est-à-dire pas de relation entre les variables réponses d'une ligne à l'autre dans le jeu de données. Il est donc important que les données ne présentent pas de répétition avant d'appliquer le GLM de type Poisson.

Il faut également que les réponses soient distribuées selon une loi de Poisson de paramètre $\lambda$. Cela se vérifie en comparant la distribution des comptages observés avec la distribution théorique de la loi de Poisson de paramètre $\lambda$ égal à la moyenne des comptages observés.

Il ne doit pas y avoir de surdispersion. Théoriquement la variance des variables réponses est égale à leur moyenne. Il y a surdispersion lorsque la variance des variables réponses est supérieure à la variance théorique. Cela entraîne que l'erreur standard des paramètres du modèle est sous-estimée. Tout cela peut conduire à une p-value très faible et un rejet de l'hypothèse statistique du lien entre les variables explicatives et la variable réponse. Les principales raisons de la surdispersion sont l'absence d'au moins une variable explicative forte, la corrélation entre les variables réponses, la présence importante des valeurs zéro par rapport à ce qui s'attend d'une distribution de Poisson de paramètre $\lambda$. En cas de surdispersion, il sera intéressant de remplacer la distribution de Poisson par une distribution binomiale négative, mieux adaptée à ce type de données.

Pour vérifier si le modèle GLM présente une surdispersion, il faut calculer

$$\phi = \frac{\text{variance observée}}{\text{variance théorique}} = \frac{\text{variance observée}}{\text{moyenne}} = \frac{\text{variance observée}}{\lambda}.$$

En pratique, $\phi$ est estimé avec le ratio de la déviance résiduelle (la déviance est définie dans la section 2.2.2) sur le nombre de degrés de libertés du modèle soit

$$\phi = \frac{\text{déviance résiduelle}}{\text{nombre de degrés de libertés}}.$$

Si $\phi > 1$, alors il y a surdispersion.

Enfin il faut vérifier plusieurs hypothèses sur les résidus. Soit le i-ème résidu $\hat{\epsilon}_i$ tel que $\hat{\epsilon}_i = Y_i - \hat{Y}_i$, avec $Y_i$ la variable à expliquer de l'individu i et $\hat{Y}_i$ la prédiction du GLM. Il convient d'introduire les résidus de déviance. Le i-ème résidu de déviance est $r_i^D = Signe(\hat{\epsilon}_i) \times \sqrt{d_i}$ avec $d_i$ la contribution de la i-ème observation à la déviance du modèle, telle que Déviance $= \sum_{i=1}^{n} d_i$.

Il faut que les résidus soient distribués selon une loi de Poisson. Mais les résidus de la réponse étant difficiles à interpréter pour un GLM Poisson, il est plus judicieux de regarder si les résidus de la déviance sont indépendants, linéaires et suivent une distribution normale, afin d'étudier la qualité du choix du GLM Poisson.

Une fois le choix de modélisation validé, le modèle GLM est calibré et il est possible de valider la performance et la qualité de prédiction du modèle via différents tests (déviance, validation croisée).

Avant cela, il est nécessaire de rappeler les choix effectués pour modéliser les données TTD au Portugal, à savoir ce qui change avec l'ancienne modélisation, quelles fréquences vont être modélisées, quel type de calibration des GLMs va être choisi et pourquoi.

### 2.1.3 Rappel des choix de modélisation avant la calibration des GLMs

Les anciennes hypothèses ont été validées en 2018 sur une expérience s'arrêtant fin 2017. Il y a maintenant 5 années d'expérience supplémentaires dans le portefeuille, et il est classique chez AXA Partners de remettre à jours les hypothèses GLM tous les 4/5 ans.

Il convient de rappeler que l'ensemble des produits Credit Protection Insurance et Lifestyle Protection sont concernés en TTD au Portugal : Personal Loan, Mortgage Loan, Car Loan, Credit Card, Waiver of Premium et Income Protection.

Lors de la précédente étude des fréquences en 2018, un seul GLM Poisson a été calibré pour les deux types de sinistres (accident et maladie) avec comme variable explicative le genre et l'âge par produit. Il a été observé dans le chapitre 1 deux tendances bien distinctes de fréquence par âge entre accident et maladie (1.6) :

- pour les données maladie, une tendance exponentielle croissante avec l'âge de la fréquence proche entre les hommes et les femmes pour tous les produits ;
- pour les données accident, une tendance horizontale avec l'âge avec une réelle distinction entre les hommes et les femmes pour tous les produits.

Ces tendances sont communes à tous les produits "non Bulk" (Personal Loan, Mortgage Loan et Car Loan). Il a été choisi en conséquent de ne pas faire deux GLMs accident et maladie par produit, mais d'intégrer la variable produit directement dans les GLMs afin d'avoir le même coefficient pour le genre et pour l'âge par produit.

Les deux modélisations sont donc les suivantes :

- la modélisation de la couverture maladie se fait via un GLM de comptage Poisson, de fonction de lien logarithmique et basé sur les variables explicatives : âge et produit couvert ;
- la modélisation de la couverture accident se fait via un GLM de comptage Poisson, de fonction de lien logarithmique et basé sur les variables explicatives : genre et produit couvert.

La variable réponse étant de type comptage, un GLM avec une distribution Poisson semble adapté. Il faut maintenant valider statistiquement que la distribution observée du nombre de sinistres est bien une distribution de Poisson avant la calibration des GLMs. Pour faire ce test, la base de données utilisée est filtrée sur les deux principales contraintes de données : les données de type "Bulk" sont exclues (1.5.1) et le partenaire "F" est exclu (1.5.3).

Ce test est séparé en deux étapes, d'abord une comparaison visuelle de la distribution théorique avec celle observée, puis un test d'ajustement du $\chi^2$.

### 2.1.4 Validation du choix de la modélisation Poisson avant calibration des GLMs

Il y a 5 059 526 données pour la couverture maladie, et 5 050 504 données pour la couverture accident. Il se trouve que les échantillons de données présentent une très grande quantité de valeurs nulles pour le nombre de sinistres (tableaux 2.2).

| Nombre de sinistres | Occurrences observées |
|---|---|
| 0 | 5 041 931 |
| 1 | 17 096 |
| 2 | 489 |
| 3 | 9 |
| 4 | 1 |

(a) Distribution du nombre de sinistres dans les données maladie

| Nombre de sinistres | Occurrences observées |
|---|---|
| 0 | 5 038 569 |
| 1 | 11 619 |
| 2 | 310 |
| 3 | 6 |
| 4 | 0 |

(b) Distribution du nombre de sinistres dans les données accident

TABLE 2.2 : Distribution des sinistres dans les données

La moyenne du nombre de sinistres est de 0.00357 pour la couverture maladie et de 0.00243 pour la couverture accident. Ainsi, si le nombre de sinistres pour chacune des couvertures suit une distribution de Poisson, en théorie le paramètre $\lambda$ des deux distributions Poisson est égal à ces moyennes. Une première vérification simple est de regarder si les espérances empiriques obtenues pour les deux jeux de données sont égales aux variances empiriques respectives. Il y a pour la maladie une variance de 0.00377 et une variance de 0.00255 pour l'accident. Les variances et les espérances empiriques sont bien proches.

Maintenant vérifions visuellement et via des tests si les variables réponses suivent des distributions de Poisson.

#### Vérification visuelle

5 059 526 tirages d'une loi de Poisson de paramètre $\lambda = 0.00357$ sont simulés pour la maladie et 5 050 504 tirages d'une loi de Poisson de paramètre $\lambda = 0.00243$ sont simulés pour l'accident (tableaux 2.3).

Pour 0 et 1 sinistres (soit 99.99% des données maladie et accident), la modélisation via la distribution de Poisson semble pertinente.

Un test d'ajustement du $\chi^2$ est maintenant mis en place.

| Nombre de sinistres | Occurrences observées | Occurrences théoriques |
|---|---|---|
| 0 | 5 041 931 | 5 041 453 |
| 1 | 17 096 | 18 040 |
| 2 | 489 | 32 |
| 3 | 9 | 0 |
| 4 | 1 | 0 |

(a) Distribution du nombre de sinistres dans les données maladie et la distribution théorique

| Nombre de sinistres | Occurrences observées | Occurrences théoriques |
|---|---|---|
| 0 | 5 038 569 | 5 038 262 |
| 1 | 11 619 | 12 227 |
| 2 | 310 | 15 |
| 3 | 6 | 0 |
| 4 | 0 | 0 |

(b) Distribution du nombre de sinistres dans les données accident et la distribution théorique

TABLE 2.3 : Distribution des sinistres dans les données et les distributions de Poisson théoriques associées

### Test d'ajustement du $\chi^2$

Il convient de rappeler que le test du $\chi^2$ est un test statistique qui permet de vérifier si les observations suivent une loi de probabilité particulière, en l'occurence une distribution de Poisson.

Soit $X_m$ la variable aléatoire des observations maladie, $\lambda_m = 0.00357$ le paramètre de la loi de Poisson théorique associée à la maladie, $X_a$ la variable aléatoire des observations accident et $\lambda_a = 0.00243$ le paramètre de la loi de Poisson théorique associée à l'accident.

Les hypothèses des deux tests sont les suivantes :

$H_0$ : "$X_i$ suit la loi $\mathcal{P}(\lambda_i)$" contre $H_1$ : "$X_i$ ne suit pas la loi $\mathcal{P}(\lambda_i)$" avec $i \in \{m, a\}$.

Pour les deux couvertures, la fonction chisq.test() du package `stats` du logiciel R CORE TEAM (2022) donne une p-value du test inférieure à 2.2e-16 donc l'hypothèse $H_0$ devrait être rejetée. Cependant le message d'erreur suivant s'affiche : "L'approximation du Chi-2 est peut-être incorrecte".

Il peut être intéressant de procéder alors à un test du $\chi^2$ de niveau $\alpha = 0.05$ manuel pour vérifier le rejet de l'hypothèse $H_0$ (SAPORTA (2006)).

Pour la maladie, il faut vérifier que $\chi^2 = \sum_{i=0}^{2} \frac{(n_i - T_i)^2}{T_i} > \chi^2_{2;\alpha=0.05}$ (2 car il y a 3 degrés de liberté dans ce test), avec les éléments précisés ci-dessous (tableau 2.4).

| i : Nombre de sinistres | 0 | 1 | ≥ 2 |
|---|---|---|---|
| Ti : Observations théoriques | 5041453 | 18040 | 32 |
| ni : Observations réelles | 5041931 | 17096 | 499 |

TABLE 2.4 : Les observations réelles et les observations théoriques pour les données maladie

Cela donne $\chi^2 = 6865 > 5.99 = \chi^2_{2;0.05}$, donc l'hypothèse $H_0$ est largement rejetée. Les observations du nombre de sinistres pour les données maladie ne suivent pas une distribution de Poisson. Il se montre de façon analogue que ce n'est pas non plus le cas pour les données accident.

Ce test a du sens puisqu'il vérifie les critères de Cochran de 1954 (STEELE (2003)) :

- $T_i \geq 1$, pour chaque i ;
- il y a un maximun de 20% des valeurs $T_i$ qui sont moins grandes que 5.

Cependant, il peut être considéré que l'échec de la vérification du test vient du fait que la moyenne du nombre de sinistres est très proche de zéro avec 99.65% de valeurs nulles dans les données maladie et

99.76% dans les données accident. Il a bien été vu en amont que les espérances et variances empiriques étaient proches et que les distributions du nombre de sinistres dans les bases de données semblaient, visuellement, suivre une loi de Poisson de paramètre égale aux moyennes empiriques.

C'est pourquoi malgré la réponse négative du test du $\chi^2$, le GLM est tout de même calibré via une distribution de Poisson.

## 2.2 Modélisation et validation des nouvelles équations de fréquences

Regardons maintenant en détail les résultats des deux GLMs, à savoir les valeurs des paramètres obtenues pour expliquer la fréquence (la variable à expliquer du GLM est le nombre de sinistres et l'exposition est *offsetée* pour calculer la fréquence), la qualité prédictive des modèles et le calibrage obtenu pour un produit en particulier : le Personal Loan. Enfin un calcul de prime pure est proposé pour ce produit ainsi que la méthodologie d'ajout des produits "Bulk" aux équations.

### 2.2.1 Calibration des coefficients des GLMs

En amont, les données sont séparées en deux jeux de données, un accident et un maladie et des filtres sont appliqués sur les données.

La période est comprise entre 2005 et 2019 pour avoir un large spectre de données et éviter les années covid. Les années covid sont retirées car il a été observé une baisse significative de la fréquence de tous les produits. Cela est lié au confinement et à l'application des gestes barrières qui ont réduit les accidents du travail et les maladies saisonnières. Avec le retour à la vie normale courant 2022, il est supposé que les fréquences vont de nouveau augmenter et rejoindre les niveaux antérieurs à 2020.

Les données pour les âges supérieurs à 60 ans sont retirées, car les fréquences sont plafonnées à partir de cet âge là et les données qui correspondent à des sinistres produits alors que le délai de carence n'est pas terminé sont également retirées(1.2.4).

La fonction glm() du package `stats` de R Core Team (2022) permet d'obtenir les GLMs suivants pour expliquer la fréquence.

Les résultats du GLM maladie sont regardés en premier (tableau 2.5).

| Nom de la variable explicative | Coefficient associé | p-value |
|---|---|---|
| Produit Car Loan (et intercept) | -6,8352643 | <2e-16 |
| Produit Mortgage Loan | 0,0071672 | 0.808 |
| Produit Personal Loan | -0,4930276 | <2e-16 |
| Âge | 0,0360684 | <2e-16 |

Table 2.5 : Variables, coefficients et p-values obtenus pour le GLM de la couverture maladie

Ce tableau associe à chaque variable explicative le coefficient obtenu lors de la calibration du GLM. La dernière colonne correspond à la p-value du test statistique de la significativité de chacune des variables explicatives. Le test est le suivant :

$H_0$ : "Le coefficient associé à cette variable explicative est nul c'est-à-dire que la variable n'explique pas la variable réponse" contre $H_1$ : "La variable est utile pour expliquer la variable réponse".

Généralement un niveau significatif $\alpha = 0.05$ est choisi, et si la p-value obtenue est inférieure à ce niveau, l'hypothèse $H_0$ est rejetée et la variable explicative est considérée comme statistiquement significative.

Pour la couverture maladie, les produits Car Loan et Personal Loan ainsi que l'âge ont une p-value très faible (<2e-16) ce qui implique que ces variables explicatives sont significatives. Le produit Mortgage Loan a quant à lui une p-value $= 0.808 > \alpha = 0.05$, donc cette variable explicative n'est pas considérée comme significative dans l'explication de la fréquence de la maladie. En revanche, d'un point de vue interne à AXA Partners, cette variable est gardée car il est nécessaire de segmenter l'explication de la fréquence par produit.

Les résultats du GLM accident sont maintenant étudiés (tableau 2.6).

| Nom de la variable explicative | Coefficient associé | p-value |
|---|---|---|
| Produit Car Loan (et intercept) | -5,46392 | <2e-16 |
| Produit Mortgage Loan | -0,11019 | 0,00195 |
| Produit Personal Loan | -0,6206 | <2e-16 |
| Genre femme | -0,7547 | <2e-16 |

TABLE 2.6 : Variables, coefficients et p-values obtenus pour le GLM de la couverture accident

A la place de l'âge il y a maintenant une variable explicative pour le genre femme, cette valeur est nulle s'il s'agit des hommes. La colonne des p-values indique que toutes les variables explicatives choisies sont statistiquement significatives (<0.05), même le produit Mortgage Loan.

**Ainsi, cette première analyse conclut sur l'importance statistique de toutes les variables explicatives choisies dans les GLMs.** En outre, il est possible de vérifier l'absence d'interaction entre les variables âge et produit pour le GLM maladie et entre les variables genre et produit pour le GLM accident, en calibrant ces GLMs avec interaction entre les variables explicatives. En effet il en résulte des coefficients quasiment identiques avec les GLMs précédents, et des coefficients très faibles pour les variables d'interaction.

Il convient à présent d'analyser la qualité prédictive des deux modélisations et la robustesse des variables explicatives en expliquant la déviance et en procédant à une validation croisée. Les GLMs maintenant calibrés, il est également possible de procéder à deux dernières vérifications du choix de la modélisation Poisson, étudier la présence de surdispersion et les hypothèses sur les résidus.

### 2.2.2 Choix de modélisation et robustesse des équations des GLMs

Il est important dans un premier temps de définir la déviance.

#### Introduction de la déviance

La déviance est une mesure de la qualité de l'ajustement d'un modèle linéaire généralisé. Ou plutôt, c'est une mesure du mauvais ajustement. Plus la déviance indiquée est élevée, et plus l'ajustement est mauvais. Elle mesure la variation de la log-vraisemblance (MIDI et al. (2010)).

La fonction glm() du package `stats` du logiciel R CORE TEAM (2022) donne deux valeurs de la déviance, la "Null deviance" (déviance nulle) et la "Residual deviance" (déviance résiduelle) avec :

- Null Deviance = 2(logL(**Modèle saturé**) - logL(**Modèle nul**)),
  où degré de liberté = degré de liberté du **Modèle saturé** - degré de liberté du **Modèle nul** ;
- Residual Deviance = 2(logL(**Modèle saturé**) - logL(**Modèle proposé**)),
  où degré de liberté = degré de liberté du **Modèle saturé** - degré de liberté du **Modèle proposé**.
  Avec logL la log-vraisemblance, **Modèle saturé** le modèle possédant autant de paramètres que d'observations et qui estime donc exactement les données, **Modèle nul** le modèle ne contenant que l'intercept et **Modèle proposé** le modèle qui a été choisi.

Il s'obtient une déviance nulle de 178 828 pour 5 059 525 degrés de liberté et une déviance résiduelle de 175 941 pour 5 059 522 degrés de liberté pour le GLM maladie. Par différence la déviance entre le modèle nul et le modèle proposé est de 2 887 pour le GLM maladie.

Il s'obtient une déviance nulle de 126 470 pour 5 050 503 degrés de liberté et une déviance résiduelle de 124 924 pour 5 050 500 degrés de liberté pour le GLM accident. Par différence la déviance entre le modèle nul et le modèle proposé est de 1 546 pour le GLM accident.

Les valeurs de déviance obtenues permettent de vérifier la présence ou non de surdispersion.

#### Surdispersion

Il faut calculer le ratio de la déviance résiduelle sur le nombre de degrés de libertés du modèle GLM maladie et du modèle GLM accident :

- pour la maladie, la déviance résiduelle est égale à 175 941 pour 5 059 522 degrés de liberté soit $\phi = \frac{\text{déviance résiduelle}}{\text{nombre de degrés de libertés}} = \frac{175941}{5059522} = 0.035 < 1$. Il n'y a donc pas de surdispersion dans le modèle ;
- pour l'accident, la déviance résiduelle est égale à 124 924 pour 5 050 500 degrés de liberté soit $\phi = \frac{\text{déviance résiduelle}}{\text{nombre de degrés de libertés}} = \frac{124924}{5050500} = 0.025 < 1$. Il n'y a donc pas de surdispersion dans le modèle.

Il n'y a pas de surdispersion dans les deux modèles GLM, il n'est donc pas nécessaire de choisir une autre structure d'erreur comme une structure binomiale négative. La distribution de Poisson reste adaptée. Regardons à présent l'étude des résidus des modèles.

#### Etude de la distribution Poisson des résidus

Pour vérifier si les résidus suivent bien une distribution de Poisson, il convient de procéder à un test de Kolmogorov-Smirnov. Cela se fait via la fonction ks.test() du package `stats` du logiciel R Core Team (2022) et il s'obtient pour les deux GLMs des p-values inférieures à 2.2e-16 d'où le rejet de la distribution Poisson des résidus. Cependant, l'hétérogénéité et l'hétéroscédasticité des réponses rendent difficile l'étude des résidus, d'où l'étude par la suite des résidus de déviance.

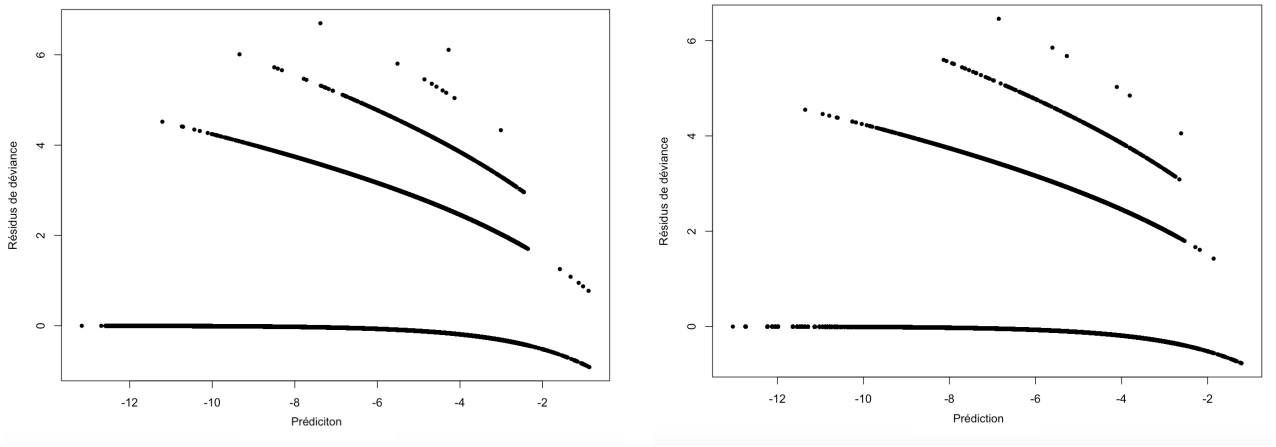
#### Etude de l'indépendance des résidus de déviance

Il convient d'utiliser le test de Ljung-Box sur les résidus de déviance pour vérifier s'ils sont indépendants.

Via la fonction ljungbox.test() du package `tseries` de R Core Team (2022), il s'obtient une p-value inférieure à 2.2e-16 pour les deux modèles GLM. Cela signifie qu'il est raisonnable de penser que les résidus de déviance ne sont pas indépendants. Cela peut entraîner des erreurs dans les estimations des paramètres du modèle et dans les prévisions.

### Etude de la linéarité des résidus de déviance

L'étude de la linéarité des résidus de déviance d'un modèle d'un GLM Poisson est importante pour vérifier si l'hypothèse de linéarité de la relation entre la variable réponse et les variables explicatives est satisfaite. Il convient pour cela de tracer le nuage de points $\{(\hat{\eta}_i, \hat{\epsilon}_i^D)\}_{i=1}^n$, avec $\hat{\epsilon}_i^D$ le i-ème résidu de déviance et $\hat{\eta}_i = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j X_{ij}$, où les $\hat{\beta}_j$ sont les paramètres obtenus après calibration du GLM. Cela donne les graphiques suivants (figure 2.1).




(a) Nuage de points pour le GLM maladie

(b) Nuage de points pour le GLM accident

Figure 2.1 : Nuages de points des deux GLMs

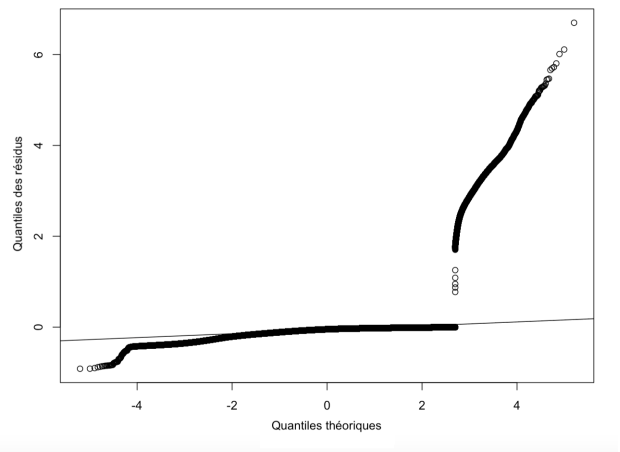
Il s'observe pour les deux GLMs une tendance dans les résidus par rapport à la prédiction. Il s'en déduit qu'il n'y a pas de linéarité de la relation entre la variable réponse et les variables explicatives.

### Etude de la normalité des résidus de déviance standardisés

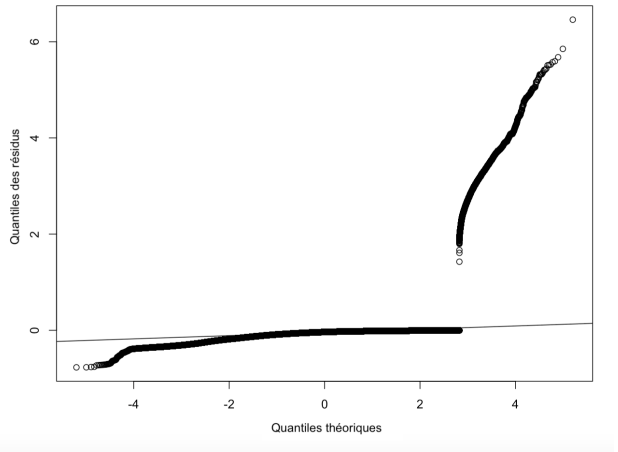
La normalité des résidus de déviance permet d'évaluer dans quelle mesure l'hypothèse de la distribution Poisson de la réponse est respectée. Il faut utiliser un graphique quantile-quantile pour vérifier si les résidus de déviance standardisés suivent une distribution normale centrée réduite. Sur le graphique, cela se vérifie si les points sont alignés sur la diagonale (figure 2.2).

Les résidus de déviance standardisés ne suivent pas une distribution normale centrée réduite dans les deux GLMs. Il est donc compliqué de confirmer par ce biais que la distribution de la réponse est bien une distribution de Poisson.

**Ainsi les résidus ne suivent pas une distribution de Poisson et les résidus de déviance ne sont ni indépendants, ni linéaires et ni distribués normalement. Il est supposé ici que c'est la très grande quantité de valeurs zéro qui est à l'origine des résultats négatifs des précédents tests. L'utilisation d'un GLM semble donc peu adaptée pour les données**




(a) Diagramme quantile-quantile des résidus de déviance standardisés du GLM maladie




(b) Diagramme quantile-quantile des résidus de déviance standardisé du GLM accident

FIGURE 2.2 : Diagramme quantile-quantile des résidus de déviance standardisés des GLMs

**incapacités. Il est possible d'étudier alors l'utilisation d'un modèle ZIP, mieux adapté à ce type de données ou l'utilisation d'un GLM avec une distribution binomiale négative.**

Le modèle ZIP combine deux processus, un premier qui génère des zéros, et un second qui est une distribution de Poisson dont certaines valeurs peuvent être nulles. Soit $X$ la variable aléatoire qui suit une loi de Poisson de paramètre $\lambda$. La distribution ZIP de paramètres $\pi$ et $\lambda$, notée ZIP($\pi$, $\lambda$), a pour distribution

$$\mathbb{P}(X=k) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & \text{si } k = 0 \\ (1-\pi)\frac{\lambda^k}{k!}e^{-\lambda} & \text{si } k \in \{1, 2, ...\} \end{cases} \text{, avec } 0 \leq \pi \leq 1 \text{ et } \lambda \geq 0.$$

Le paramètre $\pi$ est la probabilité de zéros supplémentaires.

Les résultats obtenus via la modélisation ZIP (B) et la modélisation GLM binomiale négative (C) sont très proches de ceux de la modélisation GLM Poisson. De plus, il est possible de comparer les trois modèles avec les critères AIC (Akaike's Information Criterion) et BIC (Bayesian Information Criterion). Ces critères permettent de comparer les différents modèles en tenant compte à la fois de leur qualité d'ajustement aux données et de leur complexité. En général, un modèle est considéré comme meilleur si son AIC ou son BIC est plus petit que celui des autres modèles. Il se retrouve les AIC et BIC suivants (tableau 2.7).

| | GLM Poisson | | GLM Binomiale négative | | ZIP | |
|---|---|---|---|---|---|---|
| | Accident | Maladie | Accident | Maladie | Accident | Maladie |
| AIC | 148999 | 211450 | 206524 | 210847 | 147787 | 209717 |
| BIC | 149052 | 211503 | 206591 | 210914 | 147895 | 209825 |

TABLE 2.7 : AIC et BIC des trois modèles pour chaque couverture

Les trois modèles ont pour la couverture maladie des AIC et BIC très proches. Cependant pour la couverture accident, le GLM Poisson et le modèle ZIP ont des AIC et BIC semblables et inférieurs à ceux du GLM de distribution binomiale négative. Ces deux modèles sont, selon ces critères, les plus

optimaux.

Au vu de ces différents résultats, la modélisation GLM est gardée par la suite.

Maintenant que les choix de modélisation sont terminés, il est possible de passer à la validation de la performance et du pouvoir prédictif des GLMs. Cela commence par le test anova() du package `stats` du logiciel R Core Team (2022) en spécifiant le test du $\chi^2$.

### Test anova

La fonction anova() permet de quantifier l'apport de chaque variable dans le modèle. En effet elle mesure la diminution de la déviance par l'ajout de chaque variable explicative. Grâce au GLM, il est possible de communiquer facilement sur l'impact de chaque variable explicative sur la fréquence, il est donc préféré un modèle plus simple que trop complexe. Ce test anova permet de se rapprocher du meilleur compromis entre le biais (écart entre la moyenne réelle et moyenne théorique) et la variance (complexité du modèle).

Pour le GLM maladie, la variable explicative produit a deux degrés de liberté (Personal Loan et Mortgage Loan, Car Loan étant l'intercept) pour une déviance égale à 547.76 (tableau 2.8). Il convient de faire un test du $\chi^2$ pour regarder l'importance explicative de cette variable et il s'obtient une p-value inférieure à 2.2e-16, donc la variable explicative produit réduit assez la déviance pour être significative dans le GLM. De même pour la variable explicative âge, de déviance égale à 2 339.30, le test du $\chi^2$ donne une p-value inférieure à 2.2e-16, d'où la significativité de cette variable explicative.

| Variable explicative | Degré de liberté | Déviance | Degré de liberté résiduel | Déviance résiduelle | p-value |
|---|---|---|---|---|---|
| Produit | 2 | 547,76 | 5 059 523 | 125 910 | <2e-16 |
| Âge | 1 | 2 339,3 | 5 059 522 | 124 924 | <2e-16 |

Table 2.8 : Résultat de la fonction anova() du logiciel R Core Team (2022) avec test du $\chi^2$ pour analyser le pouvoir explicatif de chaque variable dans le GLM maladie

Pour le GLM accident, la variable explicative produit a deux degrés de liberté (les produits Personal Loan et Mortgage Loan, Car Loan étant l'intercept) pour une déviance égale à 560.17 (tableau 2.9). Il convient de faire un test du $\chi^2$ pour regarder l'importance explicative de cette variable et il s'obtient une p-value inférieure à 2.2e-16, donc la variable explicative produit réduit suffisamment la déviance pour être significative dans le GLM accident. De même pour la variable explicative genre, qui a une déviance égale à 985.19, le test du $\chi^2$ donne une p-value inférieure à 2.2e-16, d'où la significativité de cette variable explicative.

| Variable explicative | Degré de liberté | Déviance | Degré de liberté résiduel | Déviance résiduelle | p-value |
|---|---|---|---|---|---|
| Produit | 2 | 560,17 | 5 050 501 | 125 910 | <2e-16 |
| Genre | 1 | 985,19 | 5 050 500 | 124 924 | <2e-16 |

Table 2.9 : Résultat de la fonction anova() du logiciel R Core Team (2022) avec test du $\chi^2$ pour analyser le pouvoir explicatif de chaque variable dans le GLM accident

**Dans les deux GLMs, les variables explicatives choisies réduisent toutes suffisamment la déviance pour être considérées comme nécessaires malgré l'ajout de complexité aux modèles qu'elles induisent**.

Procédons à présent à une validation croisée pour chaque couverture afin de vérifier qu'il n'y a pas de sur-aprentissage dans les modèles. Elle est séparée en deux étapes : d'abord une vérification visuelle puis le calcul de l'EDR.

### La validation croisée

La validation croisée est une méthode statistique permettant d'évaluer les performances prédictives d'un modèle et vérifier que celui-ci ne sur-apprend pas les données. Il y a du sur-apprentissage lorsque le modèle prend trop de paramètres en considération ce qui réduit la généralisation du modèle et son application à de nouvelles données. En effet quand un modèle est entraîné sur certaines données, il faut s'attendre à ce qu'il fonctionne sur de nouvelles données. La validation croisée permet de vérifier cette hypothèse. Dans cette étude, la méthode *Train-Test Split* est utilisée afin de vérifier la robustesse des variables explicatives. La première validation croisée est celle du GLM maladie.

Pour commencer, le jeu de données maladie est séparé en deux groupes, un *"Train set"* qui comprend 80% des données et un *"Validation set"* qui comprend les 20% de données restantes.

Dans un premier temps, il convient de regarder graphiquement si les deux échantillons ont une tendance et des valeurs proches par âge (figure 2.3) avec en noir la fréquence totale, en vert pointillé la fréquence du *"Train set"* et en bleu pointillé la fréquence du *"Validation set"*. C'est bien le cas.

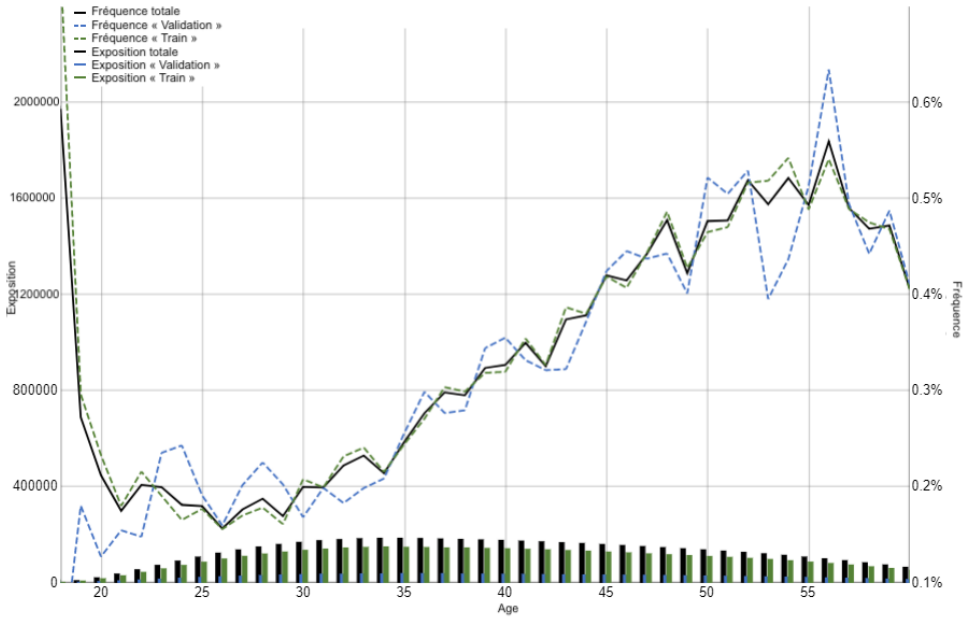



FIGURE 2.3 : Fréquence *"Train"* vs Fréquence *"Validation"* vs Fréquence totale en fonction de l'âge des données maladie

L'étape suivante est la vérification visuelle. Le GLM est entraîné sur le jeu de données *"Train set"* puis est appliqué sur le *"Validation set"*. Il est possible de comparer graphiquement la fréquence observée du *"Validation set"* par rapport à la fréquence prédite sur ce même jeu de données (figure 2.4) avec en rouge la prédiction et en bleu l'observation.

Cela confirme visuellement que la prédiction est alignée avec l'observation par âge et par conséquent que les variables explicatives âge et produit sont robustes dans le GLM maladie pour prédire la

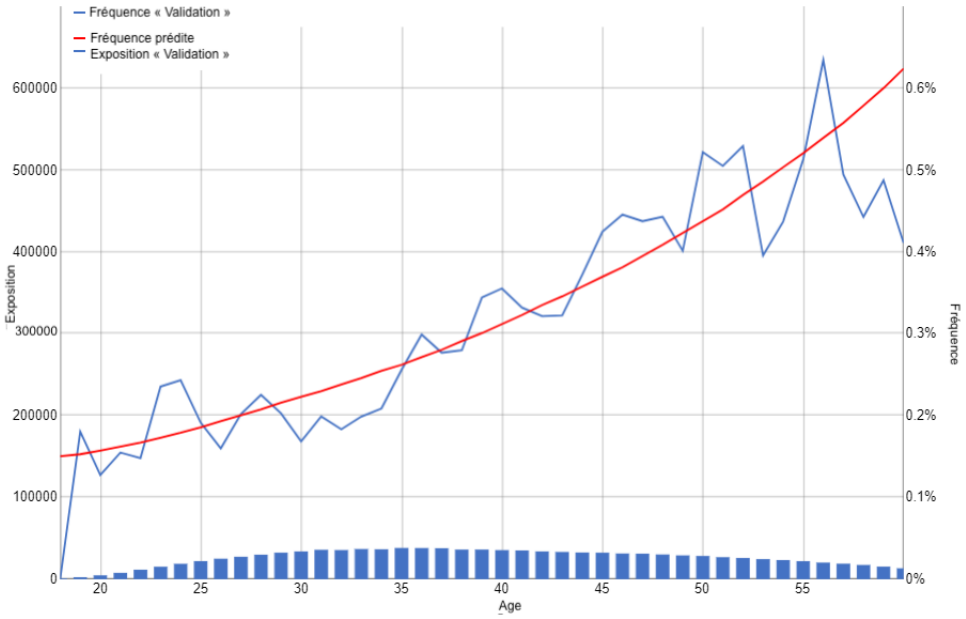



FIGURE 2.4 : Fréquence observée du *"Validation set"* par rapport à la prédiction du GLM maladie selon l'âge

fréquence par âge.

Enfin il est nécessaire de faire la comparaison des EDR (*Explained Deviance Ratio*). L'EDR se calcule avec la formule $\text{EDR} = 1 - \frac{\text{Deviance(Modele choisi)}}{\text{Deviance(Modele nul)}}$.

C'est un critère largement utilisé en interne chez AXA Partners car il présente l'avantage d'être un critère de déviance normalisé entre 0 et 1 (0 pouvoir explicatif nul, 1 prédiction parfaite).

Il faut cependant faire attention à une chose, son niveau dépend de la granularité des données. Plus les données vont être agrégées, et plus le niveau va se rapprocher de 1. Il est donc utilisé ici en comparant l'EDR de toutes les données prédites par le GLM calibré sur les données *Train set* avec l'EDR de toute les données prédites par le GLM calibré sur les données *Validation set*. **Si les deux EDR sont proches, alors les variables explicatives sont robustes et il n'y a pas de sur-apprentissage.**

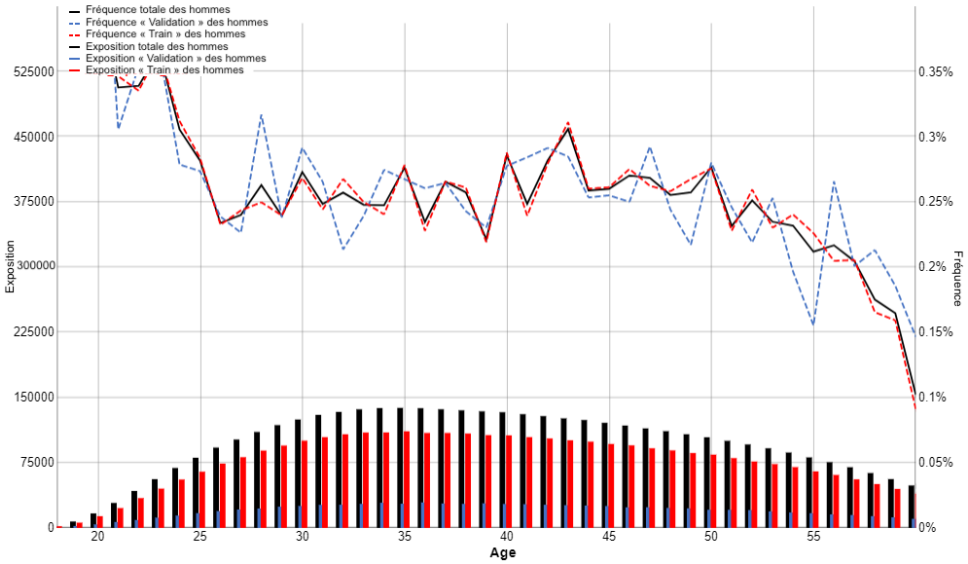
Les valeurs obtenues sont :

- EDR des données prédites sur le *"Train set"*$= 1 - \frac{175941.7}{178828.4} = 1.614\%$ ;
- EDR des données prédites sur le *"Validation set"* $= 1 - \frac{175947.8}{178829.1} = 1.611\%$.

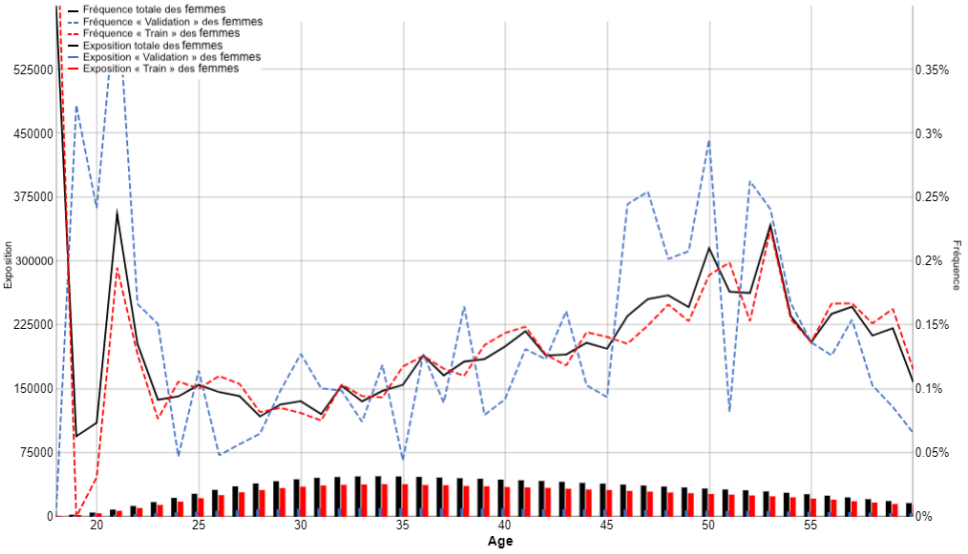
Les deux EDR étant très proches, la robustesse des deux variables explicatives âge et produit est confirmée, et le modèle ne sur-apprend pas les données.

La méthodologie de la validation croisée reste la même pour le GLM accident. Il convient de regarder dans un premier temps l'évolution de la fréquence des *"Train set"* et *"Validation set"* par genre en fonction de l'âge (figure 2.5) avec en noir la fréquence totale, en bleu pointillé le *"Validation set"* et en rouge pointillé le *"Train set"* dans le graphique du haut pour les hommes et dans le graphique du bas pour les femmes. Il se voit que pour les deux genres, les deux jeux de données proposent des

fréquences proches pour tous les âges (malgré quelques fluctuations observées chez les femmes entre 45 et 50 ans).




(a) Fréquence *"Train"* vs Fréquence *"Validation"* vs Fréquence totale en fonction de l'âge des données accident pour les hommes

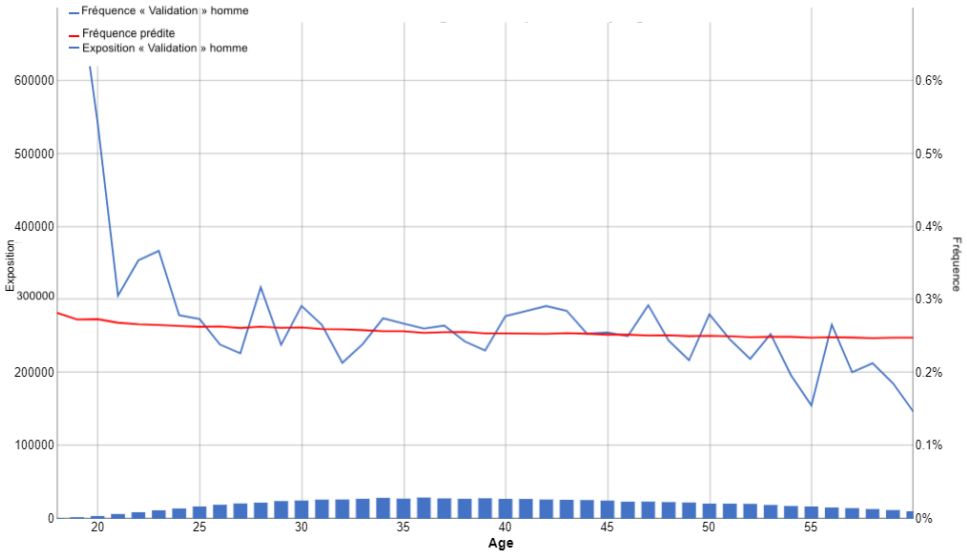



(b) Fréquence *"Train"* vs Fréquence *"Validation"* vs Fréquence totale en fonction de l'âge des données accident pour les femmes

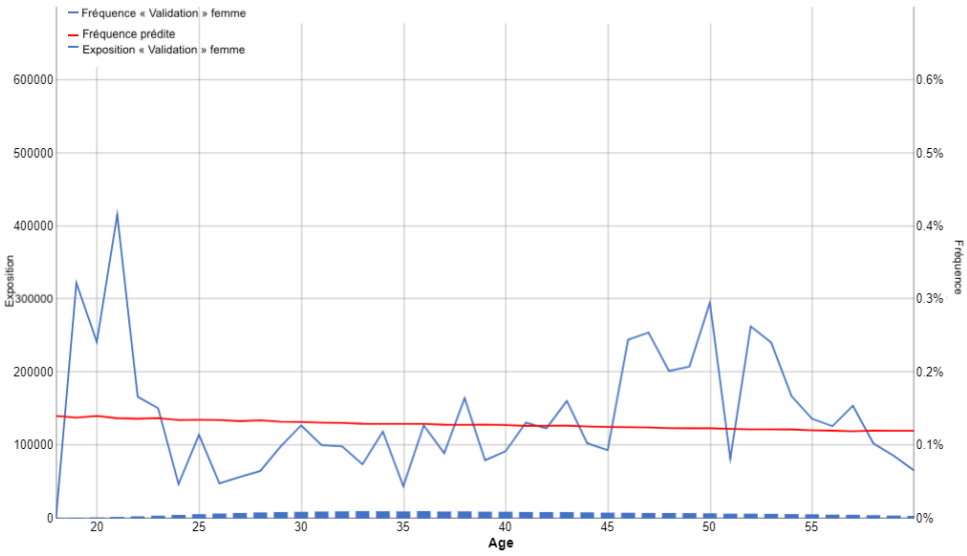
FIGURE 2.5 : Fréquence *"Train"* vs Fréquence *"Validation"* vs Fréquence totale en fonction de l'âge des données accident

L'étape suivante est la vérification visuelle. Le GLM est entraîné sur le jeu de données *"Train set"* puis est appliqué sur le *"Validation set"*. Il est possible de comparer graphiquement la fréquence observée du *"Validation set"* par rapport à la fréquence prédite sur ce même jeu de données pour les hommes en haut et pour les femmes en bas (figure 2.6) avec en rouge la fréquence prédite et en bleu la fréquence observée.

Il est confirmé visuellement que les prédictions pour chaque genre sont alignées avec les observations




(a) Fréquence observée *"Validation"* par rapport à la prédiction du GLM accident pour les hommes




(b) Fréquence observée *"Validation"* par rapport à la prédiction du GLM accident pour les femmes

FIGURE 2.6 : Fréquence observée du *"Validation set"* par rapport à la prédiction du GLM accident

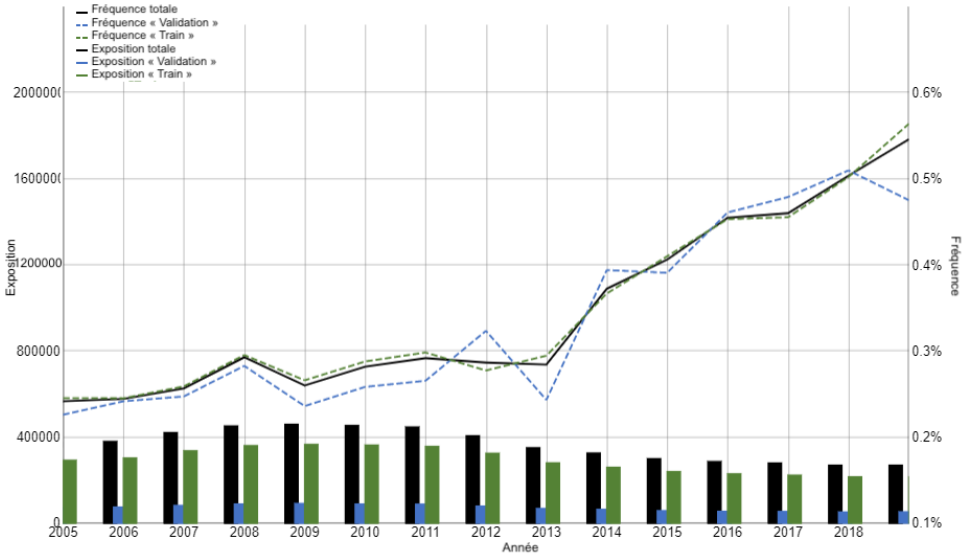
par âge et par conséquent que les variables explicatives genre et produit sont robustes dans le GLM accident pour prédire la fréquence par âge.

Enfin la comparaison des EDR (*Explained Deviance Ratio*) est faite. L'EDR de toutes les données prédites par le modèle calibré sur les données *"Train set"* et l'EDR de toutes les données prédites par le modèle calibré sur les données *"Validation set"* sont calculés. Il est vérifié si les deux EDR sont proches :

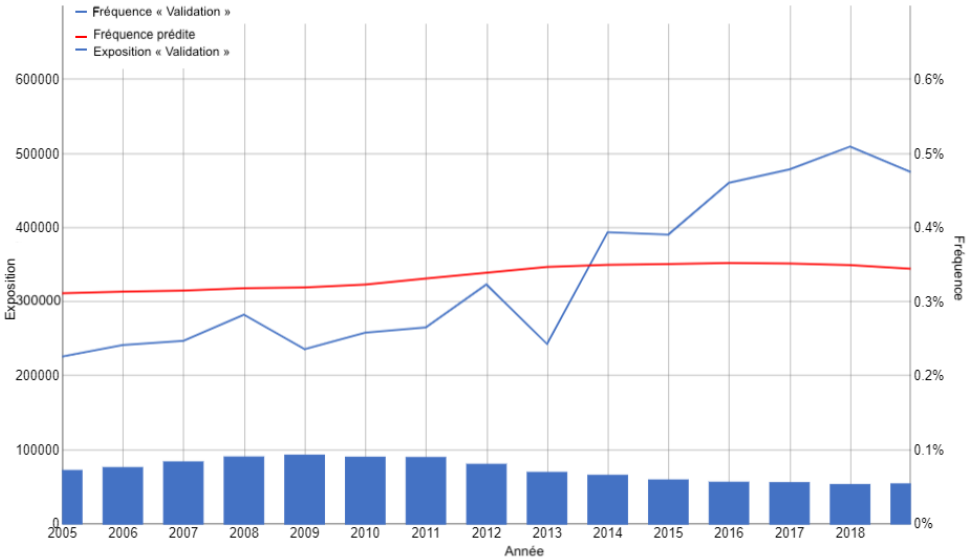
- EDR des données prédites sur le *"Train set"*$= 1 - \frac{124925.2}{124938.2} = 1.222\%$ ;
- EDR des données prédites sur le *"Validation set"* $= 1 - \frac{124938.2}{126475} = 1.215\%$.

Les deux EDR étant très proches, la robustesse des deux variables explicatives genre et produit est confirmée, et le modèle ne sur-apprend pas les données.

Cependant, si la validation croisée est regardée à présent par année d'accident et non par âge, il est remarqué que le GLM maladie (il en est de même pour le GLM accident) ne capte pas la croissance avec les années (figure 2.7, avec en haut la fréquence *"Train"* en vert vs fréquence *"Validation"* en bleu vs fréquence totale en noir en fonction de l'année des données maladie et en bas la fréquence observée du *"Validation set"* en bleu par rapport à la prédiction du GLM maladie en rouge).




(a) Fréquence *"Train"* vs Fréquence *"Validation"* vs Fréquence totale par année des données maladie




(b) Fréquence observée du *"Validation set"* par rapport à la prédiction du GLM maladie

FIGURE 2.7 : Graphiques de la validation croisée par année d'incident pour le GLM maladie

Le même phénomène est observé pour la couverture accident. **Les deux GLMs ne prédisent pas**

**la croissance observée de la fréquence avec les années d'incident.** Les contraintes internes qui s'appliquent sur la modélisation empêchent d'ajouter des variables explicatives supplémentaires qui pourraient expliquer cette croissance.

L'étude d'un produit en particulier, le Personal Loan, va à nouveau le montrer (A pour les produits Mortgage Loan et Car Loan, le Car Loan ne présente lui pas de croissance de la fréquence avec les années). Il est implémenté en conséquent une méthodologie pour ce produit (cette méthodologie est commune aux produits Mortgage Loan et Car Loan).

### 2.2.3 Choix du modèle final pour le produit Personal Loan

Dans un premier temps, les résultats obtenus des GLMs pour le produit Personal Loan sont analysés.

#### Les résultats par âge et genre en distinguant maladie et accident

Cela va permettre de confirmer ou non visuellement la prédiction des GLMs par âge et genre (selon la couverture choisie).

Pour la maladie, il convient de regarder les fréquences des hommes et des femmes selon l'âge et la prédiction du GLM (figure 2.8) avec en rouge pointillé les fréquences observées des hommes, en bleu pointillé les fréquences observées des femmes et en vert la fréquence prédite par le GLM maladie.

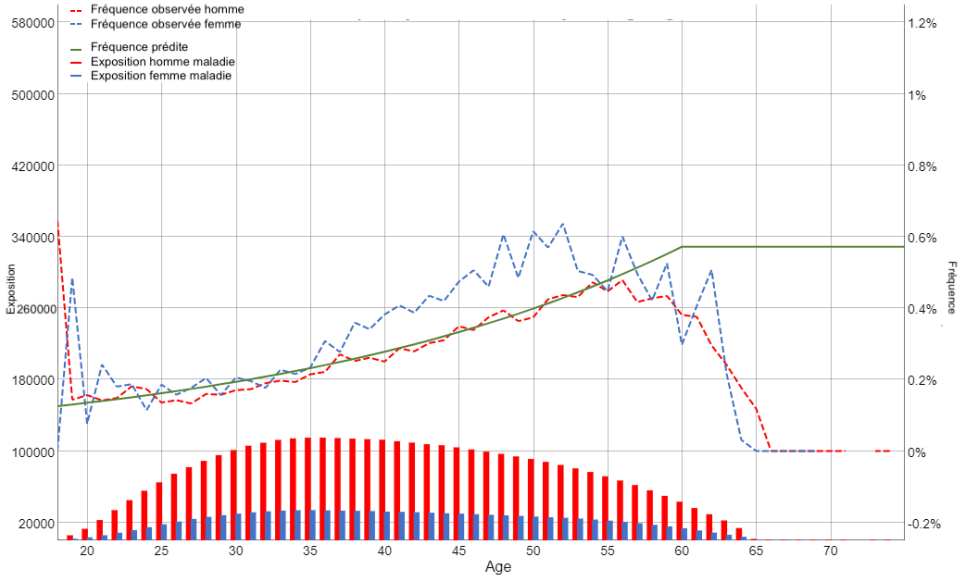



FIGURE 2.8 : Fréquences observées par genre vs Fréquence prédite pour le produit Personal Loan en couverture maladie

L'âge étant la seule variable explicative à l'échelle d'un seul produit en couverture maladie, il se retrouve une prédiction croissante avec l'âge, alignée avec les observations. Cette prédiction est beaucoup plus proche des observations masculines compte-tenu de l'exposition des hommes qui est largement supérieure à celle des femmes (barres rouges des hommes plus grandes que celles bleues des femmes).

Il se remarque également qu'entre 37 et 55 ans, les fréquences observées sont sensiblement différentes

selon le genre, ce qui n'est pas pris en compte par le modèle. La courbe prédite est forcée à stagner à partir de 60 ans car AXA Partners ne veut pas prendre en compte la baisse de fréquence liée aux vieux âges et la faible exposition disponible pour ces derniers.

Pour l'accident, les fréquences des hommes et des femmes selon l'âge et la prédiction du GLM (figure 2.9) sont présentées avec en rouge pointillé les fréquences observées des hommes, en bleu pointillé les fréquences observées des femmes et en lignes pleines rouges et bleues les fréquences prédites par le GLM accident pour les deux genres.

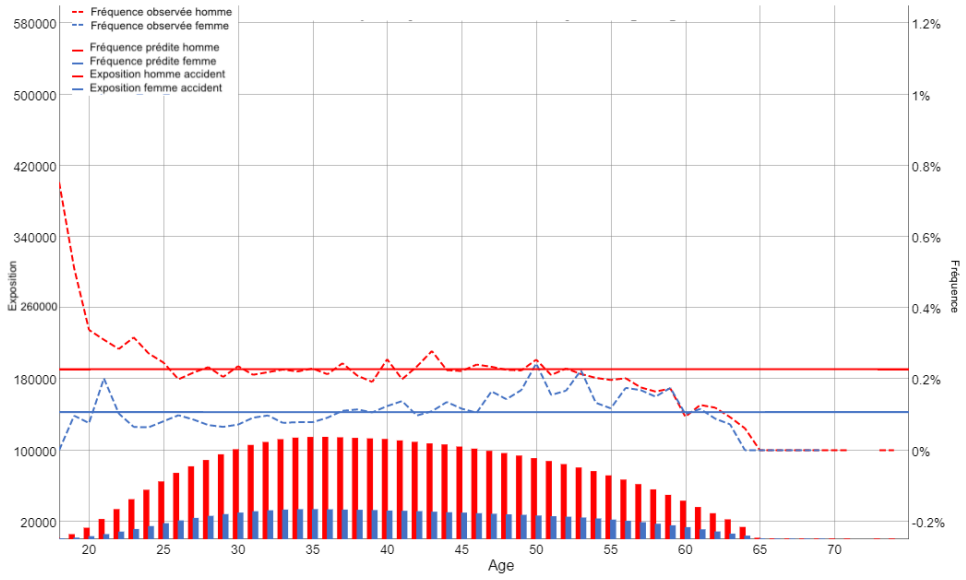



FIGURE 2.9 : Fréquences observées vs Fréquences prédites par genre et par âge pour le produit Personal Loan en couverture accident

Le genre étant la seule variable explicative à l'échelle d'un seul produit en couverture accident, il est logique de retrouver une prédiction horizontale (ne dépendant pas de l'âge) alignée avec les observations masculines et féminines. Comme pour la couverture maladie, l'exposition majoritaire est celle des hommes.

Cependant il se remarque que les courbes des observations, celle des hommes comme celle des femmes, ne sont pas complètement stagnantes selon les âges. Notamment chez les hommes s'observe une fréquence très élevée entre 18 et 23 ans et cette fréquence décroit légèrement avec l'âge à partir de 50 ans.

Les résultats obtenus des deux GLMs sont maintenant réunis pour les comparer à la précédente modélisation GLM, et voir les changements qui sont survenus.

#### Les résultats par âge et genre, les deux couvertures confondues

Tout d'abord, les résultats des fréquences par âge sont analysés en séparant les deux genres et en regroupant maladie et accident.

Pour les hommes, la fréquence prédite par les nouveaux GLMs (courbe rouge pleine) selon l'âge est comparée à la fréquence observée (courbe rouge pointillée) et à la fréquence du précédent modèle

validé en 2018 (courbe noire) (figure 2.10).

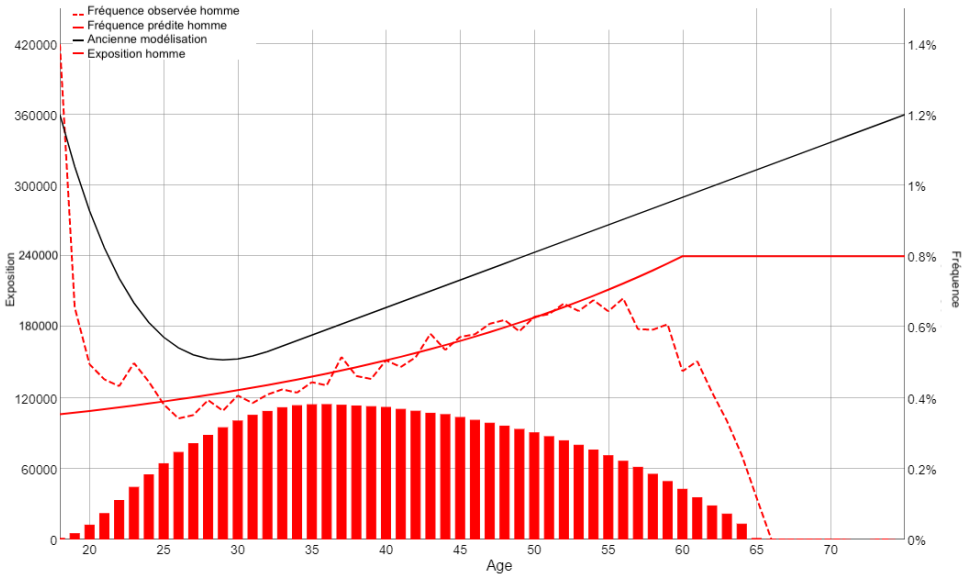



FIGURE 2.10 : Fréquence observée, fréquence prédite et ancienne fréquence modélisée par âge pour les hommes du produit Personal Loan, accident et maladie ensemble

En réunissant les GLMs maladie et accident, la courbe de fréquence prédite est complètement alignée avec les observations par âge pour les hommes sauf avant l'âge de 25 ans et après 55 ans. cette prédiction est en dessous de la précédente modélisation, et ne prend plus en compte la fréquence élevée pour les hommes d'un âge compris entre 18 et 23 ans. La nouvelle modélisation perd donc en précision sur cet intervalle mais reste plus précise pour le reste des âges.

Pour les femmes, la fréquence prédite par les nouveaux GLMs (courbe bleue pleine) selon l'âge est comparée à la fréquence observée (courbe bleue pointillée) et à la fréquence du précédent modèle validé en 2018 (courbe noire) (figure 2.11).

En réunissant les GLMs maladie et accident, la courbe de fréquence prédite obtenue n'est pas complètement alignée avec les observations par âge pour les femmes. Elle se trouve en dessous de la précédente modélisation, avec une croissance liée à l'âge de la fréquence prédite moins prononcée que les observations réelles, notamment entre 35 et 55 ans. Les fluctuations entre 18 et 23 ans ne sont également pas prises en compte. Cette modélisation manque donc aussi de précision pour les femmes, comme l'ancienne modélisation.

Finalement, il convient de regarder la fréquence prédite obtenue, accident et maladie ensemble, hommes et femmes ensemble, selon l'année d'incident qui est comparée avec la fréquence observée et la précédente modélisation.

### Les résultats par année d'incident

La fréquence prédite (courbe verte) est comparée à la fréquence observée (courbe bleue) et à la précédente fréquence modélisée (courbe noire) en fonction de l'année d'incident, sans distinction de genre et de couverture (figure 2.12).

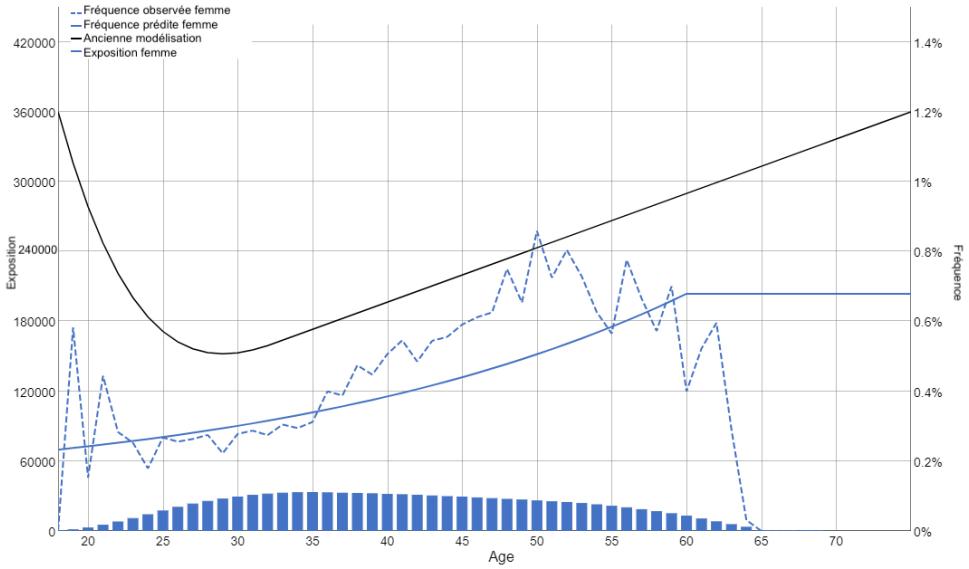



FIGURE 2.11 : Fréquence observée, fréquence prédite et ancienne fréquence modélisée par âge pour les femmes du produit Personal Loan, accident et maladie ensemble

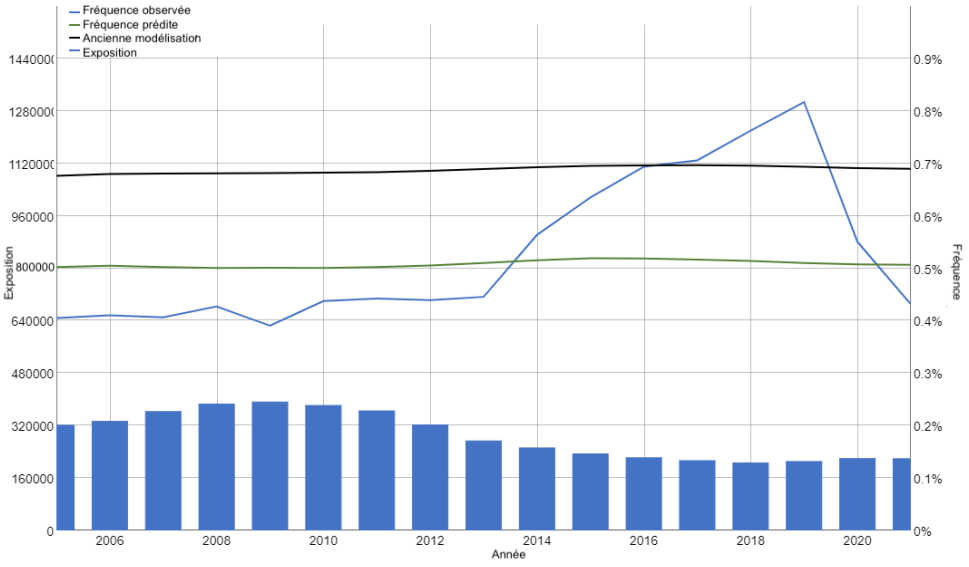



FIGURE 2.12 : Fréquence observée, fréquence prédite et ancienne fréquence modélisée tous genres et toutes couvertures confondus pour le produit Personal Loan

Il se remarque une croissance globale de la fréquence avec les années atteignant un pic en 2019 (courbe bleue). Cette croissance est due à l'augmentation de la conscience que les Portugais ont des couvertures d'assurance dont ils disposent. La baisse après 2019 correspond à la période de covid et n'est pas prise en compte dans la modélisation. Le GLM ne modélise absolument pas cette croissance et stagne aux alentours de 0.5% (courbe verte). Enfin la modélisation est bien inférieure à la fréquence de la précédente modélisation (courbe noire), qui elle non plus ne prend pas en compte la hausse observée.

**Ainsi bien qu'ils soient relativement pertinents pour modéliser les fréquences par genre et âge pour les deux types de couverture, les deux GLMs ne modélisent absolument pas la croissance avec les années de la fréquence pour le produit Personal Loan**, comme le présageait les résultats de la validation croisée (2.2.2). Il en va de même pour l'autre produit "non Bulk" Mortgage Loan (A).

En outre, l'utilisation d'une distribution ZIP à la place d'une distribution de Poisson n'apporte pas de changement à la prédiction (B). La régression ZIP se fait via la fonction zeroinfl() du package `pscl` du logiciel R Core Team (2022). Il en va de même avec un GLM de distribution binomiale négative, les fréquences et les variations obtenues sont sensiblement les mêmes (C).

Sous les contraintes internes qui empêchent l'ajout de variables explicatives supplémentaires aux GLMs (1.6), il faut procéder à la méthodologie suivante pour prendre en compte la hausse de la fréquence observée. Il est nécessaire d'ajouter du poids au coefficient du produit Personal Loan dans les deux GLMs afin que la prédiction s'aligne sur le pic de fréquence de l'année 2019 (figure 2.12). Cette méthodologie est également appliquée aux autres produits "non Bulk" (D).

#### Méthodologie d'ajustement du coefficient du produit Personal Loan

Il est souhaité d'augmenter le coefficient du produit Personal Loan dans les GLMs afin d'avoir une fréquence prédite alignée avec l'observation de 2019, soit une fréquence accident et maladie ensemble de 0.82%. Cela donne :

- pour le GLM maladie un coefficient pour la variable "Personal Loan" de -0.0623 (contre -0.4930 précédemment) ;
- pour le GLM accident un coefficient pour la variable "Personal Loan" de -0.1899 (contre -0.6206 précédemment).

Il est possible de comparer à nouveau la fréquence prédite avec les observations par âge selon le genre et la couverture (figure D.1 en annexe) ainsi que la nouvelle fréquence prédite, accident et maladie ensemble, homme et femme ensemble, par rapport à la fréquence observée par année (figure D.2 en annexe). Sur ce dernier graphique, il se remarque que la fréquence prédite par les GLMs est bien alignée avec l'année 2019, mais ne prend toujours pas en compte la croissance observée avec les années.

D'un point de vue interne à la compagnie, ces changements sont parfaitement acceptés, à défaut de pouvoir ajouter des variables explicatives supplémentaires.

Il faut procéder à des modifications similaires pour les deux autres produits Mortgage Loan et Car Loan et il s'obtient les coefficients suivants pour les deux GLMs (tableaux 2.10).

Il se remarque d'ailleurs que pour les deux GLMs, toutes les variables explicatives sont encore considérées comme statistiquement significatives avec des p-values $< \alpha = 0.05$.

Regardons à présent l'intégration des produits "Bulk" aux équations des GLMs.

### 2.2.4 Ajout des produits "Bulk" aux GLMs

**Il est nécessaire de rappeler que les produits "Bulk" n'avaient pas pu être introduits à cause d'informations détaillées sur l'âge et le genre manquantes dans les données (1.5.1).**

| Nom de la variable explicative | Coefficient associé | p-value |
|---|---|---|
| Produit Car Loan (et intercept) | -6,7795714 | <2e-16 |
| Produit Mortgage Loan | 0,3776003 | <2e-16 |
| Produit Personal Loan | -0,0623415 | 0,00424 |
| Âge | 0,0360684 | <2e-16 |

(a) Nouveaux coefficients pour le GLM maladie

| Nom de la variable explicative | Coefficient associé | p-value |
|---|---|---|
| Produit Car Loan (et intercept) | -5,40823 | <2e-16 |
| Produit Mortgage Loan | 0,26024 | 2,55e-13 |
| Produit Personal Loan | -0,18992 | 4,47e-14 |
| Genre femme | -0,7547 | <2e-16 |

(b) Nouveaux coefficients pour le GLM accident

TABLE 2.10 : Nouveaux coefficients des deux GLMs

Dans la précédente modélisation faite par AXA Partners, ces produits n'étaient pas intégrés au GLM, il y avait une fréquence commune pour tous les genres et tous les âges pour chaque produit "Bulk" (1.6). La méthodologie d'ajout de ces produits aux GLMs est détaillées en annexe (F), seuls les résultats finaux sont présentés ici.

Tous les produits TTD vendus au Portugal sont maintenant des variables explicatives de la fréquence d'incident dans les deux GLMs maladie et accident. Les fréquences finales sont :

- pour la maladie, $Frequence = \exp(-6.780 + 0.036 \times Age - 0.062 \times \mathbf{1}_{\text{Personal Loan}} + 0.378 \times \mathbf{1}_{\text{Mortgage Loan}} - 0.684 \times \mathbf{1}_{\text{Credit Card}} - 1.206 \times \mathbf{1}_{\text{Waiver of Premium}} + 0.875 \times \mathbf{1}_{\text{Income Protection}})$ ;
- pour l'accident, $Frequence = \exp(-5.408 - 0.755 \times \mathbf{1}_{\text{Genre féminin}} - 0.190 \times \mathbf{1}_{\text{Personal Loan}} + 0.260 \times \mathbf{1}_{\text{Mortgage Loan}} - 1.174 \times \mathbf{1}_{\text{Credit Card}} - 1.807 \times \mathbf{1}_{\text{Waiver of Premium}} + 0.154 \times \mathbf{1}_{\text{Income Protection}})$.

Avec l'intercept (-6.780 et -5.408) qui correspond au coefficient du produit Car Loan.

Un exemple de calcul de la prime pure (unique et périodique annuelle) est maintenant introduit, qui est repris dans les deux modélisations suivantes pour comparer les modèles d'un point de vue tarification.

## 2.3 Calcul de la prime pure pour le produit Personal Loan

Le cadre de l'exemple est le suivant :

- le produit vendu est le Personal Loan ;
- les caractéristiques du portefeuille sont 1000 assurés pour un âge moyen de 43 ans et une proportion de 75% d'hommes (F.1) ;
- les caractéristiques du prêt sont un prêt de 20000€ sur 5 ans (de 2023 à 2027) avec taux d'intérêt annuel de 5% soit un remboursement chaque début de mois de 374.88€ (un taux d'intérêt annuel de 5% implique un taux mensuel de 0,41%). Il faut procéder à 60 versements M d'où l'égalité $20000 = \sum_{i=0}^{59} \frac{M}{(1+0.41\%)^i} \Rightarrow M = 374,88€$ (car les mensualités sont actualisées) ;
- taux actuariel au 31/08/2022 : 0.75% ;

- durée moyenne d'incapacité pour le produit Personal Loan de 6 mois ;
- comme hypothèses supplémentaires, il est supposé que les 1000 assurés ont un contrat de 1 ans (d'où une exposition totale de 1000 par année) et qu'il n'y a ni décès, ni rachat.

Il est souhaité ici de calculer la prime pure unique et annuelle. L'exposition reste constante sur les 5 années, à savoir 1000. La fréquence est également constante avec les années et il s'obtient pour le produit Personal Loan, avec une moyenne d'âge de 43 ans et 75% d'hommes une fréquence de 0.824% (maladie et accident ensemble) pour chaque année soit un nombre annuel de sinistres de 8.24. Voici un tableau récapitulatif (tableau 2.11).

| Année | 2023 | 2024 | 2025 | 2026 | 2027 |
|---|---|---|---|---|---|
| Exposition | 1000 | 1000 | 1000 | 1000 | 1000 |
| Fréquence | 0,824% | 0,824% | 0,824% | 0,824% | 0,824% |
| Nombre de sinistres | 8,2 | 8,2 | 8,2 | 8,2 | 8,2 |

TABLE 2.11 : Récapitulatif de l'exposition, de la fréquence et du nombre de sinistres par année

Le montant du prêt est de 20000€, à rembourser sur 5 ans avec un taux d'intérêt de 5% soit un remboursement de 374.88€ par mois. Avec une durée moyenne d'incapacité de 6 mois, le coût total que l'assureur doit supporter par an est égal à Nombre de sinistres × Remboursement mensuel du prêt × Durée de l'incapacité $= 8.2 \times 374.88 \times 6 = 18525.02$€. Avec 1000 assurés, cela donne un coût de 18.53€ par assurés tous les ans.

Le calcul de la valeur actualisée de ces coûts annuels donne $VA = \sum_{i=0}^{4} \frac{18.53}{(1+0.75\%)^i} = 91.26$€. Ces coûts étant constants sur les 5 années, il en est déduit directement la prime pure unique et la prime pure périodique. En effet prime pure unique = VA, prime pure périodique annuelle = coût par assuré tous les ans (tableau 2.12).

| Prime pure unique | Prime pure annuelle |
|---|---|
| 91,26 € | 18,53 € |

TABLE 2.12 : Prime pure unique et prime pure annuelle

La fréquence utilisée ici est celle obtenue après les ajustements faits sur les GLMs pour être aligné avec le pic de fréquence de 2019. Avec le modèle GLM sans ajustements, la prime pure unique s'élève à 56.51€ et la prime pure annuelle à 11.47€.

### Conclusion du chapitre 2

Il a pu être vu dans ce chapitre la méthodologie mise en place pour calibrer les GLMs maladie et accident de la couverture TTD au Portugal à partir de la base de données créée et validée dans le chapitre 1. La pertinence et la robustesse du choix des variables explicatives ont également été vérifiées, après avoir validé le choix d'une distribution de Poisson (malgré des limites observées). Cependant il a été observé des hausses de fréquences avec les années d'incident non prises en compte par cette nouvelle modélisation. L'ajout de variables explicatives n'étant pas possible à cause des contraintes internes qui s'appliquent sur cette modélisation, il a été mis en place une méthodologie pour s'aligner sur la fréquence observée la plus élevée. Enfin les produits "Bulk" ont été ajoutés aux GLMs, la calibration

des GLMs étant impossible avec ces produits, et il a été calculé dans un exemple les primes pures unique et annuelle associées aux équations de fréquence pour le produit Personal Loan.

Que se passe-il maintenant si les contraintes imposées par AXA Partners qui encadrent la modélisation GLM de ce chapitre 2 sont retirées ? L'ajout de nouvelles variables explicatives au GLM permet-t-il une vraie amélioration du modèle sans trop le complexifier ? Qu'en est-il de proposer une modélisation complètement différente via un *boosting* ?

# Chapitre 3

# La nouvelle modélisation GLM et la modélisation Machine Learning Boosting

Maintenant que les contraintes internes sont retirées, il est intéressant de regarder si l'ajout de nouvelles variables explicatives à la modélisation GLM va permettre de mieux prendre en compte les hausses de la fréquence avec les années précédemment observées et non modélisées. Une fois ces variables étudiées et ajoutées aux données, les GLMs sont à nouveau calibrés, la robustesse des variables explicatives est vérifiée. Puis il convient de procéder à la modélisation *boosting* pour la comparer aux deux modélisations GLM. Cela permet de conclure sur le choix de la meilleure méthode pour modéliser les données incapacité au Portugal.

## 3.1 De nouvelles variables explicatives

Les variables explicatives choisies sont des variables macroéconomiques dont l'intérêt et un premier travail sur leur corrélation sont présentés.

### 3.1.1 Les variables macroéconomiques

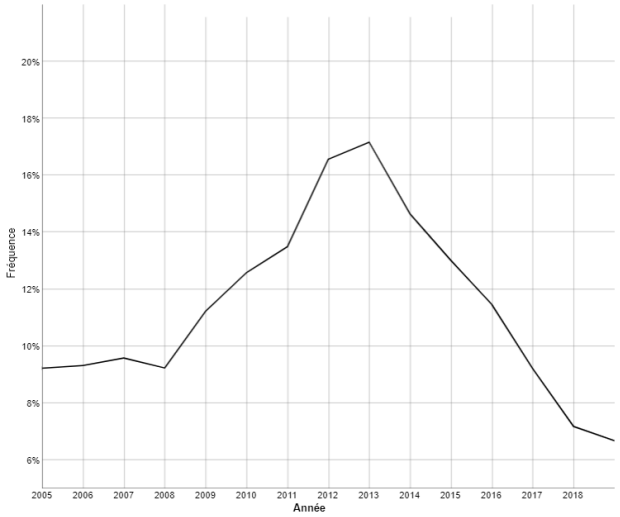
Il a pu être vu dans le chapitre 1 que la crise économique de 2013 a entrainé une augmentation de la conscience des Portugais pour les couvertures d'assurance qu'ils ont à disposition (1.4). Les Portugais ont davantage déclaré leurs sinistres en TTD et il a pu être observé une hausse de la fréquence générale, notamment pour les produits Personal Loan et Mortgage Loan.

C'est pourquoi des variables macroéconomiques sont ajoutées aux données :

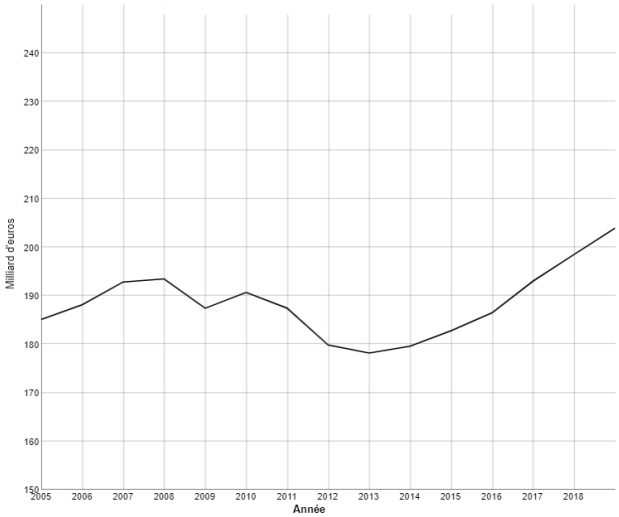
- le taux de chômage car il a pu être observé dans le chapitre 1 qu'il atteint un pic en 2013 avant de décroître jusqu'en 2019 (1.4). Ce pic marque le début de la conscience des Portugais pour leur couverture d'assurance. Cette variable est exprimée en % ;
- le PIB (Produit Intérieur Brut) qui mesure la production de richesse annuelle d'un pays, exprimé en milliards d'€. Le PIB est souvent fortement inversement corrélé au taux de chômage ;

- l'IPC (Indice des Prix à la Consommation) qui mesure l'évolution du niveau moyen des prix des biens et services consommés par les ménages, exprimé sur une base de 100 pour 2012. Il s'agit d'une mesure de l'inflation. Il augmente avec les années au Portugal. La hausse des prix pousse les Portugais à davantage chercher des moyens de compenser leur perte de pouvoir d'achat, d'où une augmentation des connaissances dans les produits d'assurance auxquels ils sont éligibles. Cela vient compléter l'explication de la crise économique de 2013.

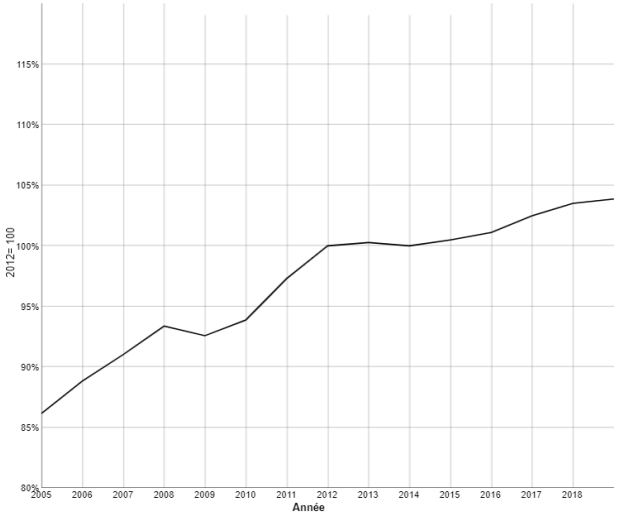
Ces données sont récupérées sur la période 2005-2019 et leur évolution est étudiée (figure 3.1).




(a) Evolution du taux de chômage au Portugal sur la période 2005-2019




(b) Evolution du PIB portugais sur la période 2005-2019




(c) Evolution de l'IPC au Portugal sur la période 2005-2019

FIGURE 3.1 : Evolution des trois variables macroéconomiques sur la période 2005-2019

Ces variables sont comparées dans un premier temps et leur corrélation est étudiée afin de ne pas ajouter de variables répétitives aux données.

### 3.1.2 Test de corrélation et variables retenues

Les fonctions cor(), cor.test() et corrplot() du package `stats` du logiciel R Core Team (2022) vont être utilisées pour calculer et tester les corrélations entre les trois variables.

Il convient de commencer par déterminer la matrice de corrélation via la fonction cor() qui par défaut calcule la corrélation linéaire de Pearson. Il est nécessaire de rappeler le calcul d'un estimateur du coefficient de corrélation d'un échantillon $\{(x_i, y_i), 1 \leq i \leq n\}$ de n réalisations indépendantes de variables aléatoires X et Y par la méthode de Pearson (Saporta (2006)),

$\hat{r}_p = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$.

Avec :

- $\hat{\sigma}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$, l'estimateur de la covariance ;
- $\hat{\sigma}_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, et $\hat{\sigma}_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$, les estimateurs des écarts-types ;
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, les espérances des variables X et Y.

Le coefficient de corrélation ainsi calculé est compris entre -1 et 1. 1 signifie une corrélation parfaite, -1 une corrélation inversement parfaite, et 0 l'absence totale de corrélation.

Il s'obtient la matrice de corrélation (tableau 3.1) et le corrélogramme (figure 3.2).

| | taux de chômage | PIB | IPC |
|---|---|---|---|
| taux de chômage | 1 | -0,8724179 | 0,1406703 |
| PIB | -0,8724179 | 1 | 0,1334429 |
| IPC | 0,1406703 | 0,1334429 | 1 |

Table 3.1 : Matrice de corrélation des variables macroéconomiques

A première vue, les variables taux de chômage et PIB ont l'air fortement inversement corrélées avec un coefficient de corrélation égal à -0.872. La variable IPC ne semble pas corrélée avec les deux autres puisque les coefficients de corrélations obtenus sont proches de 0.

Un test de significativité de la corrélation est nécessaire pour le vérifier (Saporta (2006)). Le test de Pearson est choisi (les tests de Kendall et Spearman sont également possibles) et est effectué via la fonction cor.test(). Le test est le suivant :

$H_0$ : "Pas de corrélation entre les deux variables" contre $H_1$ : "Corrélation entre les deux variables".

Les p-values ci-dessous sont obtenues (tableau 3.2).

Il s'obtient une p-value inférieure à 0.05 pour le test de la corrélation entre le taux de chômage et le PIB. L'hypothèse $H_0$ est rejetée et les deux variables sont considérées comme significativement corrélées. En revanche il s'obtient une p-value supérieure à 0.05 pour les autres tests de corrélation. Il en est déduit qu'il n'y a pas de corrélation entre le taux de chômage et l'IPC, et entre le PIB et l'IPC.

Finalement, seules les variables taux de chômage et IPC sont ajoutées à la base de données, la variable PIB étant fortement corrélée à celle du taux de chômage et donc superflue. Il est maintenant possible de calibrer les nouveaux GLMs. La validité du choix de la distribution de Poisson n'est pas à nouveau

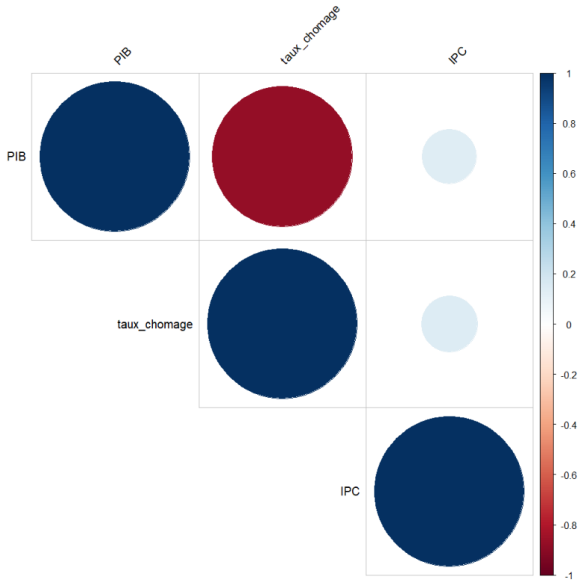



FIGURE 3.2 : Corrélogramme des variables macroéconomiques, les corrélations positives sont en bleu et les corrélations négatives en rouge ; plus le cercle est grand et opaque, plus il y a de corrélation

| Pearson | taux de chômage | PIB | IPC |
|---|---|---|---|
| taux de chômage | | 2.215e-05 | 0.617 |
| PIB | 2.215e-05 | | 0.6354 |
| IPC | 0.617 | 0.6354 | |

TABLE 3.2 : p-values des tests de Pearson

étudiée mais il est vérifié via une analyse de la déviance et une validation croisée que les GLMs ont des pouvoirs prédictifs forts.

## 3.2 Les nouveaux GLMs enrichis des variables macroéconomiques

### 3.2.1 Coefficients obtenus et test de leur significativité statistique

La séparation GLM maladie et GLM accident est gardée avec leurs précédentes variables explicatives, âge et produit pour le GLM maladie et genre et produit pour le GLM accident, et sont ajoutées aux deux GLMs les variables taux de chômage et IPC. Les coefficients et les p-values associées obtenues sont les suivants.

Le GLM maladie est étudié en premier (tableau 3.3 où le produit Car Loan correspond à l'intercept).

Il y a dans ce tableau les différents coefficients estimés via un maximum de vraisemblance pour chaque variable explicative et la dernière colonne correspond à la p-value du test statistique de la significativité de chacune des variables. Comme pour la précédente modélisation GLM, il convient de prendre un niveau significatif $\alpha = 0.05$ et si la p-value est inférieure à ce niveau, la variable explicative est considérée comme statistiquement significative.

| Nom de la variable explicative | Coefficient associé | p-value |
|---|---|---|
| Produit Car Loan (et intercept) | -10,1900 | <2e-16 |
| Produit Mortgage Loan | -0,11990 | 6,25e-05 |
| Produit Personal Loan | -0,50340 | <2e-16 |
| Âge | 0,03494 | <2e-16 |
| IPC | 0,04014 | <2e-16 |
| taux de chômage | -0,04103 | <2e-16 |

TABLE 3.3 : Coefficients et p-values obtenus pour le nouveau GLM de la couverture maladie

Il se trouve que la p-value est inférieure à $\alpha = 0.05$ pour toutes les variables explicatives, d'où leur pertinence explicative.

Pour le GLM accident, il s'agit du même procédé (tableau 3.4, où le produit Car Loan correspond à l'intercept).

| Nom de la variable explicative | Coefficient associé | p-value |
|---|---|---|
| Produit Car Loan (et intercept) | -7,928916 | <2e-16 |
| Produit Mortgage Loan | -0,200542 | 2,80e-16 |
| Produit Personal Loan | -0,630668 | <2e-16 |
| Genre femme | -0,765955 | <2e-16 |
| IPC | 0,029759 | <2e-16 |
| taux de chômage | -0,034337 | <2e-16 |

TABLE 3.4 : Coefficients et p-values obtenus pour le nouveau GLM de la couverture accident

Comme pour le GLM maladie, ce tableau donne les coefficients et les p-values associées pour chaque variable explicative. Il s'obtient des p-values inférieures à $\alpha = 0.05$ pour toutes les variables explicatives d'où leur significativité statistique dans la calibration du GLM accident.

Il convient d'analyser plus en profondeur la performance des deux modélisations et la robustesse des variables explicatives en étudiant les résidus, en regardant les résultats du test de déviance et en procédant à une validation croisée.

### 3.2.2 Etude des résidus des GLMs

#### Etude de la distribution Poisson des résidus

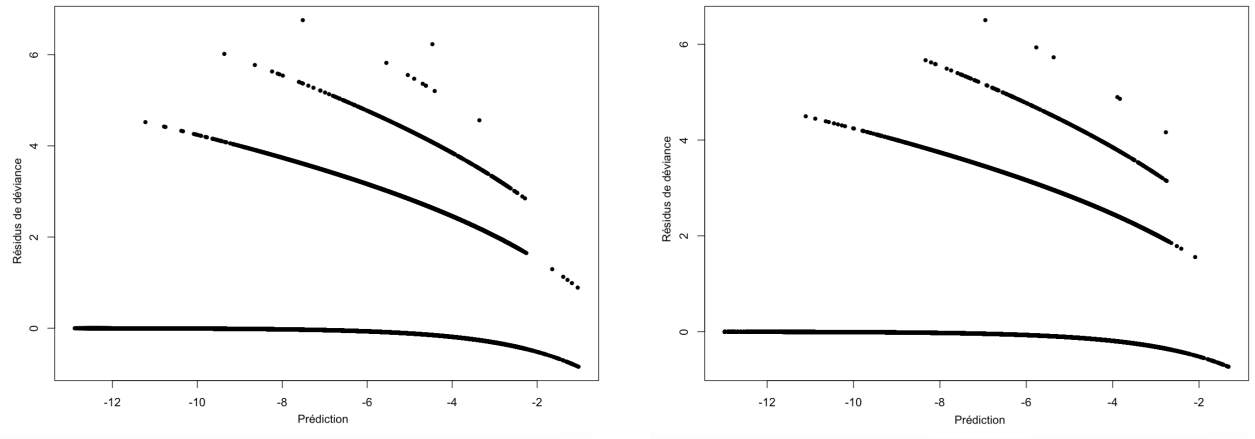
Les résidus des deux nouveaux GLMs doivent suivre une distribution de Poisson. Cela se vérifie à l'aide d'un test de Kolmogorov-Smirnov. Via la fonction ks.test() du package `stats` du logiciel R CORE TEAM (2022), il s'obtient pour les deux GLMs une p-value inférieure à 2.2e-16 et donc le rejet de la distribution Poisson des résidus. Comme pour la précédente modélisation GLM, l'hétérogénéité et l'hétéroscédasticité des réponses rendent difficile l'étude des résidus, d'où l'étude par la suite des résidus de déviance.

### Etude de l'indépendance des résidus de déviance

Les résidus de déviance des deux nouveaux GLMs doivent être indépendants. Cela se vérifie avec le test de Ljung-Box. Via la fonction ljungbox.test() du package `tseries` de R Core Team (2022), il s'obtient pour les deux GLMs une p-value inférieure à 2.2e-16 d'où le rejet de l'hypothèse nulle selon laquelle les résidus de déviance sont indépendants. Cela peut entraîner des erreurs dans les estimations des paramètres du modèle et dans les prévisions.

### Etude de la linéarité des résidus de déviance

L'étude de la linéarité des résidus de déviance permet de vérifier si l'hypothèse de linéarité de la relation entre la variable réponse et les variables explicatives est satisfaite. Comme pour la précédente modélisation GLM, il convient de tracer le nuage de points $\{(\hat{\eta}_i, \hat{\epsilon}_i^D)\}_{i=1}^n$, avec $\hat{\epsilon}_i^D$ le i-ème résidu de déviance et $\hat{\eta}_i = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j X_{ij}$, où les $\hat{\beta}_j$ sont les paramètres obtenus après calibration du GLM. Cela donne les graphiques suivants (figure 3.3).




(a) Nuage de points pour le GLM maladie

(b) Nuage de points pour le GLM accident

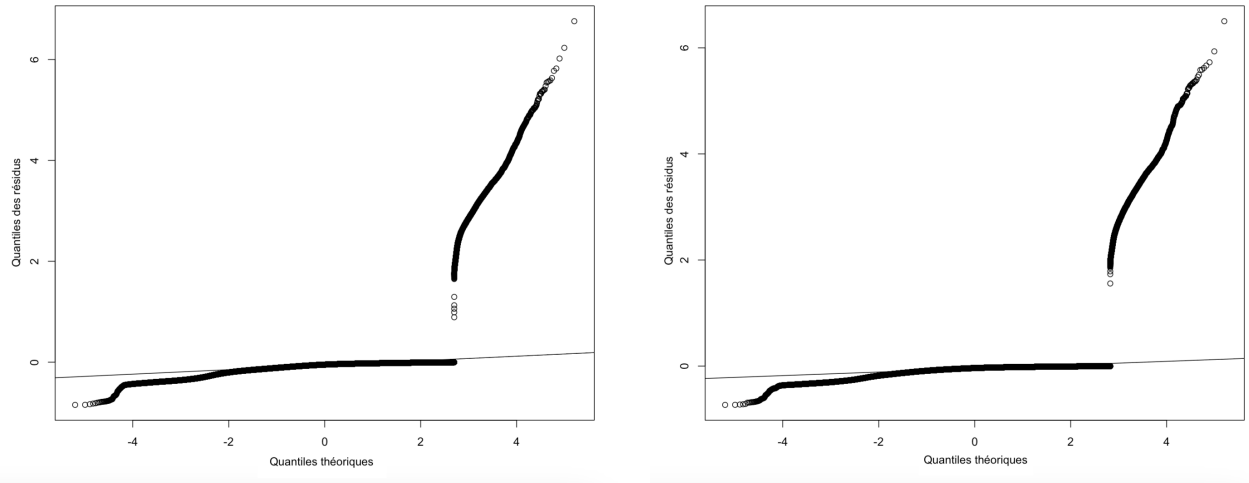
Figure 3.3 : Nuages de points des deux GLMs

Il s'observe pour les deux nouveaux GLMs une tendance dans les résidus par rapport à la prédiction. Il s'en déduit qu'il n'y a pas de linéarité de la relation entre la variable réponse et les variables explicatives.

### Etude de la normalité des résidus de déviance standardisés

Comme pour la précédente modélisation GLM, la normalité des résidus de déviance permet d'évaluer dans quelle mesure l'hypothèse de la distribution Poisson de la réponse est respectée. Il faut utiliser un graphique quantile-quantile pour vérifier si les résidus de déviance standardisés suivent une distribution normale centrée réduite. Sur le graphique, cela se vérifie si les points sont alignés sur la diagonale (figure 3.4).

Les résidus de déviance standardisés ne suivent pas une distribution normale centrée réduite dans les deux nouveaux GLMs. Il est donc compliqué de confirmer par ce biais que la distribution de la réponse est bien une distribution de Poisson.




(a) Diagramme quantile-quantile des résidus de déviance standardisés du GLM maladie

(b) Diagramme quantile-quantile des résidus de déviance standardisé du GLM accident

FIGURE 3.4 : Diagramme quantile-quantile des résidus de déviance standardisés des GLMs

Ainsi les résidus ne suivent pas une distribution de Poisson et les résidus de déviance ne sont ni indépendants, ni linéaires et ni distribués normalement. Comme pour la précédente modélisation, la mauvaise qualité des résidus est sans doute liée à la grande proportion de valeurs zéro dans les données. Cela montre encore une fois que le GLM Poisson n'est pas très adapté à ce type de données. Malgré tout l'étude de la nouvelle modélisation GLM continue en regardant la robustesse des nouvelles équations.

### 3.2.3 Validation de la robustesse des nouvelles équations des GLMs

#### Test de déviance anova

Les déviances obtenues des GLMs sont analysées grâce à la fonction anova() du package `stats` du logiciel R CORE TEAM (2022) en spécifiant le test du $\chi^2$, comme il a été fait dans le chapitre 2 (2.2.2).

Pour le GLM maladie, la variable explicative produit a deux paramètres (les produits Personal Loan et Mortgage Loan, Car Loan étant l'intercept) pour une déviance égale à 773.41 (tableau 3.5). Il est fait un test du $\chi^2$ pour regarder l'importance prédictive de cette variable et il est obtenu une p-value inférieure à 2.2e-16, donc elle réduit suffisamment la déviance pour être considérée comme significative dans le nouveau GLM maladie. De même pour les variables explicatives âge, IPC et taux de chômage, il est obtenu une p-value inférieure à 2.2e-16, d'où la pertinence de garder ces trois variables explicatives.

| Variable explicative | Degré de liberté | Déviance | Degré de liberté résiduel | Déviance résiduelle | p-value |
|---|---|---|---|---|---|
| Âge | 1 | 2 113,65 | 5 059 524 | 176 715 | <2e-16 |
| Produit | 2 | 773,41 | 5 059 522 | 175 941 | <2e-16 |
| IPC | 1 | 660,75 | 5 059 521 | 175 281 | <2e-16 |
| Taux de chômage | 1 | 295,85 | 5 059 520 | 174 985 | <2e-16 |

TABLE 3.5 : Résultat de la fonction anova() de R CORE TEAM (2022) avec test du $\chi^2$ pour analyser le pouvoir explicatif de chaque variable dans le GLM maladie

Pour le GLM accident, la variable explicative produit a deux paramètres (les produits Personal Loan et Mortgage Loan, Car Loan étant l'intercept) pour une déviance égale à 720.12 (tableau 3.6). Il est fait un test du $\chi^2$ pour regarder l'importance prédictive de cette variable et il est obtenu une p-value inférieure à 2.2e-16, donc elle réduit suffisamment la déviance pour être considérée comme significative dans le nouveau GLM accident. De même pour les variables explicatives genre, IPC et taux de chômage, il est obtenu une p-value inférieure à 2.2e-16, d'où la pertinence de garder ces trois variables explicatives.

| Variable explicative | Degré de liberté | Déviance | Degré de liberté résiduel | Déviance résiduelle | p-value |
|---|---|---|---|---|---|
| Genre | 1 | 825,24 | 5 050 502 | 125 645 | <2e-16 |
| Produit | 2 | 720,12 | 5 050 500 | 124 924 | <2e-16 |
| IPC | 1 | 234,6 | 5 050 499 | 124 690 | <2e-16 |
| Taux de chômage | 1 | 132,3 | 5 050 498 | 124 558 | <2e-16 |

TABLE 3.6 : Résultat de la fonction anova() de R CORE TEAM (2022) avec test du $\chi^2$ pour analyser le pouvoir explicatif de chaque variable dans le GLM accident

Procédons maintenant à une validation croisée pour étudier la robustesse des variables explicatives et voir s'il y a du sur-apprentissage.

### Validation croisée

La méthodologie appliquée est la même que pour la modélisation du chapitre précédent. Les données des deux couvertures sont séparées en deux jeux de données, un de 80% appelé *"Train set"* et un de 20% appelé *"Validation set"*.

Pour le GLM maladie, le GLM est entraîné sur le jeu de données *"Train set"* et est appliqué sur le *"Validation set"*.

Il est regardé dans un premier temps si les deux échantillons ont une tendance et des valeurs proches par âge (figure 3.5) avec en noir la fréquence totale, en vert pointillé la fréquence du *"Train set"* et en bleu pointillé la fréquence du *"Validation set"*.

Les deux jeux de données sont répartis de manière homogène.

Maintenant la fréquence *"Validation"* observée est comparée graphiquement avec la fréquence prédite selon l'âge dans le graphe du haut et selon l'année d'accident dans le graphe du bas (figure 3.6).

Il peut se voir graphiquement que la prédiction via le GLM entraîné sur le *"Train set"* des données du *"Validation set"* est alignée avec les observations de ces dernières que ça soit par âge et par année d'incident. Ainsi le nouveau GLM, grâce à l'ajout des variables explicatives macroéconomiques, prend maintenant en compte la croissance observée par année d'incident en plus de celle par âge. Visuellement, cela confirme que toutes les variables explicatives de ce GLM sur les données maladie ont un pouvoir prédictif nécessaire au modèle qui ne sur-apprend pas les données. La robustesse des variables explicatives choisies est vérifiée via le calcul de l'EDR.

La méthodologie est rappelée. Le GLM est entraîné sur le *"Train set"* et prédit toutes les données, puis le GLM est entraîné sur le *"Validation set"* et prédit toutes les données. Le calcul des deux EDR peut se faire et s'ils sont proches, alors les variables explicatives choisies dans le GLM sont robustes et il n'a pas de sur-apprentissage des données :

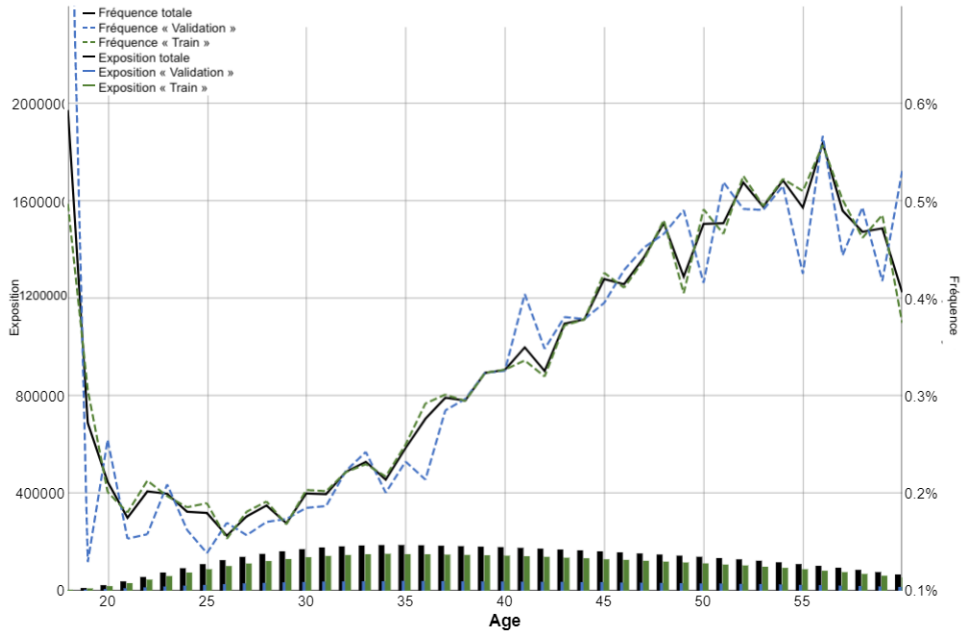



FIGURE 3.5 : Fréquence *"Train"* vs Fréquence *"Validation"* vs Fréquence totale en fonction de l'âge des données maladie

- EDR des données prédites sur le *"Train set"* $= 1 - \frac{174986.5}{178828.9} = 2.15\%$ ;
- EDR des données prédites sur le *"Validation set"* $= 1 - \frac{175011.3}{178836.3} = 2.14\%$.

Les deux EDR étant très proches, cela confirme la robustesse des variables explicatives âge, produit, taux de chômage et IPC dans le GLM maladie, et le modèle ne sur-apprend pas les données.

Pour le GLM accident, le GLM est entraîné sur le jeu de données *"Train set"* et est appliqué sur le *"Validation set"*.

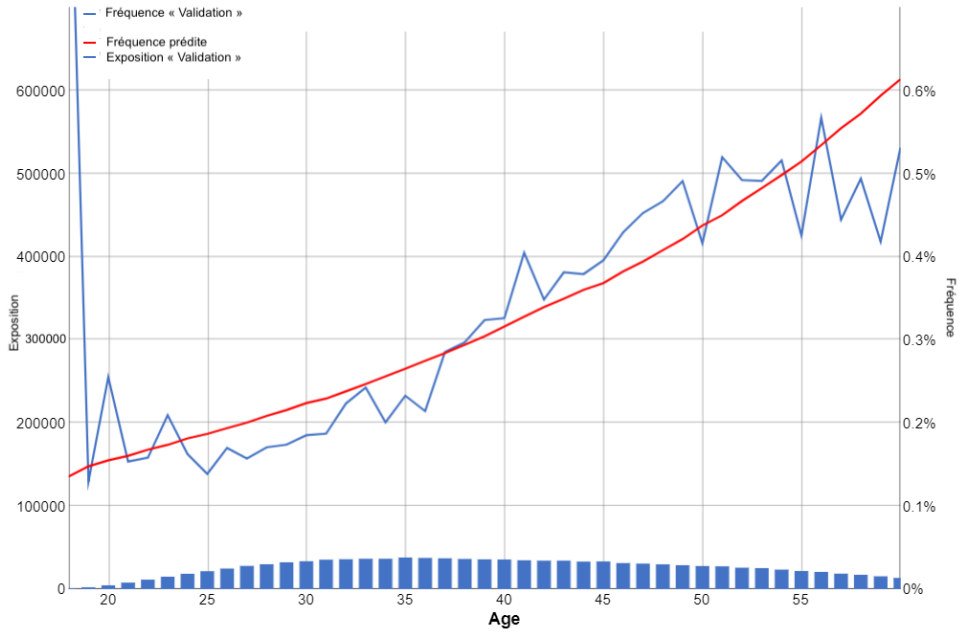
Il est regardé dans un premier temps si les deux échantillons ont bien une tendance et des valeurs proches par âge pour chaque genre (figure 3.7), en haut les hommes et en bas les femmes, avec en noir la fréquence totale, en bleu pointillé la fréquence *"Validation"* et en rouge pointillé la fréquence "Train". C'est bien le cas malgré quelques écarts chez les femmes à partir de 50 ans.

Puis il convient de comparer graphiquement la fréquence *"Validation"* observée avec la fréquence prédite selon l'âge pour les hommes dans le graphe du haut et pour les femmes dans le graphe du bas (figure 3.8), avec en bleu la fréquence des données *"Validation"* et en rouge la prédiction du GLM faite sur ces données.

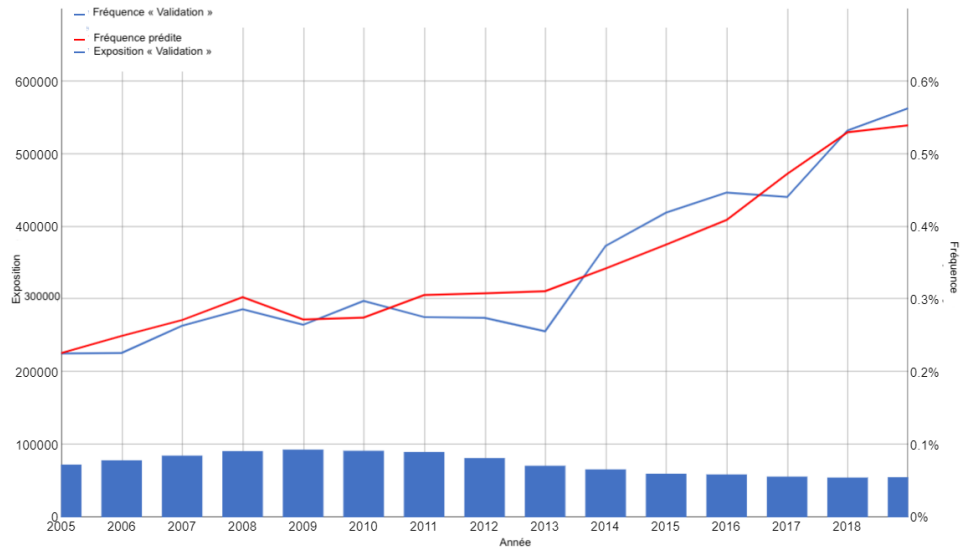
Il se voit graphiquement que la prédiction via le GLM entraîné sur le *"Train set"* des données du *"Validation set"* est alignée avec ces dernières par âge pour les deux genres.

Maintenant il faut regarder si comme pour le GLM maladie, la croissance de la fréquence avec les années d'incident est bien prédite (figure 3.9), en haut les hommes et en bas les femmes avec en bleu la fréquence observée par année et en rouge la prédiction du GLM accident.

Pour les hommes, cette croissance avec les années est bien prédite en étant proche des fréquences du *"Validation set"*. Cependant c'est moins le cas chez les femmes où malgré une croissance de la




(a) Fréquence observée *"Validation"* par rapport à la prédiction du nouveau GLM maladie par âge



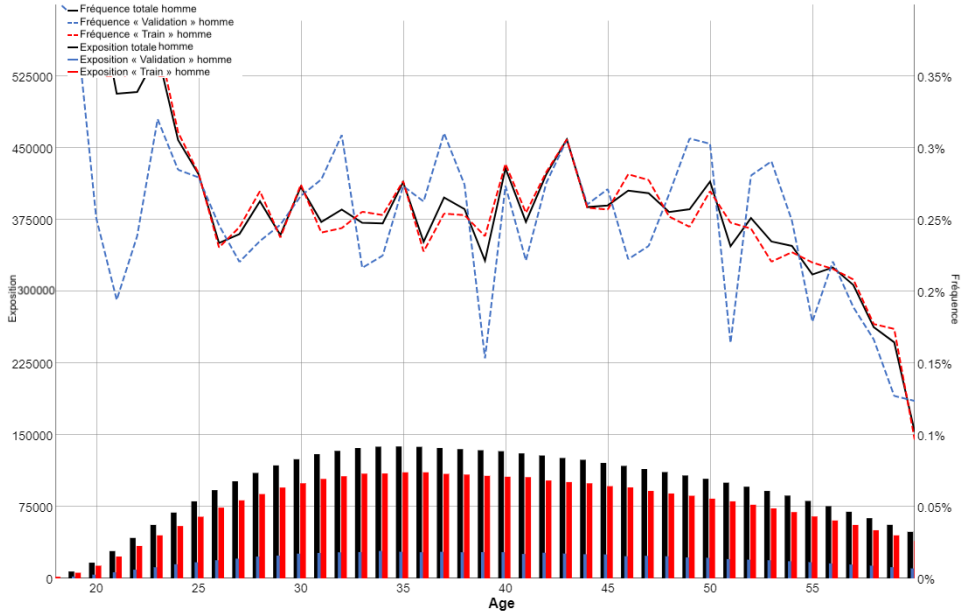

(b) Fréquence observée *"Validation"* par rapport à la prédiction du nouveau GLM maladie par année d'incident

FIGURE 3.6 : Fréquence observée du *"Validation set"* par rapport à la prédiction du nouveau GLM maladie

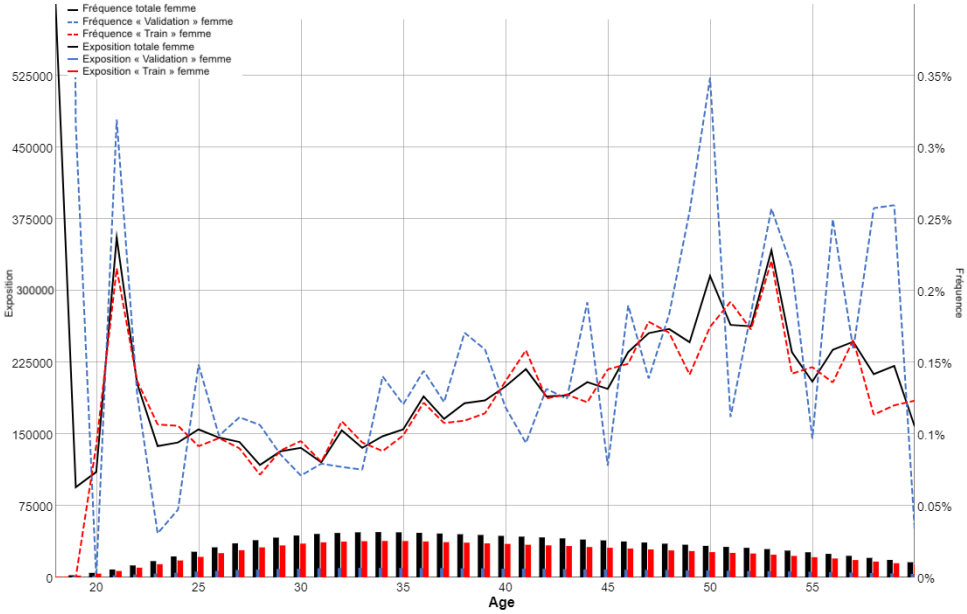
fréquence prédite, celle-ci n'a pas la même tendance notamment à partir de l'année 2013.

Regardons à présent comme pour le GLM maladie les valeurs des EDR :

- EDR des données prédites sur le *"Train set"* $= 1 - \frac{124561.4}{126469.7} = 1.51\%$ ;
- EDR des données prédites sur le *"Validation set"* $= 1 - \frac{124615.3}{126469.8} = 1.47\%$.




(a) Fréquence *"Train"* vs Fréquence *"Validation"* vs Fréquence totale en fonction de l'âge des données accident des hommes

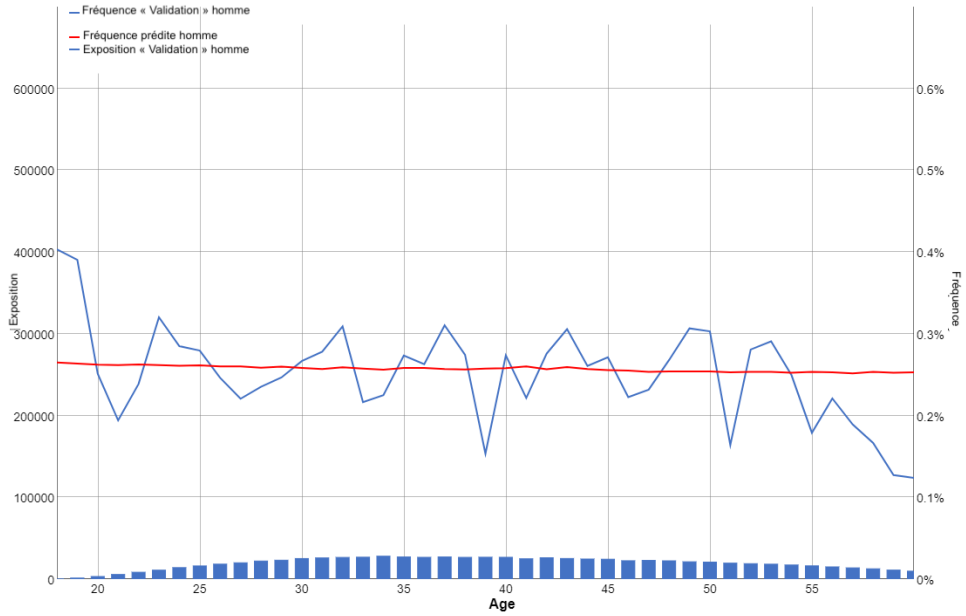



(b) Fréquence *"Train"* vs Fréquence *"Validation"* vs Fréquence totale en fonction de l'âge des données accident des femmes

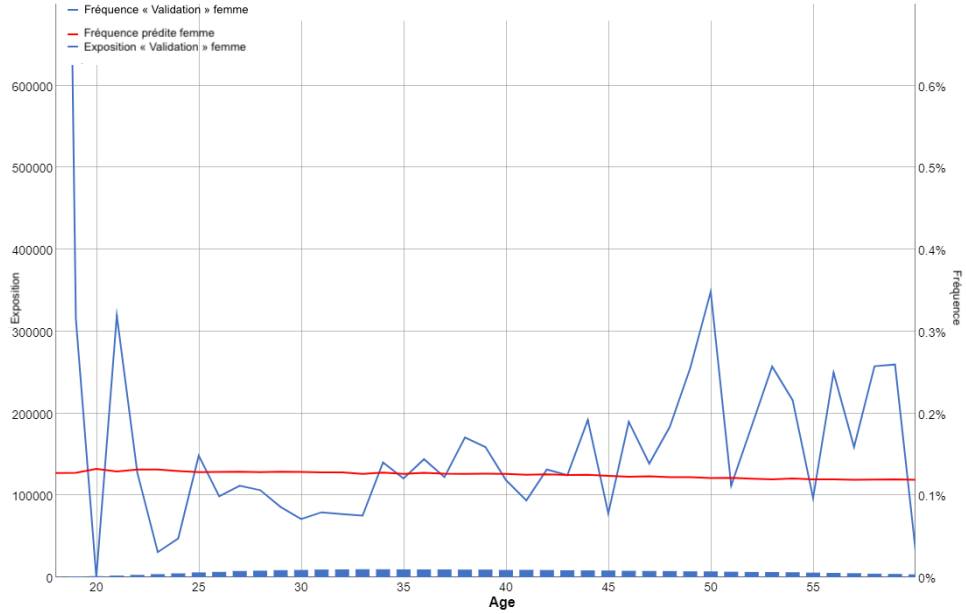
FIGURE 3.7 : Fréquence *"Train"* vs Fréquence *"Validation"* vs Fréquence totale en fonction de l'âge des données accident par genre

Il se trouve que les EDR sont proches et cela permet de conclure que les variables explicatives choisies pour le GLM accident, à savoir le genre, le produit, le taux de chômage et l'IPC sont robustes dans la modélisation et le modèle ne sur-apprend pas les données.

Il est intéressant de regarder à présent les résultats des GLMs à l'échelle d'un produit, le Personal Loan. Dans un premier temps, il est analysé graphiquement l'ajustement des GLMs par rapport aux



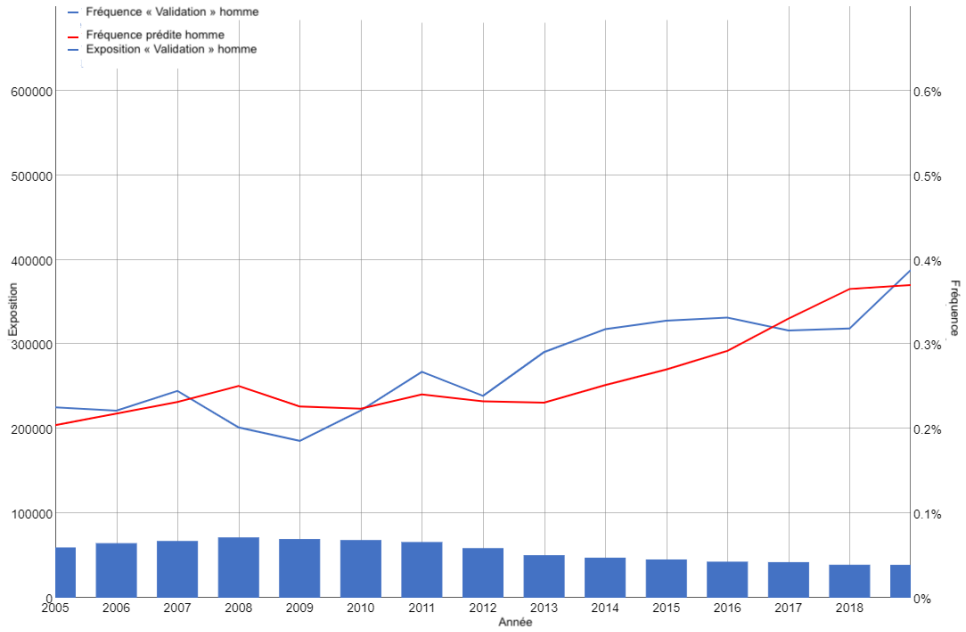

(a) Fréquence observée du *"Validation set"* par rapport à la prédiction du nouveau GLM accident par âge pour les hommes



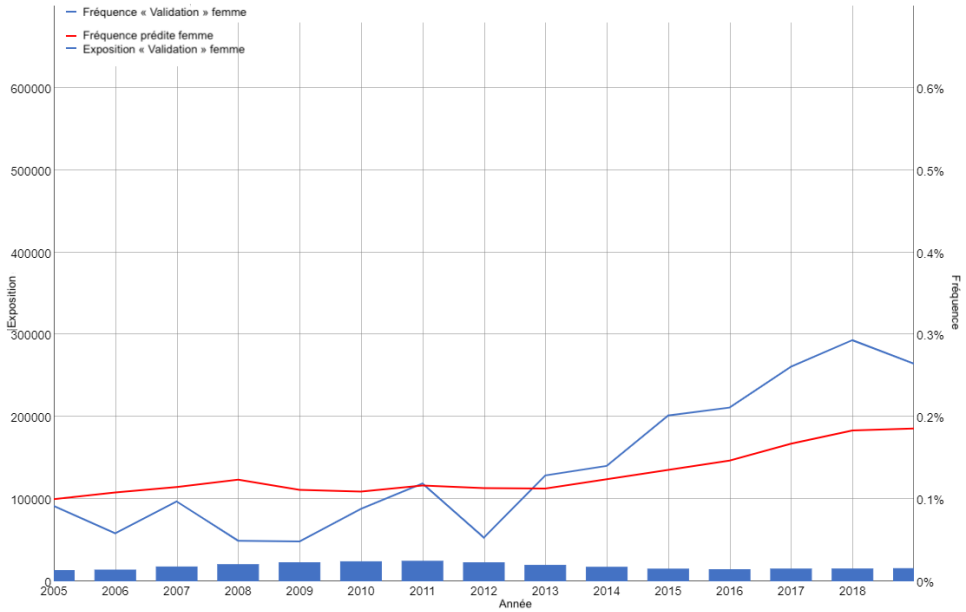

(b) Fréquence observée du *"Validation set"* par rapport à la prédiction du nouveau GLM accident par âge pour les femmes

FIGURE 3.8 : Fréquence observée du *"Validation set"* par rapport à la prédiction du nouveau GLM accident par âge pour chaque genre

observations selon l'âge. Puis il est regardé l'ajustement des deux GLMs ensemble par rapport aux observations par année.




(a) Fréquence observée du *"Validation set"* par rapport à la prédiction du nouveau GLM accident par année pour les hommes




(b) Fréquence observée du *"Validation set"* par rapport à la prédiction du nouveau GLM accident par année pour les femmes

FIGURE 3.9 : Fréquence observée du *"Validation set"* par rapport à la prédiction du nouveau GLM accident par année pour les deux genres

### 3.2.4 Modèle pour le produit Personal Loan

#### Résultats des GLMs par couverture selon l'âge et le genre

Le graphique du haut présente la fréquence prédite (courbe verte) par le GLM maladie comparée aux fréquences observées des hommes (courbe rouge pointillée) et des femmes (courbe bleue pointillée) selon l'âge, et le graphique du bas présente les fréquences prédites par le GLM accident (courbes pleines rouges et bleues) en fonction du genre comparées aux observations masculines (courbe rouge pointillée) et féminines (courbe bleue pointillée) selon l'âge (figure 3.10).

L'ajustement n'a pas changé par rapport aux précédents GLMs (avant modification de la valeur du coefficient du produit Personal Loan dans les équations), la fréquence prédite par le GLM est croissante par âge et alignée avec les observations pour la maladie, et la fréquence prédite pour l'accident est séparée par genre et alignée avec les observations hommes et femmes.

#### Résultats des GLMs par année d'incident, maladie et accident ensemble

Il est regardé à présent si la prédiction accident et maladie ensemble s'aligne bien avec les observations par année d'incident et la hausse observée, ce qui n'était pas le cas avec la précédente modélisation GLM (2.12).

La courbe bleue correspond aux observations par année d'incident, la courbe verte la prédiction des deux GLMs ensemble et la courbe noire la fréquence du modèle de 2018 (figure 3.11).

Cette fois, l'ajustement des GLMs (courbe verte) s'aligne avec les observations (courbe bleue) et prédit bien la croissance observée par année d'incident.

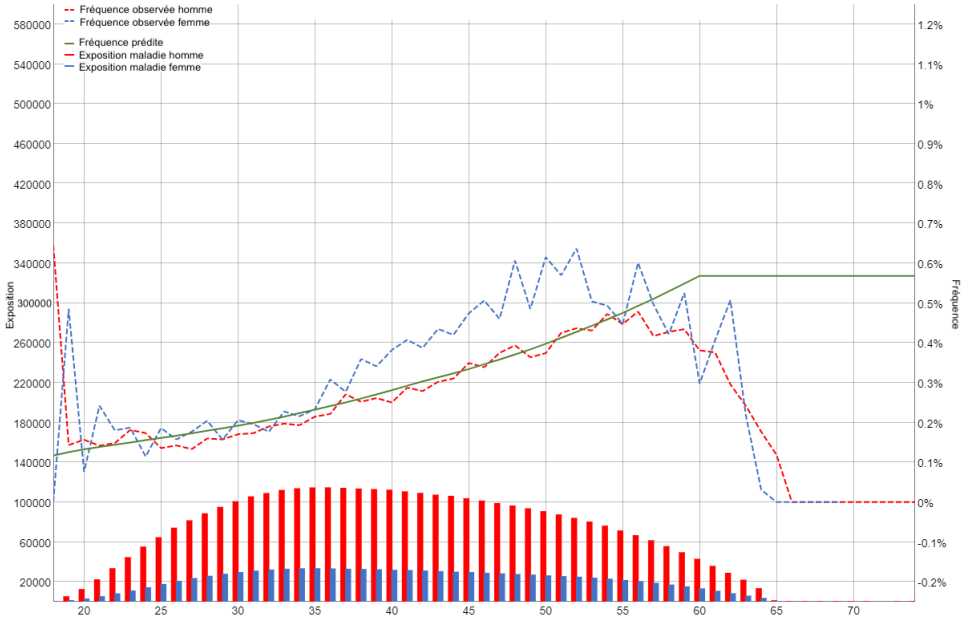
**Cela confirme que l'apport de nouvelles variables explicatives aux GLMs vient enrichir la modélisation et prendre en compte la hausse de la fréquence avec les années d'observations pour le produit Personal Loan**. Il en est de même pour l'autre produit "non Bulk" Mortgage Loan. En revanche pour le produit "non Bulk" Car Loan, la fréquence prédite croît fortement à partir de 2015 alors que les observations restent stagnantes (E). Cela montre une limite de cette nouvelle modélisation GLM, qui à cause des variables macroéconomiques va indiquer une croissance de la fréquence pour tous les produits, même ceux qui n'en ont pas.

Concernant les produits "Bulk", il faudrait procéder à une méthodologie similaire à la modélisation du chapitre 1 pour les ajouter aux équations des GLMs.

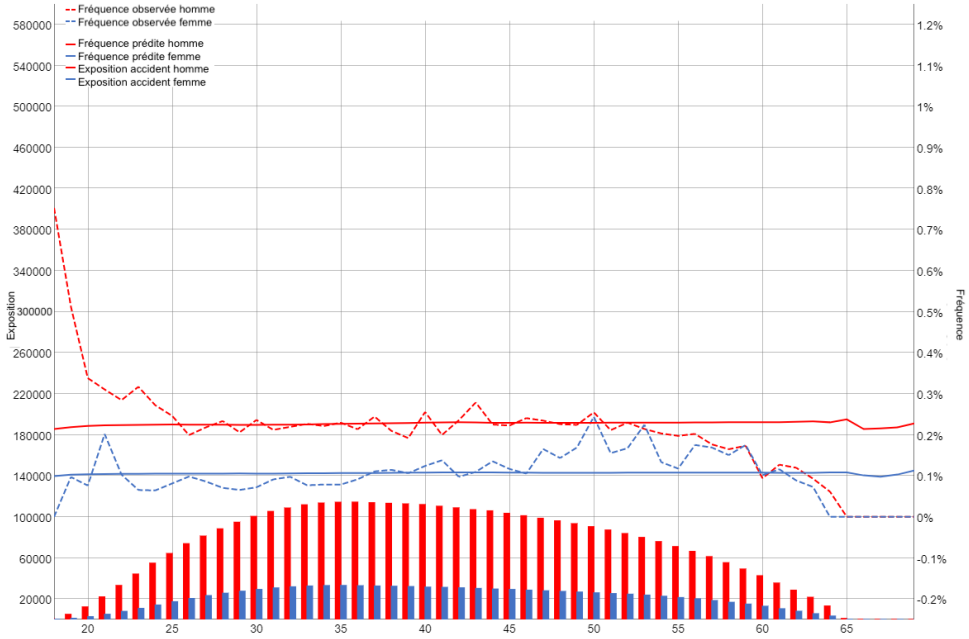
Il est décidé ici de ne pas le faire et les équations suivantes sont obtenues :

- pour la maladie, $Frequence = \exp(-10.190 + 0.035 \times Age - 0.503 \times \mathbf{1}_{\text{Personal Loan}} - 0.120 \times \mathbf{1}_{\text{Mortgage Loan}} + 0.040 \times IPC - 0.041 \times \text{taux de chômage})$ ;
- pour l'accident, $Frequence = \exp(-7.929 - 0.766 \times \mathbf{1}_{\text{Genre féminin}} - 0.631 \times \mathbf{1}_{\text{Personal Loan}} - 0.201 \times \mathbf{1}_{\text{Mortgage Loan}} + 0.030 \times IPC - 0.034 \times \text{taux de chômage})$.

L'exemple de calcul de la prime pure (unique et périodique annuelle) est repris avec les fréquences obtenues dans cette modélisation.




(a) Fréquences observées vs Fréquence prédite pour le produit Personal Loan en couverture maladie selon l'âge




(b) Fréquences observées vs Fréquences prédites pour le produit Personal Loan en couverture accident selon l'âge

FIGURE 3.10 : Fréquences observées vs Fréquences prédites pour le produit Personal Loan des deux couvertures

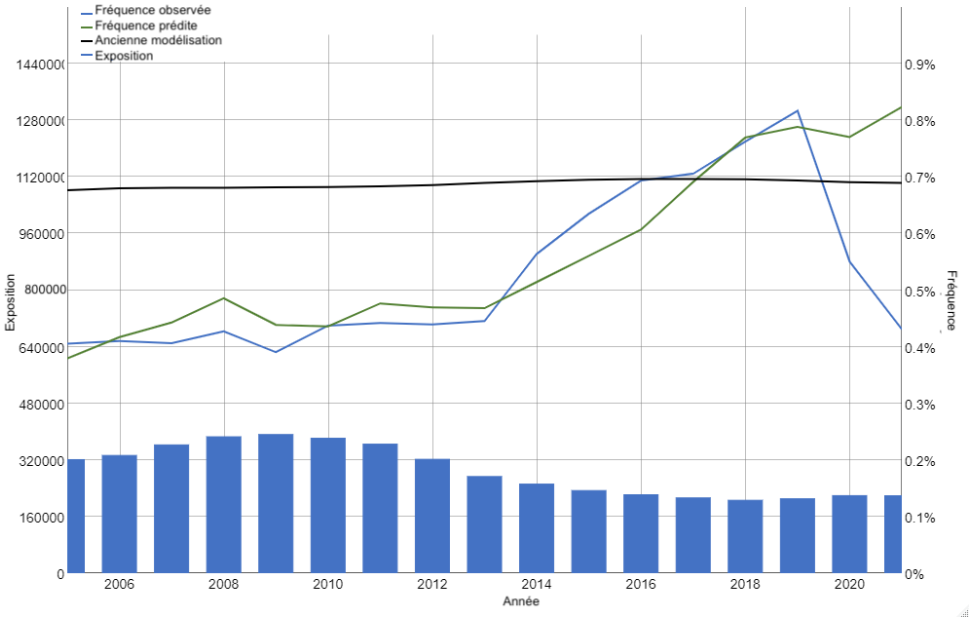



FIGURE 3.11 : Fréquence observée, fréquence prédite et ancienne fréquence du modèle de 2018 tous genres et toutes couvertures confondus pour le produit Personal Loan

### 3.2.5 Calcul de la prime pure pour le produit Personal Loan

Le cadre de l'exemple est le même que pour la précédente modélisation (2.3). Il est supposé que l'exposition reste également constante sur les 5 années, mais cette fois, la fréquence augmente avec les années compte-tenu de l'utilisation des variables macroéconomiques taux de chômage et IPC, pour lesquelles sont disponibles des prédictions pour la période 2023-2027. Voici un tableau récapitulatif (tableau 3.7).

| Année | 2023 | 2024 | 2025 | 2026 | 2027 |
|---|---|---|---|---|---|
| Exposition | 1000 | 1000 | 1000 | 1000 | 1000 |
| Taux de chômage | 5,34% | 5,12% | 5,01% | 5,19% | 5,37% |
| IPC | 113,9 | 114,6 | 115,6 | 117,2 | 119,1 |
| Fréquence | 1,202% | 1,243% | 1,298% | 1,364% | 1,455% |
| Nombre de sinistres | 12,0 | 12,4 | 13,0 | 13,6 | 14,6 |

TABLE 3.7 : Récapitulatif de l'exposition, du taux de chômage, de l'IPC, de la fréquence et du nombre de sinistres par année

Le montant du prêt est de 20000€, à rembourser sur 5 ans avec un taux d'intérêt de 5% soit un remboursement de 374.88€ par mois. Avec une durée moyenne d'incapacité de 6 mois, le coût total $C_i$ que l'assureur doit supporter par année i est égal à Nombre de sinistres l'année i×Remboursement mensuel du prêt× Durée de l'incapacité. Cela donne $C_{2023} = 27029.71$, $C_{2024} = 27951.51$, $C_{2025} = 29185.62$, $C_{2026} = 30676.68$ et $C_{2027} = 32730.67$.

Par assuré, cela donne $c_{2023} = 27.03$, $c_{2024} = 27.95$, $c_{2025} = 29.19$, $c_{2026} = 30.68$ et $c_{2027} = 32.73$.

Il est calculé la valeur actualisée de ces coûts annuels $VA = \sum_{i=2023}^{2027} \frac{c_i}{(1+0.75\%)^{i-2023}} = 145.29$€. Il en

est déduit la prime pure unique qui est égale à la VA. Pour la prime pure annuelle, il faut calculer la valeur actualisée des primes versées par l'assuré

$VA^{\text{assuré}} = \sum_{i=2023}^{2027} \frac{\text{prime pure annuelle}}{(1+0.75\%)^{i-2023}}$.

Par équité actuarielle il doit y avoir $VA = VA^{\text{assuré}}$

$\Rightarrow$ Prime pure annuelle $= \frac{VA}{\sum_{i=2023}^{2027} \frac{1}{(1+0.75\%)^{i-2023}}} = 29.64€$ (tableau 3.8).

| Prime pure unique | Prime pure annuelle |
|---|---|
| 145,29 € | 29,64 € |

TABLE 3.8 : Prime pure unique et prime pure annuelle

Les deux types de primes ont augmenté avec cette nouvelle modélisation de la fréquence :

- la prime pure unique passe de 91.26€ (premier modèle) à 145.29€ ($\times$1.6) ;
- la prime pure annuelle passe de 18.53€ (premier modèle) à 29.64€ ($\times$1.6).

La différence importante entre les primes vient du fait que dans le premier modèle, il a été supposé que la fréquence stagne à 0.824% par an, alignée avec le pic de fréquence de 2019. Alors que dans le deuxième modèle, la fréquence continue d'augmenter fortement après 2019 sur la période 2023-2027 (liée aux prédictions récupérées du taux de chômage et de l'IPC, notamment l'IPC qui augmente fortement au cours des 5 prochaines années). Par exemple pour l'année 2027, la premier modèle (après ajustement) prédit une fréquence de 0.824% contre 1.455% dans le deuxième modèle.

Le deuxième modèle est plus prudent, reste à savoir comment la fréquence va réellement se comporter dans les années futures : va-t'elle recommencer à augmenter comme elle le faisait avant 2019 (vision du second modèle) ou va-t'elle stagner en dessous du pic observé de 2019 (vision du premier modèle).

En outre, si les primes obtenues avec le premier modèle sans ajustement sont comparés aux primes obtenues avec le deuxième modèle, il s'obtient un rapport de $\times$2.58. L'écart entre ces deux modèles est beaucoup trop important, et illustre une vraie sous-estimation de la fréquence par le premier modèle (sans ajustement) et sans doute une surestimation de la fréquence future par le second modèle. Il est donc intéressant d'implémenter une troisième modélisation complètement différente pour comparer les résultats et tirer des conclusions plus fines.

C'est ainsi que le *machine learning* et la méthode du *boosting* sont introduits. Cette méthode est appliquée sur le jeu de données enrichi des variables explicatives macroéconomiques et sa pertinence prédictive est analysée.

## 3.3 La modélisation *boosting* des données incapacité

Il est intéressant dans un premier temps rappeler ce que sont le *machine learning* et le *boosting*.

### 3.3.1 Définition du machine learning

Le *machine learning* est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre sans avoir à être programmés en amont pour cela. Les algorithmes de *machine learning* vont avoir la capacité de déterminer des répétitions dans des flux de données et en dériver des prédictions. Plus les données vont être fournies, et plus l'algorithme va en tirer des tendances, d'où le lien fort entre le *machine learning* et le *Big Data* (technologie de stockage numérique d'un nombre extraordinaire de données). En d'autres termes, plus les données sont enrichies et variées, et plus les algorithmes de *machine learning* vont apprendre et être performants, à l'inverse des modèles traditionnels qui ont besoin que les données ne dépassent pas un certain volume. Une fois un algorithme de *machine learning* entraîné sur un jeu de données, celui-ci pourra retrouver les tendances déterminées dans de nouvelles données (Azencott (2018)).

Quatre étapes se succèdent dans le développement d'un modèle de *machine learning*(DataScientest (2019)) :

- le choix et la préparation des données d'entraînement. Ce sont les données utilisées pour l'apprentissage de l'algorithme à résoudre le problème pour lequel il est choisi. Elles doivent en amont être nettoyées et préparées pour éviter tout biais dans la modélisation ;
- le choix du type d'algorithme sur lequel les données sont entraînées. Il dépend du type et de la taille des données d'entraînement et du type de problème que l'utilisateur souhaite résoudre ;
- l'entraînement de l'algorithme de *machine learning*. L'algorithme est exécuté et les résultats obtenus sont comparés avec ceux qu'il y aurait dû avoir, puis le poids affecté à chaque variable est modifié pour améliorer la précision des prédictions de l'algorithme. Ce processus itératif continue jusqu'à atteindre un niveau de satisfaction explicité en amont ;
- l'application du modèle sur de nouvelles données. Ces nouvelles données servent également à améliorer la qualité prédictive du modèle qui apprend aussi avec elles.

Il existe une diversité d'algorithmes de *machine learning* précisés ici :

- les algorithmes de régression étudient les relations entre les données ;
  - les algorithmes de régression linéaire (2.1.1),
  - les algorithmes de régression logistique, utilisés lorsque la variable à expliquer est binaire,
- les algorithmes d'association, qui associent une conclusion spécifique à un ensemble de conditions ;
- les réseaux de neurones artificiels. L'algorithme se présente sous forme de plusieurs couches. La première couche correspond à l'entrée des données, les couches cachées traitent ces données récupérées et la dernière couche transmet divers résultats avec un poids affecté à chacun.

Enfin, voici les différents types de *machine learning* :

- l'apprentissage supervisé où l'algorithme connaît déjà la réponse attendue de lui. Le travail s'effectue avec des données étiquetées, c'est-à-dire que les propriétés et les classifications des données sont mises en avant, en vue d'indiquer à l'algorithme que ce sont les cibles qu'il devra prédire. Cette méthode permet de faire des classifications (affecter une classe à un objet) et des régressions (affecter une valeur à un objet) ;

- l'apprentissage non supervisé où les réponses que l'utilisateur souhaite prédire ne sont pas présentes dans le jeu de données, il faut donc que l'algorithme crée ses propres réponses. Cette méthode sert à faire du *clustering* (grouper des objets dans des classes les plus homogènes possibles, classes choisies par l'algorithme et non par les hommes comme en apprentissage supervisé) ;

- l'apprentissage par renforcement : l'algorithme apprend seul de ses erreurs pour atteindre un but via diverses approches.

Dans le cas de cette étude, il faut expliquer la variable réponse qu'est la fréquence de sinistres par diverses variables explicatives présentes dans la base de données. L'algorithme de *machine learning* utilisé pour cela est de type apprentissage supervisé puisqu'il est indiqué à l'algorithme la variable à prédire, et il s'agit d'une régression linéaire puisqu'il est souhaité de déterminer une valeur mathématique de cette variable. Il est choisi en conséquence de faire un *gradient boosting*, qui est défini dans la section suivante.

### 3.3.2 Définition du *gradient boosting*

L'algorithme de *boosting de gradient* utilise des arbres de décision.

#### L'arbre de décision

Un arbre de décision est un modèle de classification supervisé qui utilise une structure hiérarchique de nœuds et de branches pour représenter les relations entre les caractéristiques des données d'entrée et les classes de sortie. Il est utilisé pour résoudre des problèmes de classification binaire ou multiclasse.

Les arbres de décision sont construits à l'aide d'algorithmes d'apprentissage automatique qui utilisent une mesure de qualité de classification telle que la log-vraisemblance ou l'entropie pour maximiser la performance. L'algorithme de construction de l'arbre est généralement basé sur un parcours de l'arbre qui permet de trouver la meilleure séparation des données en utilisant une mesure de qualité de classification.

Les arbres de décision sont constitués de nœuds de décision et de feuilles. Les nœuds de décision représentent des tests sur les caractéristiques des données d'entrée, tandis que les feuilles représentent les classes ou les prédictions finales. A chaque nœud de décision, un test est effectué sur une caractéristique de l'entrée, en utilisant une valeur seuil. En fonction du résultat du test, il faut se déplacer à gauche ou à droite de l'arbre (exemple 3.12 avec l'âge comme variable explicative de la fréquence).

Les arbres de décision sont faciles à comprendre et à interpréter car ils permettent de visualiser les règles de décision qui ont été utilisées pour effectuer les prédictions.

C'est aussi un algorithme qui est très sensible aux variations de données dans le jeu d'entraînement. Enfin, plus l'arbre est profond et a de branches, plus il va faire du cas par cas pour chaque donnée et tendre vers un problème de sur-apprentissage.

L'utilisation d'un algorithme de *gradient boosting* va fortement limiter les défauts cités ci-dessus.

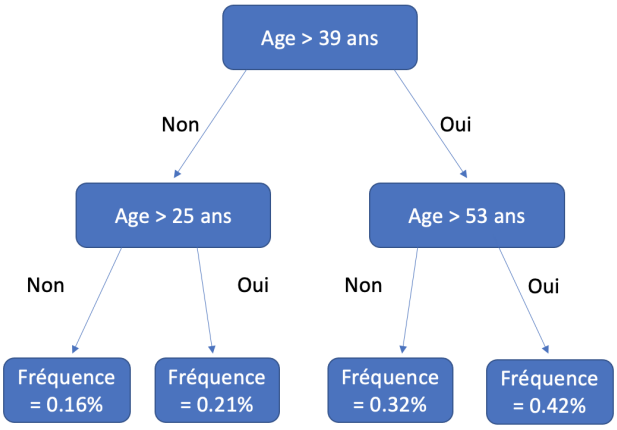



FIGURE 3.12 : Exemple d'arbre de décision en couverture maladie

### Le *gradient boosting*

Le *gradient boosting* est un algorithme d'apprentissage automatique qui utilise une technique d'optimisation pour construire un modèle de prédiction en combinant plusieurs modèles de base, appelés "estimateurs faibles", en un modèle plus performant, appelé "estimateur fort". Ces "estimateurs faibles" sont des arbres de décisions d'une faible qualité prédictive. Il est principalement utilisé pour les tâches de régression et de classification.

Le *gradient boosting* utilise une fonction de perte pour mesurer l'erreur entre les prédictions du modèle et les valeurs réelles. L'algorithme de *gradient boosting* consiste à ajouter successivement des estimateurs faibles qui minimisent cette fonction de perte en se concentrant sur les observations les plus difficiles à prédire.

Il utilise un processus itératif pour ajouter des modèles de base qui corrigent les erreurs commises par les modèles précédents. Pour chaque itération, un nouveau modèle est formé en utilisant la fonction de perte et en se concentrant sur les observations mal prédites par les modèles précédents.

La combinaison de ces modèles de base permet de construire un modèle plus performant qui peut mieux se généraliser à des données nouvelles. La profondeur et le nombre d'arbres à construire des modèles de base sont des hyperparamètres du *gradient boosting* que l'utilisateur doit spécifier en amont.

L'algorithme de *boosting* de cette étude est le *Gradient Boosting Machines* (GBM). Celui-ci utilise le principe de descente de gradient et se décline en ces étapes itératives (NATEKIN et KNOLL (2013)) :

- Entrée : le jeu de données $\{(x_i, y_i)\}_{i=1}^{n}$ et la fonction de perte $L(y_i, F(x_i))$

  où $y_i$ correspond à l'observation i de la variable réponse, i allant de 1 à n le nombre d'observations totales et $x_i$ est l'observation i de toutes les variables explicatives. La fonction $F()$ est la prédiction faite à partir de l'observation $x_i$, et $L()$ est la fonction qui évalue la capacité à prédire $y_i$. La fonction de perte utilisée ici est $L(y_i, F(x_i)) = (y_i - F(x_i))^2$.

- Étape 1 : initialisation du modèle avec la valeur constante $F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$.

  Il s'obtient $F_0(x) = \frac{y_1+...+y_n}{n}$, la moyenne des observations de la variable réponse. Ainsi le premier arbre de décision construit donne toujours la moyenne des observations comme prédiction, c'est évidemment un apprenant très faible.

- Étape 2 : pour m allant de 1 à M (M est le nombre d'arbres que l'utilisateur souhaite construire) :
  - (A) : calculer $r_{im} = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}$, pour i allant de 1 à n.
    $r_{im}$ est le pseudo résidu de chaque observation i dans l'arbre m. La partie entre crochets est le gradient, d'où le nom de *gradient boosting*. Il faut calculer la dérivée de la fonction de perte afin de déterminer l'écart entre l'observation et la prédiction de l'arbre m-1 (en effet avec la fonction de lien introduite plus haut, cela donne $r_{im} = y_i - F(x_i)$). Pour m = 1, le premier pseudo résidu $r_{i1}$ correspond à l'écart entre la variable réponse i et la prédiction de l'arbre 0, qui est la moyenne des observations pour tout $x_i$.
  - (B) : ajuster un arbre de décision à la variable $r_{im}$ et créer les régions terminales $R_{jm}$, pour j allant de 1 à $J_m$.
    L'algorithme crée un arbre de décision pour expliquer la variable résidu, et non plus la variable réponse initiale. $R_{jm}$ correspond à la feuille numéro j de l'arbre m, l'arbre m ayant $J_m$ feuilles.
  - (C) : pour j allant de 1 à $J_m$, calculer $\gamma_{jm} = argmin_\gamma \sum_{x \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$.
    Pour chaque feuille, il faut calculer $\gamma_{jm}$ qui est le meilleur résidu pour prédire la variable réponse via les observations $x_i$ de la feuille, en utilisant la précédente prédiction $F_{m-1}$.
  - (D) : mettre à jour la prédiction pour chaque échantillon de données, $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}_{x \in R_{jm}}$.
    La nouvelle prédiction $F_m$ dépend de la précédente $F_{m-1}$, à laquelle sont ajoutés les résidus $\gamma_{jm}$ obtenus pour chaque feuille de l'arbre m. $\nu$ est appelé taux d'apprentissage ; il est compris entre 0 et 1. Plus $\nu$ est petit, plus il réduit les effets de chaque arbre sur la prédiction finale, ce qui améliore la précision à long terme. $\nu$ est le troisième hyperparamètre que l'utilisateur doit spécifier avant d'implémenter l'algorithme.

Cet algorithme s'implémente sur le logiciel R Core Team (2022) via la fonction gbm() du package `gbm`, où l'utilisateur doit spécifier les trois hyperparamètres :

- le nombre d'arbres que l'algorithme va construire (en général entre 1000 et 10000) ;
- la profondeur de chaque arbre (en général entre 2 et 4) ;
- la valeur du taux d'apprentissage $\nu$ (en général entre 0.01 et 0.001).

Maintenant défini, l'algorithme peut être appliqué sur la base de données.

### 3.3.3 Implémentation du *gradient boosting* sur les données

La même base de données filtrée est gardée (pas de produits "Bulk", période comprise entre 2005 et 2019, pas le partenaire "F" pour le produit Mortgage Loan) mais il n'y a plus la séparation de la base entre couverture maladie et accident, la couverture étant considérée maintenant comme une variable explicative. De plus, il est souhaité que la variable réponse soit la fréquence et non plus le nombre de sinistres. Il est ajouté en conséquence une colonne fréquence à la base de données.

Pour résumer, la variable réponse est la fréquence, les variables explicatives sont l'âge, le genre, le produit, la couverture, le taux de chômage et l'IPC.

Afin de choisir le meilleur algorithme de *gradient boosting*, il faut jouer avec les hyperparamètres, et comparer les algorithmes obtenus via les critères de **déviance** et de **sur-apprentissage**. La meilleure

méthode pour sélectionner les paramètres est de définir une *grid-search*, qui va tester une grille de valeurs possibles pour chaque paramètre. Cela permet de trouver le meilleur ensemble de paramètres qui vont optimiser les performances du modèle en réduisant son sur-ajustement. Compte tenu de la très grande taille de la base de données de cette étude et du temps de calcul nécessaire à tester tous les paramètres, il n'est pas possible d'utiliser une *grid-search* complète ici. En conséquent, les 3 *gradient boostings* suivants sont calibrés via la fonction gbm() du package `gbm` du logiciel R Core Team (2022), dont les performances sont comparées. Le choix des paramètres de ces trois *gradient boostings* est le suivant :

- 5000 arbres, taux d'apprentissage de 0.001, profondeur d'arbre de 3 (par la suite noté 5000/0.001/3) ;
- 5000 arbres, taux d'apprentissage de 0.01, profondeur d'arbre de 3 (par la suite noté 5000/0.01/3) ;
- 5000 arbres, taux d'apprentissage de 0.01, profondeur d'arbre de 4 (par la suite noté 5000/0.01/4).

Un algorithme avec 10000 arbres (10000/0.01/3) a également été calibré mais a demandé une semaine pour tourner sur le logiciel R Core Team (2022) et les résultats obtenus n'étaient pas meilleurs que ceux des 3 précédents algorithmes.

### Critère de déviance

Il convient de calculer les prédictions pour chaque observation annuelle via les 3 algorithmes de *gradient boosting* afin de calculer la déviance de chaque modèle, qui dans ce cas correspond à la somme des carrés des résidus (SCE). $SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, avec $y_i$ l'observation i et $\hat{y}_i$ la prédiction correspondante (tableau 3.9).

| | | 5000/0.01/3 | 5000/0.001/3 | 5000/0.01/4 |
|---|---|---|---|---|
| Année | Observée | Prédiction | Prédiction | Prédiction |
| 2005 | 0,40% | 0,60% | 0,55% | 0,58% |
| 2006 | 0,41% | 0,38% | 0,47% | 0,39% |
| 2007 | 0,40% | 0,37% | 0,47% | 0,37% |
| 2008 | 0,42% | 0,45% | 0,49% | 0,44% |
| 2009 | 0,39% | 0,35% | 0,47% | 0,35% |
| 2010 | 0,43% | 0,38% | 0,48% | 0,37% |
| 2011 | 0,44% | 0,53% | 0,54% | 0,54% |
| 2012 | 0,44% | 0,43% | 0,55% | 0,43% |
| 2013 | 0,44% | 0,44% | 0,55% | 0,44% |
| 2014 | 0,56% | 0,63% | 0,57% | 0,64% |
| 2015 | 0,63% | 0,63% | 0,66% | 0,62% |
| 2016 | 0,69% | 0,70% | 0,74% | 0,69% |
| 2017 | 0,71% | 0,95% | 0,97% | 0,96% |
| 2018 | 0,77% | 1,01% | 1,01% | 1,01% |
| 2019 | 0,82% | 1,02% | 1,01% | 1,02% |
| 2020 | 0,55% | 1,00% | 1,01% | 1,01% |
| 2021 | 0,44% | 1,03% | 1,01% | 1,03% |
| | SCE = | 7,76E-05 | 7,88E-05 | 7,85E-05 |

Table 3.9 : Récapitulatif des observations et des prédictions par année pour chaque modèle ainsi que la déviance (SCE) associée

Le modèle qui réduit le plus la déviance, c'est-à-dire celui dont les prédictions sont le plus proches des observations est le modèle 5000/0.01/3 avec un SCE = 7.76E-05 (les autres modèles ont des déviances très proches). Du point de vue de la déviance, il est plus judicieux de choisir ce modèle. Il est nécessaire maintenant de procéder à une validation croisée *k-folds* pour déterminer si ce modèle (et les deux autres) présente du sur-apprentissage.

### Critère de sur-apprentissage : validation croisée *k-folds*

La validation croisée *k-folds* diffère de la validation croisée *Train-Test Split* mais l'objectif reste le même, à savoir identifier la présence de sur-apprentissage des données. Pour ce faire, il faut diviser la base de données en k échantillons, puis un des échantillons est choisi comme ensemble de validation et les $k-1$ échantillons restants comme ensemble d'apprentissage. L'ensemble d'apprentissage est utilisé pour entraîner l'algorithme qui est testé sur l'ensemble de validation. Le processus est répété k fois pour que chaque échantillon soit une fois l'ensemble de validation.

Les résultats des 3 validations croisées sont analysés sur le logiciel R Core Team (2022) via la fonction gbm.perf() du package `gbm` où le test de validation croisée est spécifié. Cette fonction va tracer un graphique représentant l'erreur quadratique moyenne de la validation croisée en fonction du nombre d'arbres utilisés dans l'algorithme (plus il y a d'arbres, et plus l'algorithme est complexe). La courbe noire correspond à l'ensemble d'apprentissage, la courbe verte l'ensemble de validation et les pointillés indiquent le nombre d'arbres optimal pour ne pas être en sur-apprentissage (figure 3.13).

Pour l'algorithme 5000/0.001/3, la courbe verte continue de décroitre après les 5000 arbres. Cela veut dire qu'il faudrait davantage d'arbres pour réduire l'erreur quadratique moyenne de l'ensemble de validation. Cela est dû à un taux d'apprentissage trop faible (0.001) qui donne lieu à de très petites améliorations incrémentielles.

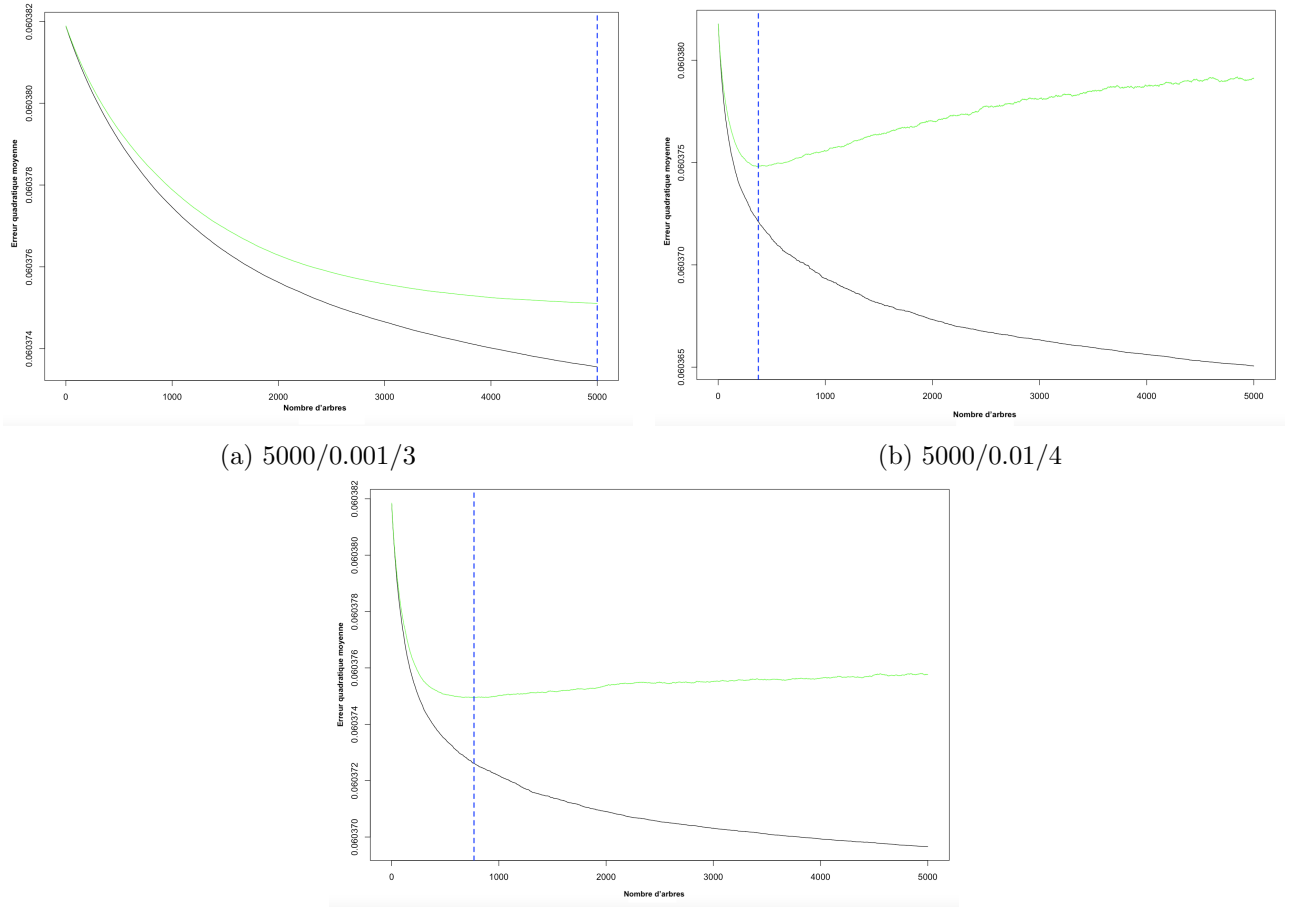
En ce qui concerne les algorithmes 5000/0.01/3 et 5000/0.01/4, ils obtiennent respectivement un nombre d'arbres optimal de 768 et de 375 (là où la courbe verte atteint son minimum). Les erreurs quadratiques correspondantes sont respectivement de 0.06037495 et de 0.0603748.

En d'autres termes, l'algorithme optimal en terme de non sur-apprentissage avec un taux d'apprentissage de 0.01 et une profondeur d'arbre de 3 doit avoir 768 arbres. Et l'algorithme optimal en terme de non sur-apprentissage avec un taux d'apprentissage de 0.01 et une profondeur d'arbre de 4 doit avoir 375 arbres. La différence du nombre d'arbres vient du fait que l'un des algorithmes a une profondeur supplémentaire par rapport à l'autre (4 vs 3) ce qui demande moins d'arbres à l'algorithme pour être performant.

### Choix de l'algorithme en fonction des critères de déviance et de validation croisée

En terme de déviance, l'algorithme 5000/0.01/3 est celui qui présente la déviance la plus faible. La validation croisée sur les deux algorithmes 5000/0.01/3 et 5000/0.01/4 présente une erreur quadratique moyenne identique pour respectivement 768 et 375 arbres, le nombre d'arbres optimal pour trouver le bon compromis entre biais et variance et ainsi éviter du sur-apprentissage. Les algorithmes 768/0.01/3 et 375/0.01/4 sont donc identiquement performants en terme de non sur-apprentissage des données.

C'est pourquoi il est nécessaire de recalculer les déviances des algorithmes 768/0.01/3 et 375/0.01/4. Il s'obtient respectivement SCE = 8.09E-05, et SCE = 7.78E-05. L'algorithme 375/0.01/4 présente une déviance plus faible.




(a) 5000/0.001/3

(b) 5000/0.01/4

(c) 5000/0.01/3

FIGURE 3.13 : Evolution de l'erreur quadratique moyenne en fonction du nombre d'arbres pour les 3 algorithmes de *gradient boosting* de l'ensemble d'apprentissage et de l'ensemble de validation

Il sera donc judicieux en cumulant ces informations de choisir l'algorithme **375/0.01/4** puisqu'il ne présente pas de sur-apprentissage et a la déviance la plus faible.

Via la fonction summary() de R CORE TEAM (2022), il est possible d'obtenir l'influence relative de chaque variable dans l'algorithme (figure 3.14) (BARSHAN et al. (2020)).

A chaque séparation de chaque arbre, la fonction gbm() calcule le critère d'amélioration MSE (Mean Squared Error, mesure la moyenne de la somme des carrés des erreurs entre la valeur prédite et la valeur réelle), puis calcule la moyenne de l'amélioration apportée par chaque variable sur tous les arbres où elle est utilisée. Plus une variable va améliorer l'erreur quadratique sur tous les arbres où elle est présente, plus son influence relative sera importante (H).

Dans l'algorithme 375/0.01/4, les variables ayant le plus d'influence sont le produit, l'âge et la couverture, mais tous les paramètres apportent leur utilité à l'algorithme.

Il est possible à présent de regarder le résultat du modèle pour le produit Personal Loan.

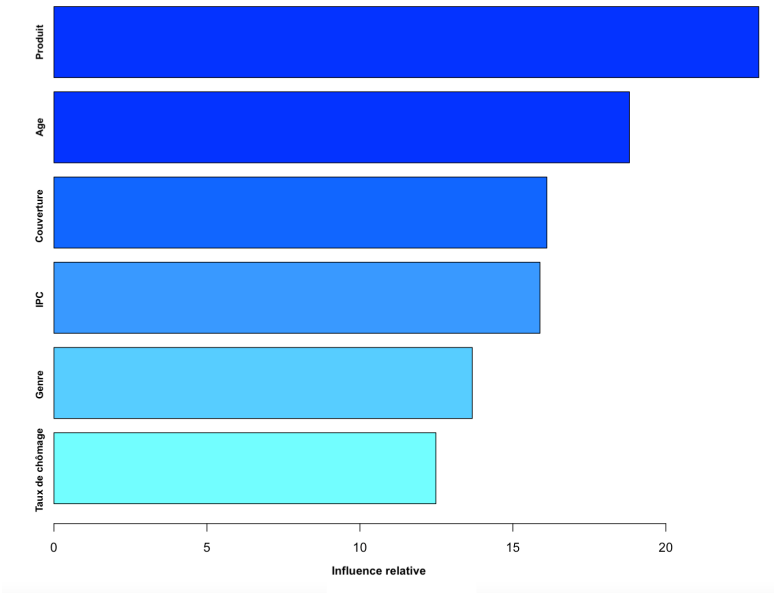



FIGURE 3.14 : Influence relative des différents paramètres du *gradient boosting* 375/0.01/4

### 3.3.4 Application du *gradient boosting* 375/0.01/4 au produit Personal Loan

#### Résultats du *gradient boosting* par couverture selon l'âge et le genre

Il est présenté dans le graphique du haut la fréquence prédite (courbe verte) par le *gradient boosting* pour la maladie comparée aux fréquences observées des hommes (courbe rouge pointillée) et des femmes (courbe bleue pointillée) selon l'âge, et dans le graphique du bas les fréquences prédites par le *gradient boosting* (courbes pleines rouges et bleues) pour l'accident en fonction du genre comparées aux observations masculines (courbe rouge pointillée) et féminines (courbe bleue pointillée) selon l'âge (figure 3.15).

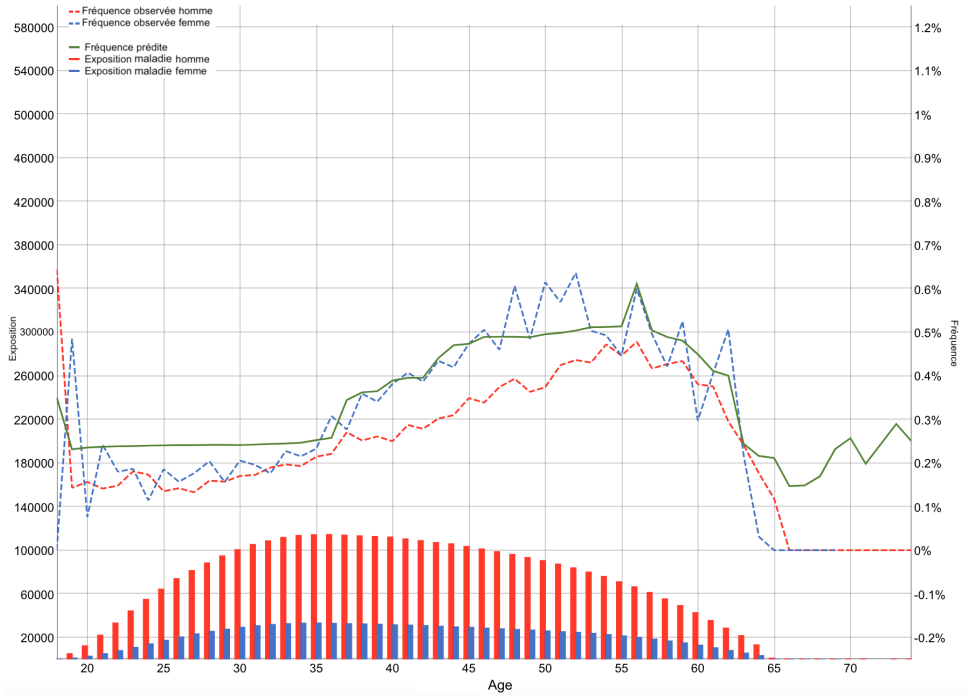
Pour la couverture maladie, la fréquence prédite suit bien la croissance observée par âge, en étant plus proche de la fréquence observée des femmes que des hommes.

En revanche pour la couverture accident, la fréquence prédite pour les hommes est bien alignée avec les observations mais ce n'est pas le cas de la prédiction des fréquences féminines. Celle-ci est bien supérieure aux observations et proche de celles masculines. Il a pu être vu que dans cet algorithme, l'influence relative du genre est l'une des plus faibles parmi les différents paramètres ce qui pourrait expliquer la mauvaise prise en compte du genre dans la prédiction de la fréquence accident pour les femmes. Cependant, il reste bien une fréquence prédite horizontale comme c'était le cas pour les précédentes modélisations GLM.

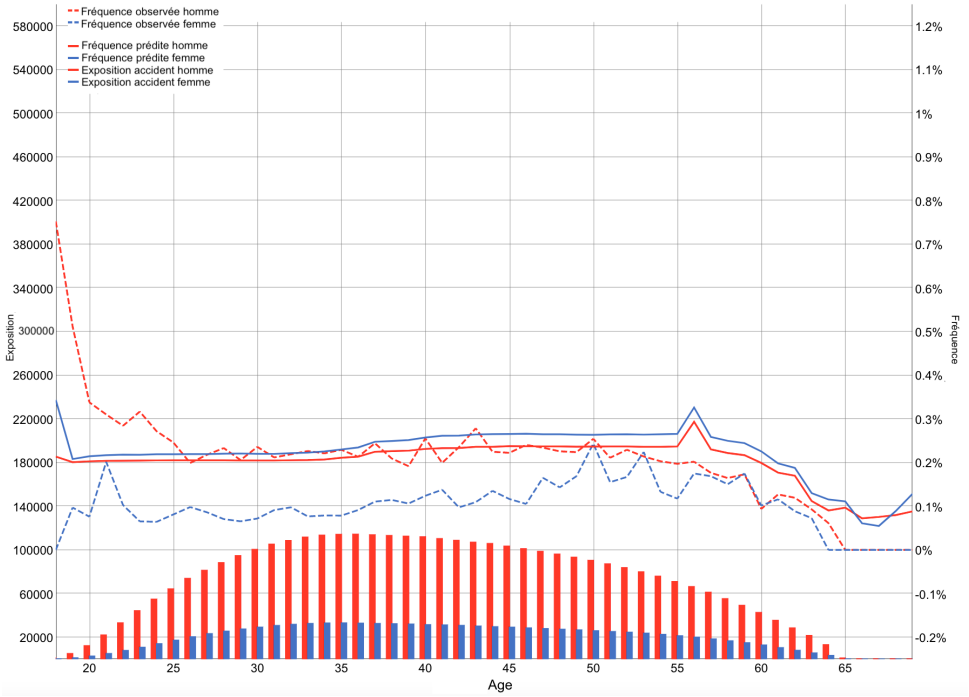
#### Résultats du *boosting* par année d'incident, maladie et accident ensemble

La courbe bleue correspond aux observations par année d'incident et la courbe verte la prédiction du *boosting* (figure 3.16).

Le *gradient boosting* prédit bien la période stable entre 2005 et 2013 ainsi que la croissance observée à partir de 2013. Cependant, à partir de l'année 2017, l'algorithme prédit une fréquence beaucoup plus




(a) Fréquences observées vs Fréquence prédite pour le produit Personal Loan en couverture maladie selon l'âge




(b) Fréquences observées vs Fréquences prédites pour le produit Personal Loan en couverture accident selon l'âge

FIGURE 3.15 : Fréquences observées vs Fréquences prédites pour le produit Personal Loan des deux couvertures

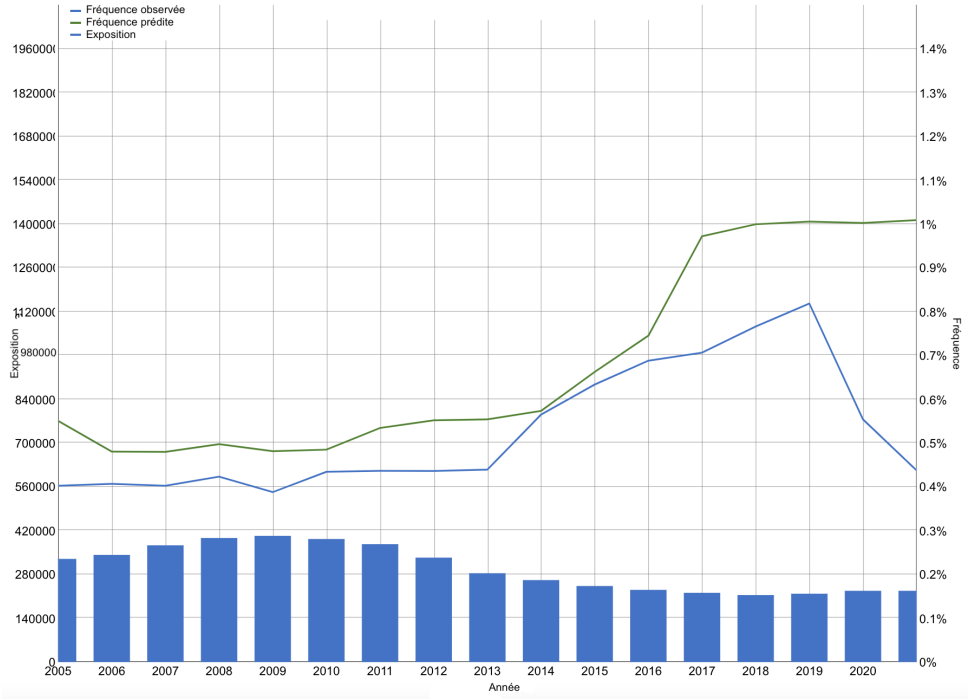



FIGURE 3.16 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Personal Loan

élevée (fréquence observée en 2019 = 0.82% VS fréquence prédite pour l'année 2019 = 1.01%). De plus, la fréquence prédite à partir de 2017 stagne là où la fréquence prédite par la deuxième modélisation GLM continue de croître 3.2.4 (il est rappelé que les données n'ont pas été entraînées sur la période 2020-2021).

L'algorithme de *gradient boosting* ne semble pas dans le cas du Personal Loan donner des résultats plus pertinents que la précédente modélisation ; les fréquences prédites que ce soit dans le cas couverture/âge ou par année ne sont pas totalement alignées avec les observations comme c'est davantage le cas avec la deuxième modélisation GLM. En revanche pour le produit Car Loan, le deuxième GLM prédit une croissance avec les années qui n'est pas observée (E.2.2), là où le *gradient boosting* prédit bien une fréquence stagnante comme les observations (G.2.2).

Ce qui départage ces deux algorithmes, c'est que le *gradient boosting* présente l'inconvénient de ne pas pouvoir être expliqué simplement à une personne ne connaissant pas le *machine learning* et les algorithmes de *gradient boosting*. Là où un GLM s'écrit via une fonction simple à comprendre prenant en compte les variables explicatives et le poids de chacune. De plus, son temps de calibration est beaucoup plus important.

En revanche, l'intérêt de cet algorithme de *gradient boosting* est qu'il va bien plus s'affiner avec le temps et l'ajout de nouvelles données à la base. Les résultats observés ici sont encourageants pour un nombre faible de variables explicatives (6 variables), une période d'observation de durée moyenne (15 années) et l'absence de *grid-search*. Il sera intéressant de refaire ce modèle ainsi que les GLMs dans quelques années afin de les comparer avec les modèles actuels, et voir lequel présente les meilleurs améliorations.

Comme pour les deux précédents modèles, l'exemple de calcul de la prime pure (unique et périodique

annuelle) est repris avec les fréquences obtenues via le *gradient boosting*.

### 3.3.5 Calcul de la prime pure pour le produit Personal Loan

Le cadre de l'exemple est le même que pour les deux précédentes modélisations (2.3). Les projections sur la période 2023-2027 des variables macroéconomiques taux de chômage et IPC sont utilisées. Il se remarque que la fréquence prédite par le *gradient boosting* reste la même chaque année. Voici un tableau récapitulatif (tableau 3.10).

| Année | 2023 | 2024 | 2025 | 2026 | 2027 |
|---|---|---|---|---|---|
| Exposition | 1000 | 1000 | 1000 | 1000 | 1000 |
| Taux de chômage | 5,34% | 5,12% | 5,01% | 5,19% | 5,37% |
| IPC | 113,9 | 114,6 | 115,6 | 117,2 | 119,1 |
| Fréquence | 1,009% | 1,009% | 1,009% | 1,009% | 1,009% |
| Nombre de sinistres | 10,1 | 10,1 | 10,1 | 10,1 | 10,1 |

TABLE 3.10 : Récapitulatif de l'exposition, du taux de chômage, de l'IPC, de la fréquence et du nombre de sinistres par année

Le montant du prêt est de 20000€, à rembourser sur 5 ans avec un taux d'intérêt de 5% soit un remboursement de 374.88€ par mois. Avec une durée moyenne d'incapacité de 6 mois, le coût total $C_i$ que l'assureur doit supporter par année i est égal à Nombre de sinistres l'année i×Remboursement mensuel du prêt× Durée de l'incapacité. Cela donne $C_{2023} = C_{2024} = C_{2025} = C_{2026} = C_{2027} = 22698.42$.

Par assuré, cela donne $c_{2023} = c_{2024} = c_{2025} = c_{2026} = c_{2027} = 22.70$€.

La valeur actualisée de ces coûts annuels est $VA = \sum_{i=2023}^{2027} \frac{c_i}{(1+0.75\%)^{i-2023}} = 111.81$€. Il en est déduit la prime pure unique qui est égale à la VA et la prime pure annuelle qui est égale au coût par assuré tous les ans (tableau 3.11).

| Prime pure unique | Prime pure annuelle |
|---|---|
| 111,81 € | 22,70 € |

TABLE 3.11 : Prime pure unique et prime pure annuelle

.

Les deux types de primes ont baissé avec cette nouvelle modélisation de la fréquence :

- la prime pure unique passe de 145.29€ (deuxième modèle GLM) à 111.81€ (\1.3) ;
- la prime pure annuelle passe de 29.64€ (deuxième modèle GLM) à 22.70€ (\1.3).

La différence importante entre les primes vient du fait que dans le deuxième GLM, la fréquence continue de croître durant la période 2023-2027. Alors que dans le *gradient boosting*, la fréquence stagne à partir de 2017 et ce jusqu'en 2027.

Voici un tableau récapitulatif des primes pures, unique et annuelle, obtenues par les trois modèles (tableau 3.12).

| | Prime pure unique | Prime pure annuelle |
|---|---|---|
| Modèle 1 : GLM âge/genre/produit | 91,26 € | 18,53 € |
| Modèle 2 : GLM modèle 1 + variables macroéconomiques | 145,29 € | 29,49 € |
| Modèle 3 : Gradient boosting 375/0.01/4 | 111,81 € | 22,70 € |

TABLE 3.12 : Récapitulatif des primes pures, unique et annuelle, calculées via les 3 modèles calibrés

Le modèle le plus prudent, celui qui propose les primes les plus élevées, est le GLM augmenté des variables macroéconomiques soit le deuxième modèle de l'étude. C'est également celui qui semble prédire le mieux les données observées au vu des graphiques comparatifs 3.10 et 3.11. Reste à savoir si ses prédictions pour les années futures ne sont pas trop surestimées et impliquent des propositions de primes pures largement supérieures à ce qu'elles devraient être.

Au vu de l'incertitude liée au futur fréquences, il serait intéressant d'étudier les primes *burning-cost* et de les comparer aux précédentes primes pures. Cela consiste à calculer la prime d'assurance nécessaire pour éliminer progressivement le risque au fil du temps, plutôt que de le couvrir entièrement à un moment donné.

## 3.4 Comparaison et limite des trois modélisations

### 3.4.1 Comparaison des trois modélisations

Les trois modélisations proposent une prédiction différente de la fréquence pour les années futures pour les produits Personal Loan et Mortgage Loan. La première modélisation GLM et le *gradient boosting* prédisent chacun une fréquence stagnante à des niveaux différents. La deuxième modélisation GLM propose une fréquence qui continue de croître.

Il est possible de comparer ces trois modèles via différents indicateurs : l'AIC et le BIC (2.2.2) pour les deux modélisations GLM ainsi que l'erreur quadratique moyenne MSE (3.3.3) pour les trois modèles. Le modèle qui réduit le plus l'AIC, le BIC et le MSE est considéré comme celui qui a une meilleure qualité d'ajustement aux données compte tenu de sa complexité. Le MSE calculé ici est celui pour le nombre de sinistres. Le *gradient boosting* étant calibré pour prédire la fréquence, il faut calculer en amont le nombre de sinistres correspondants via l'exposition. De plus, par soucis de comparaison, l'étude du MSE du *gradient boosting* est séparée entre maladie et accident comme pour les GLMs (tableau 3.13).

| | GLM | | GLM augmenté | | Gradient Boosting | |
|---|---|---|---|---|---|---|
| | Accident | Maladie | Accident | Maladie | Accident | Maladie |
| AIC | 148 999 | 211 450 | 148 636 | 210 497 | | |
| BIC | 149 052 | 211 503 | 148 716 | 210 578 | | |
| MSE | 0,002527 | 0,003729 | 0,002526 | 0,003728 | 0,002543 | 0,003735 |

TABLE 3.13 : Valeurs obtenues des indicateurs AIC, BIC et MSE pour les trois modélisations

Le GLM augmenté des variables explicatives macroéconomiques est celui qui réduit le plus les critères AIC, BIC et MSE pour les deux couvertures. Cependant les résultats obtenus sont très proches de ceux des autres modèles.

Afin de départager les trois modèles, ceux-ci sont comparés selon différents critères : qualité prédictive,

temps de calibration, interprétabilité et prudence tarifaire, avec une notation allant de - - - (mauvais) à +++ (excellent) (tableau 3.14).

| | Qualité prédictive | Temps de calibration | Interprétabilité | Prudence tarifaire |
|---|---|---|---|---|
| Modèle 1 : GLM âge/genre/produit | "-" : non prise en compte de la hausse de la fréquence avec les années, pas de sur-apprentissage | "++" ≃2 minutes | "+++" : très intuitif, facilement explicable à une personne ne connaissant pas le principe d'un GLM | "+": modèle parmi les trois qui prédit la fréquence la plus faible pour le produit Personal Loan et donc les primes pures les plus faibles, mais reste prudent car la fréquence prédite est alignée avec la fréquence la plus haute observée |
| Modèle 2 : GLM modèle 1 + variables macroéconomiques | "++" : prise en compte de la hausse de la fréquence avec les années mais modélise une hausse pour le produit Car Loan qui n'en a pas, pas de sur-apprentissage, meilleurs résultats pour les indicateurs AIC, BIC et MSE | "++" ≃7 minutes | "+++" : très intuitif, facilement explicable à une personne ne connaissant pas le principe d'un GLM | "+++": prédit une hausse de la fréquence pour le produit Personal Loan dans le futur ce qui conduit aux primes pures les plus élevées des trois modèles |
| Modèle 3 : Gradient boosting 375/0.01/4 | "++" : prise en compte de la hausse de la fréquence avec les années sauf pour le produit Car Loan qui n'en a pas, mais prédiction surestimé à partir de 2017, pas de sur-apprentissage | "- - -" : ≃ 1,5 semaine pour calibrer trois boostings avec validation croisée | "- -" : peu compréhensible pour les personnes non familiarisées avec le machine learning et le boosting | "++": prédit une fréquence stagnante pour le produit Personal Loan supérieure à la fréquence stagnante du premier modèle GLM et inférieure à celle de la seconde modélisation GLM ce qui conduit à des primes pures prudentes |

TABLE 3.14 : Comparatif des 3 modèles selon différents critères

Il se remarque qu'à tous les niveaux, le deuxième modèle GLM, celui enrichi des variables explicatives macroéconomiques, est celui qui présente les meilleurs résultats : prédiction pertinente, temps de calcul optimal, interprétabilité intuitive et primes pures les plus élevées pour l'exemple du produit Personal Loan.

Le modèle de *gradient boosting* présente également de bons résultats prédictifs et des primes pures prudentes dans l'exemple du produit Personal Loan. Il modélise également bien la fréquence du produit Car Loan, là où la deuxième modélisation GLM prédit une croissance avec les années non observée. En revanche son temps de calcul reste beaucoup trop long, sachant que pour cette étude seulement trois algorithmes de *boosting* ont été implémentés. Normalement, il faudrait en faire tourner davantage en faisant varier les hyperparamètres de manière à pouvoir sélectionner le meilleur, ce qui n'est pas le cas lors de la calibration d'un GLM. De plus son interprétabilité est difficile, de l'explication de son fonctionnement à son utilisation.

### 3.4.2 Limite des trois modélisations

De nombreux éléments viennent limiter la pertinence des modélisations réalisées dans cette études, des données de départ aux choix des modèles. Concernant les données, deux principaux problèmes se sont présentés. Tout d'abord la faible quantité de variables explicatives à disposition avant l'ajout des variables macroéconomiques (seulement trois variables, l'âge, le genre et le produit) qui induit des modèles très simples manquant de profondeur. L'autre problème observé est la très grande quantité de valeurs zéros dans les données. Cela à un impact direct sur la modélisation GLM Poisson qui n'est pas adaptée et réduit la fiabilité statistique des résultats obtenus. En conséquent, les modèles ont

davantage été comparés d'un point de vu graphique, entre eux et avec les observations. Concernant le *gradient boosting*, son intérêt principal de gérer une grande quantité de données n'a pas été pris en compte avec l'ajout de seulement trois variables explicatives supplémentaires (6 au total). Cela est dû à un volume conséquent de données qui en terme de temps de calcul empêche cet ajout d'autres variables ainsi que de procéder à une *grid-research* pour obtenir le modèle le plus adapté. En outre, le choix des variables macroéconomiques peut être sujet à débat. L'augmentation des fréquences est bien liée à la prise de conscience des Portugais des produits d'assurance qu'ils ont à disposition, mais l'explication de cette prise de conscience via des questions économiques n'est pas forcément évidente. En effet le choix de modéliser un risque incapacité via des variables macroéconomiques reste atypique voir inédit.

### Conclusion du chapitre 3

L'ajout de variables macroéconomiques a permis de prendre en compte les hausses de fréquences observées pour les produits Personal Loan et Mortgage Loan. Seulement la nouvelle modélisation GLM et le *gradient boosting* ne prédisent pas de la même manière l'évolution future de la fréquence par année. Pour le GLM, celle-ci va continuer de croître, là où le *gradient boosting* prédit une fréquence stagnante. Cela se traduit par des primes pures futures plus élevées pour le modèle GLM.

Tous les modèles de cette étude ont finalement été comparés, donnant avantage au GLM augmenté des variables explicatives. Cependant les limites observées pour ce modèle et les autres induisent de prendre les résultats avec précaution.

# Conclusion

Cette étude porte sur la pertinence du choix d'un modèle linéaire généralisé pour modéliser la fréquence incapacité. L'objectif premier de cette étude est d'analyser la qualité prédictive du GLM en gardant les contraintes propre à l'entreprise, à savoir un nombre limité de variables explicatives (âge, genre et produit) ; puis de se libérer de ces contraintes pour proposer un GLM plus fourni en variables explicatives (ajout de variables macroéconomiques) en lien avec les explications de la hausse de la fréquence avec les années observée pour les produits Personal Loan et Mortgage Loan. Finalement, les résultats de ces deux modélisations et leur pertinence sont comparées à un algorithme de *gradient boosting*.

Le premier GLM sous contraintes a donné de mauvais résultats. La fréquence prédite par année est inférieure aux observations pour les produits Personal Loan et Mortgage Loan (76% du portefeuille) et la hausse de la fréquence observée avec les années pour ces produits n'a pas été prise en compte. Des ajustements de la fréquence sont nécessaire en interne pour proposer des niveaux de fréquences plus hauts. Le mauvais ajustement de ce GLM vient du faible nombre de variables explicatives qui n'expliquent pas cette hausse liée à des questions économiques. De plus le choix d'un GLM Poisson présente des limites. Il a été choisi pour compter un nombre de sinistres alors que ceux-ci présentent dans les données un très grand nombre de valeurs zéros. Cela a impliqué un échec des tests statistiques sur la variable réponse et sur les résidus nécessaires à confirmer la pertinence du modèle choisi. Malgré tout, le choix d'une autre distribution (binomiale négative) ou d'un modèle plus adapté à ce type de données (modèle ZIP) n'ont pas apporté plus de profondeur à la modélisation.

Le problème premier étant les contraintes internes citées ci-dessus, cette étude regarde dans un second temps ce qu'il se passe lorsque ces contraintes sont levées et que de nouvelles variables explicatives macroéconomiques sont ajoutées au modèle. Le nouveau GLM a pu modéliser les hausses de fréquences par année des produits Personal Loan et Mortgage Loan et s'aligner avec les observations. Malgré tout le GLM Poisson reste mal adapté avec toutes les valeurs zéros et les résultats des tests statistiques sur les résidus montrent encore ses limites. De plus, la hausse de la fréquence prédite liée aux variables macroéconomiques a été également prise en compte pour le produit Car Loan qui n'en a pas. Cela est lié à la structure du GLM qui explique la fréquence avec les mêmes coefficients pour l'âge, le genre, le taux de chômage et l'IPC par produit. Il serait intéressant de faire dans ce cas un GLM par produit.

Il était donc intéressant de regarder un autre type de modélisation et de le comparer aux GLMs. Le choix s'est porté sur un modèle de *machine learning* non-linéaire très puissants qui est le *gradient boosting*. Celui-ci a pu prédire une hausse de la fréquence avec les années pour les produits Personal Loan et Mortgage Loan et une fréquence stagnante pour le produit Car Loan qui est alignée avec les observations. De plus la validation croisée *k-folds* a pu spécifié pour un taux d'apprentissage et une profondeur d'arbre donnés le nombre d'arbres optimal malgré la non implémentation d'une *grid-search*. Le modèle en lui-même et les résultats de la prédiction des fréquences présentent donc un grand intérêt, mais de nouvelles limites apparaissent. Les temps de calibration du *gradient boosting* sont beaucoup

trop longs, malgré un faible nombre de variables explicatives (l'âge, le genre, la couverture, le produit, le taux de chômage, l'IPC) et pas d'implémentation de *grid-search*. De plus cela reste un algorithme difficile à expliquer et qui ne peut pas être généralisé aussi facilement qu'un GLM.

C'est pourquoi après comparaison, cette étude considère le deuxième modèle GLM, celui enrichi des variables explicatives macroéconomiques, comme le plus probant. Il est celui qui présente le meilleur compromis entre qualité prédictive, temps de calibration, interprétabilité et prudence tarifaire. Le GLM reste donc un outil pertinent pour calibrer la fréquence incapacité.

Malgré tout, si le temps de calibration des algorithmes de *gradient boosting* peut-être amélioré, avec notamment le développement d'outils informatiques plus puissants (les ordinateurs quantiques dans un futur plus ou moins proche par exemple), et les connaissances des individus pour le *machine learning* approfondies, les modèles de *gradient boosting* s'imposeront petit à petit et remplaceront les modèles classiques en assurance (dont les GLMs).

De manière générale, les résultats obtenus dans cette étude ont le mérite de faire réfléchir les compagnies, ici AXA Partners, sur le choix des variables explicatives et des modèles à implémenter pour modéliser les fréquences ; et de peut-être développer de nouveaux outils de tarification pour prendre en compte des modèles plus complexes. Il ne faut cependant pas oublier l'importance de garder des modèles en adéquation avec la volonté de généralisation et de simplification de la compagnie. Il serait intéressant d'analyser l'intérêt économique de modifier les outils de tarification d'AXA Partners pour prendre en compte ces modèles plus complexes et moins généralistes.

Enfin, il serait pertinent de voir si les éléments amenés dans cette étude peuvent s'extrapoler à d'autres compagnies d'assurances et d'autres branches de l'assurance, à savoir de prendre en compte davantage de variables explicatives dans les outils internes et d'utiliser davantage de modèles de *machine learning* comme le *gradient boosting*. Il est évident que l'enrichissement des bases de données ne doit pas être laissé de côté lorsque de tels outils existent pour les exploiter. C'est déjà le cas, notamment chez AXA pour lutter contre la fraude à l'assurance des activités IARD par exemple, mais cela reste encore très marginal.

# Bibliographie


Azencott, C.-A. (2018). Introduction au Machine Learning. Dunod.

Barshan, E., Brunet, M.-E. et Dziugaite (2020). RelatIF: Identifying Explanatory Training Examples via Relative Influence. *International Conference on Artificial Intelligence and Statistics*.

DataScientest (2019). Machine Learning. url : https://datascientest.com/machine-learning-tout-savoir.

McCullagh, P. et Nelder, J. A. (1989). Generalized Linear Models. Second Edition. Statistics and Applied Probability 37. Boca Raton : Chapman et Hall.

Midi, H., Sarkar, S. K. et Rana, S. (2010). Binary response modeling and validation of its predictive ability. *WSEAS Transactions on Mathematics* 9(6).438-447.

Monnier, D. (2016). Modèles linéaires généralisés et assurance santé individuelle : Tarification et évaluation des engagements sous solvabilité II. Mémoire d'actuariat. ISUP. url : http://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/0/b47ec3b5ed859109c12580f900205a7b/$FILE/MONNIER.002.pdf/MONNIER.pdf.

Natekin, A. et Knoll, A. (2013). Gradient boosting machines, a tutorial. *frontiers in Neurobotics* 7.21.

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions TECHNIP.

SAS (1985). SAS user's guide: Statistics. Sas Inst.

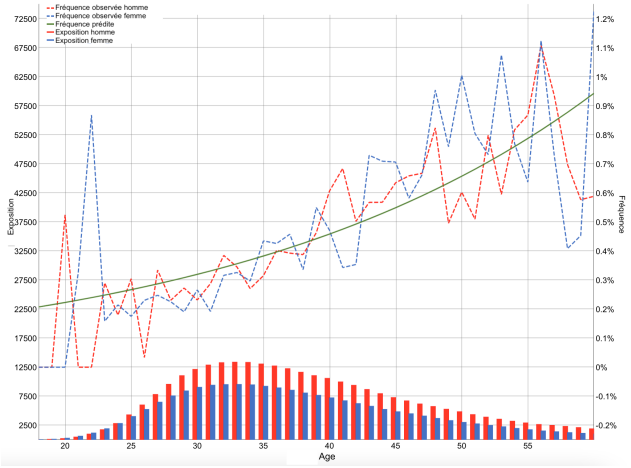
Steele, M. C. (2003). The Power of Categorical Goodness-Of-Fit Statistics. Thèse de doct. Australian School of Environmental Studies.

# Annexe A

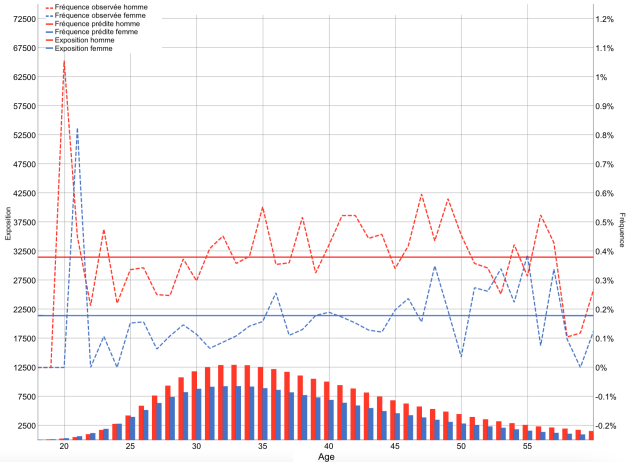
# Résultats de la 1re modélisation GLM pour les autres produits

## A.1 Produit Mortgage Loan

### A.1.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite pour le produit Mortgage Loan en couverture maladie selon l'âge




(b) Fréquences observées vs Fréquences prédites pour le produit Mortgage Loan en couverture accident selon l'âge

FIGURE A.1 : Fréquences observées vs Fréquences prédites par âge pour le produit Mortgage Loan des deux couvertures

Les prédictions sont alignées avec les observations par âge pour les deux couvertures (figure A.1).

### A.1.2 Résultats par année d'incident, maladie et accident ensemble

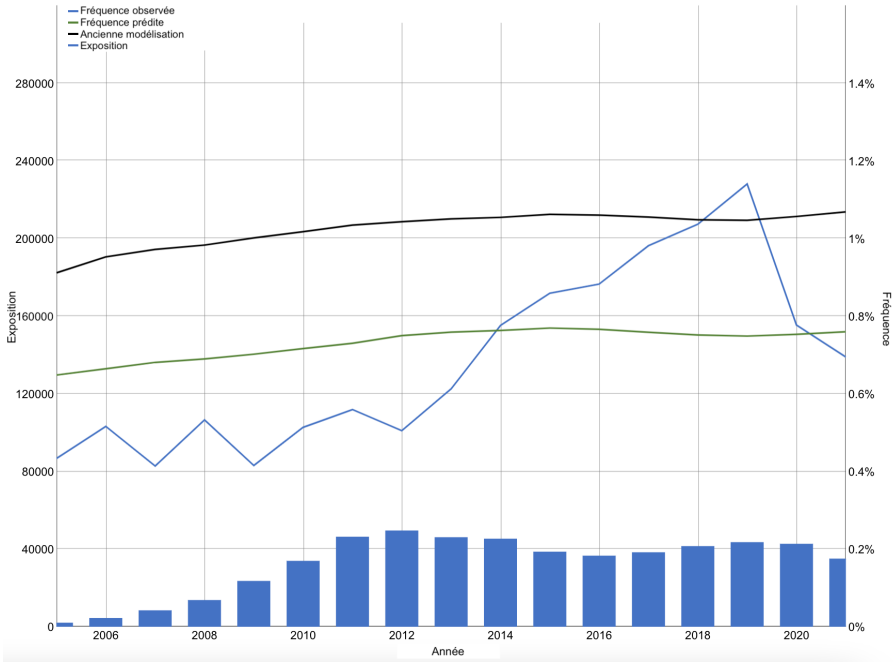
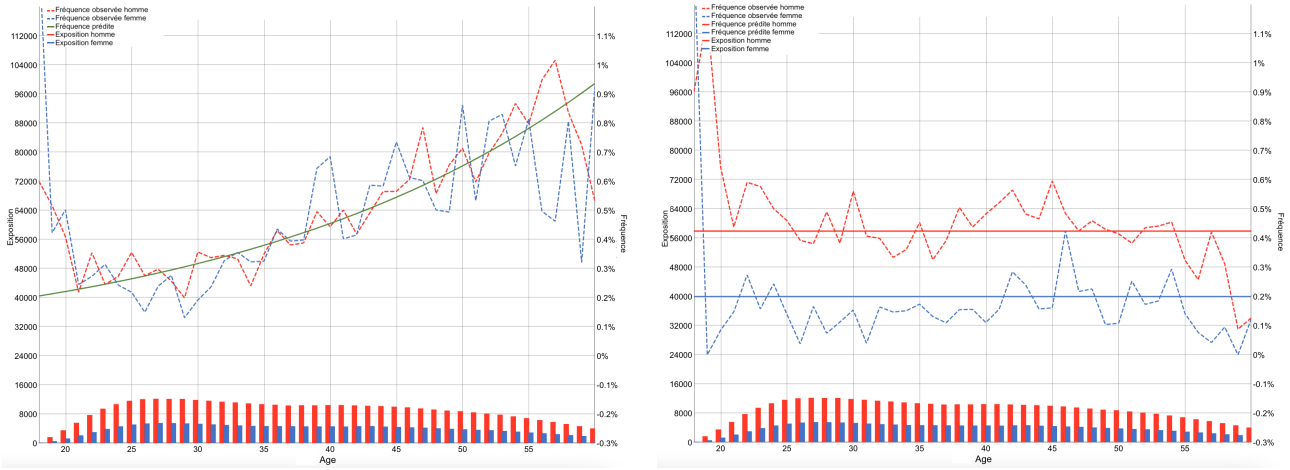



FIGURE A.2 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Mortgage Loan

La prédiction en vert est bien inférieure aux observations à partir de 2014. De plus la hausse de la fréquence avec les années n'est pas modélisée par la modélisation GLM (figure A.2).

## A.2 Produit Car Loan

### A.2.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite pour le produit Car Loan en couverture maladie selon l'âge

(b) Fréquences observées vs Fréquences prédites pour le produit Car Loan en couverture accident selon l'âge

FIGURE A.3 : Fréquences observées vs Fréquences prédites par âge pour le produit Car Loan des deux couvertures

Les prédictions sont alignées avec les observations par âge pour les deux couvertures (figure A.3).

### A.2.2 Résultats par année d'incident, maladie et accident ensemble

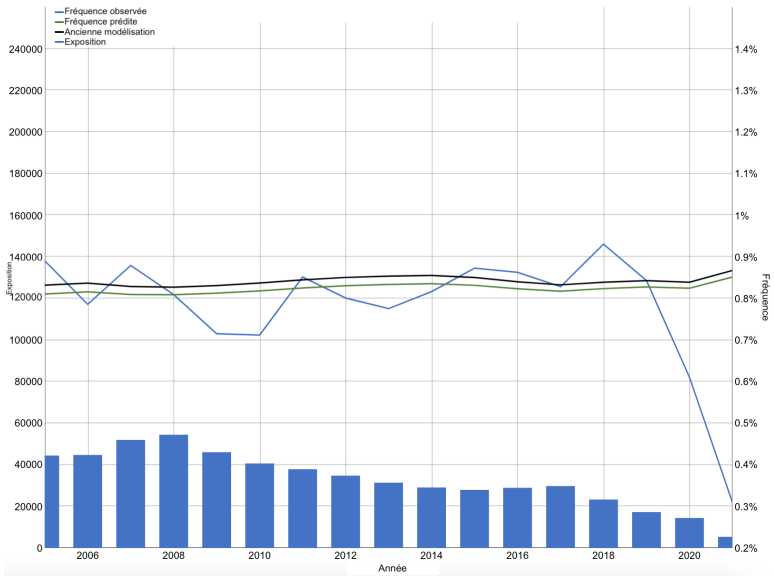



FIGURE A.4 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Car Loan

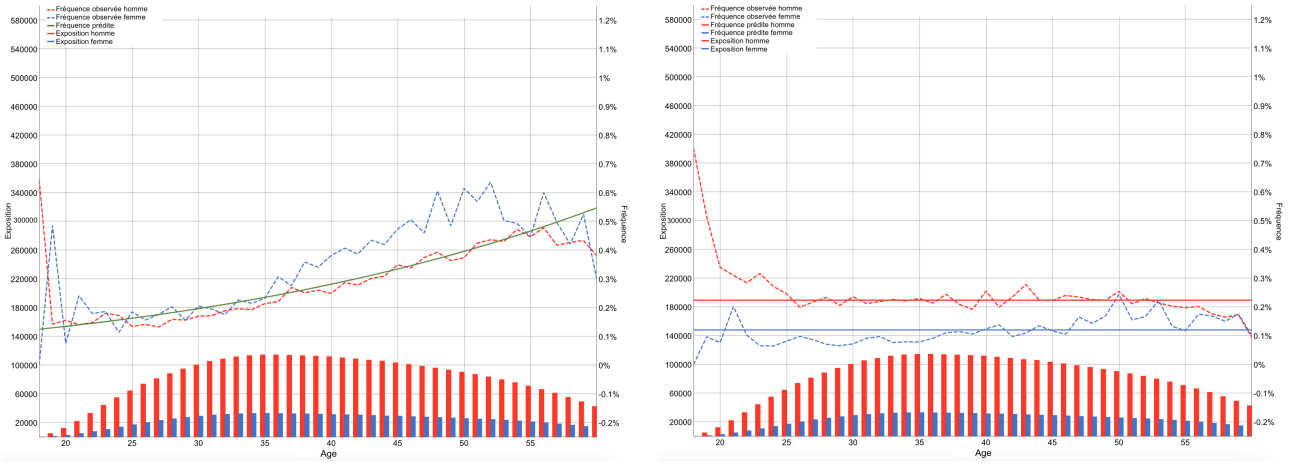
La prédiction en vert est très proche de l'ancienne modélisation et des fréquences observées (figure A.4).

# Annexe B

# Résultats de la modélisation ZIP pour tous les produits

## B.1 Produit Personal Loan

### B.1.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite pour le produit Personal Loan en couverture maladie selon l'âge

(b) Fréquences observées vs Fréquences prédites pour le produit Personal Loan en couverture accident selon l'âge

FIGURE B.1 : Fréquences observées vs Fréquences prédites par âge pour le produit Personal Loan des deux couvertures

Les prédictions sont alignées avec les observations par âge pour les deux couvertures (figure B.1).

### B.1.2 Résultats par année d'incident, maladie et accident ensemble

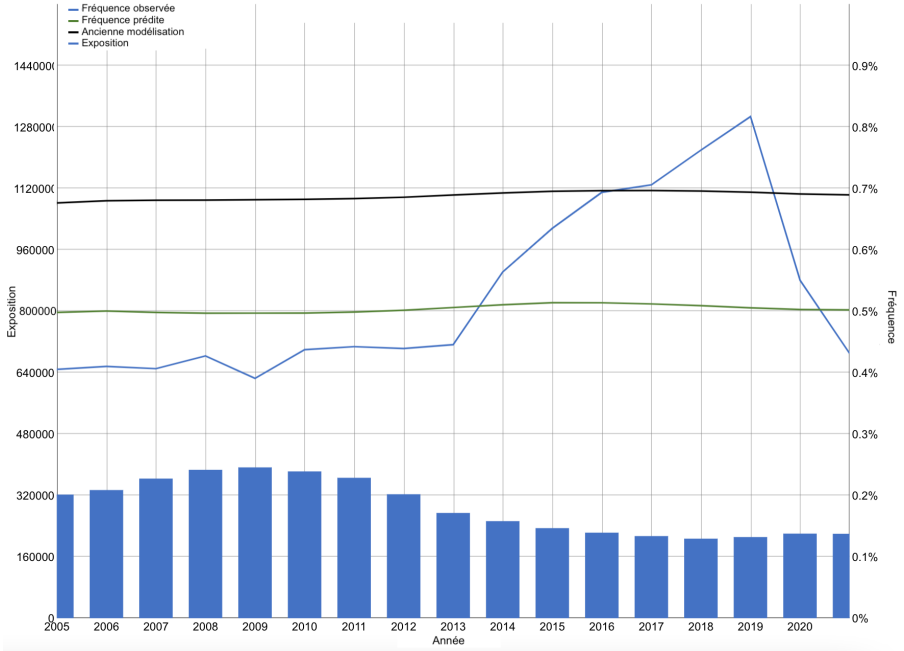
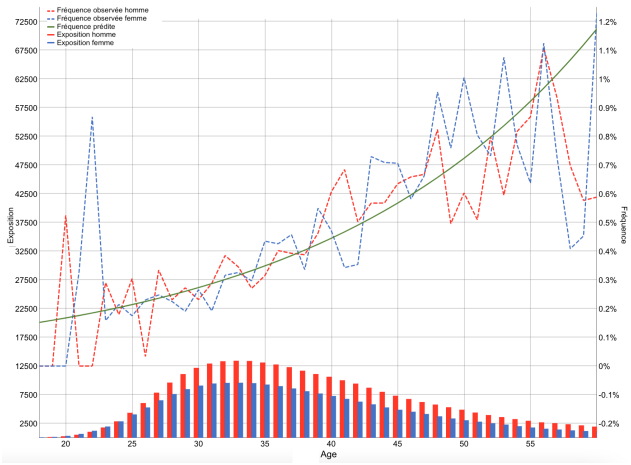



FIGURE B.2 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Personal Loan

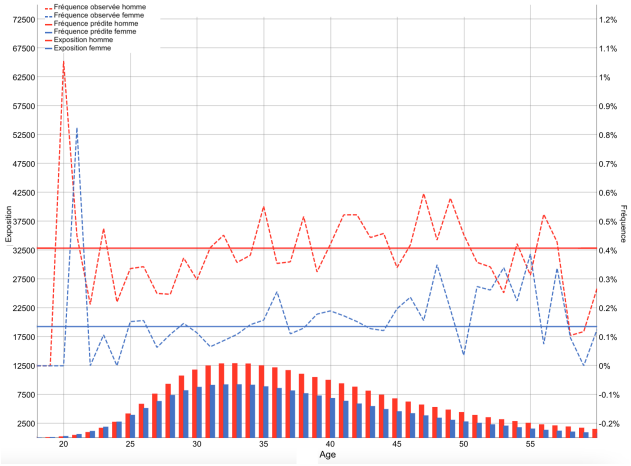
La prédiction en vert est bien inférieure aux observations à partir de 2013-2014. De plus la hausse de la fréquence avec les années n'est pas modélisée par la modélisation ZIP. La fréquence prédite est très proche de celle de la modélisation GLM avant ajustements (figure B.2).

## B.2 Produit Mortgage Loan

### B.2.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite pour le produit Mortgage Loan en couverture maladie selon l'âge




(b) Fréquences observées vs Fréquences prédites pour le produit Mortgage Loan en couverture accident selon l'âge

FIGURE B.3 : Fréquences observées vs Fréquences prédites par âge pour le produit Mortgage Loan des deux couvertures

Les prédictions sont alignées avec les observations par âge pour les deux couvertures (figure B.3).

### B.2.2 Résultats par année d'incident, maladie et accident ensemble

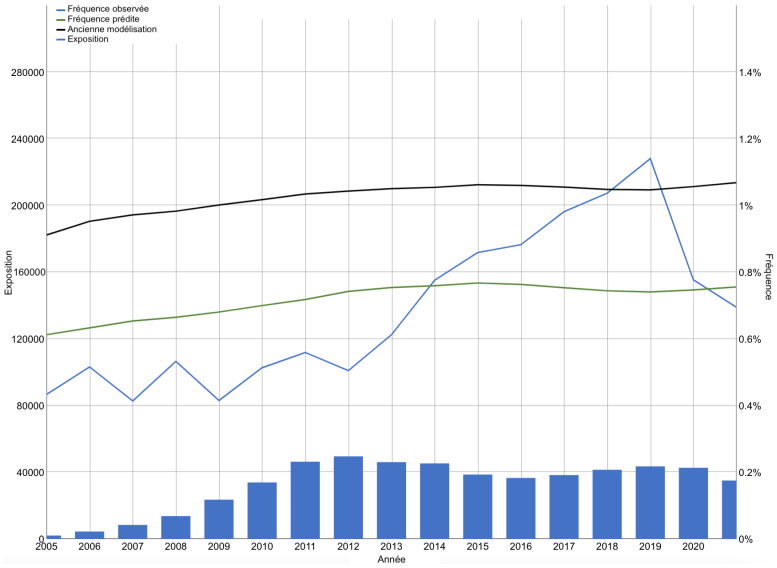
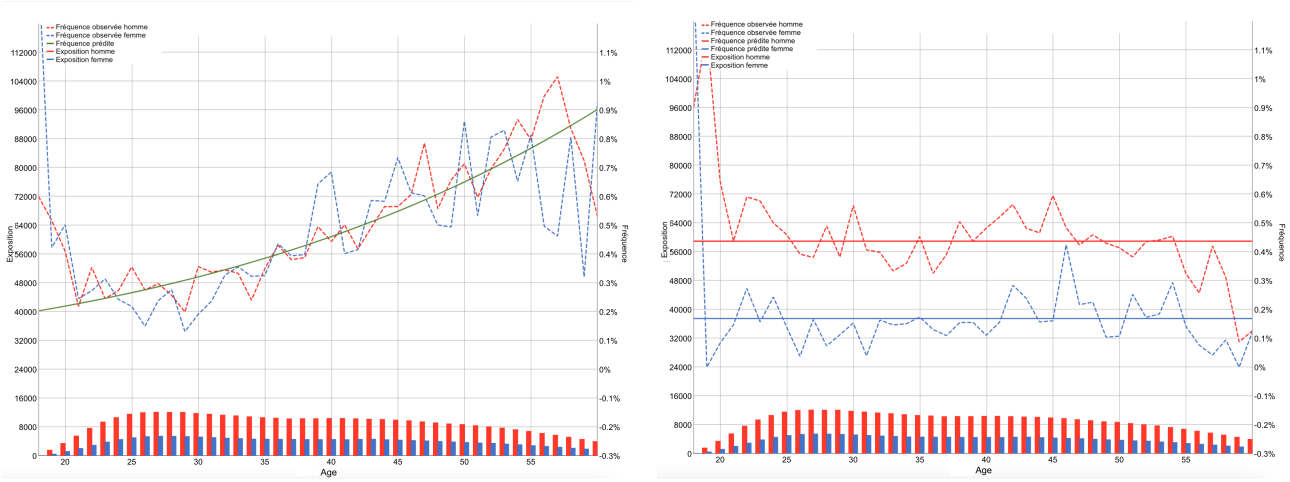



FIGURE B.4 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Mortgage Loan

La prédiction en vert est bien inférieure aux observations à partir de 2013-2014. De plus la hausse de la fréquence avec les années n'est pas modélisée par la modélisation ZIP. La fréquence prédite est très proche de celle de la modélisation GLM avant ajustements (figure B.4).

## B.3 Produit Car Loan

### B.3.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite pour le produit Car Loan en couverture maladie selon l'âge

(b) Fréquences observées vs Fréquences prédites pour le produit Car Loan en couverture accident selon l'âge

FIGURE B.5 : Fréquences observées vs Fréquences prédites par âge pour le produit Car Loan des deux couvertures

Les prédictions sont alignées avec les observations par âge pour les deux couvertures (figure B.5).

### B.3.2 Résultats par année d'incident, maladie et accident ensemble

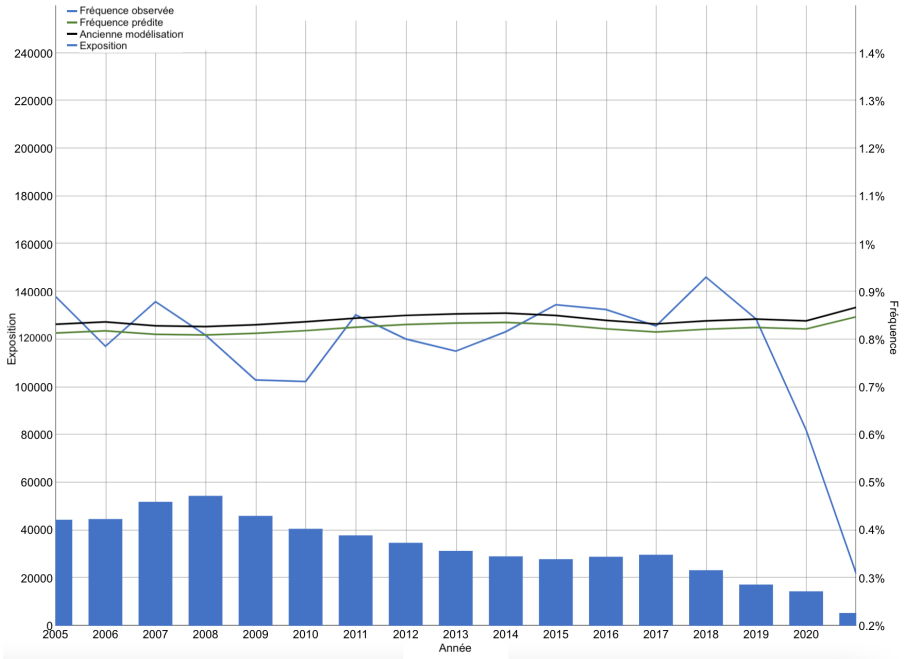



FIGURE B.6 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Car Loan

La prédiction en vert est très proche de l'ancienne modélisation et des fréquences observées (figure B.6).

# Annexe C

# Résultats de la modélisation GLM avec une distribution binomiale négative

## C.1 Coefficients des GLMs

| Nom de la variable explicative | Coefficient associé | p-value |
|---|---|---|
| Produit Car Loan (et intercept) | -6,83251 | <2E-16 |
| Produit Mortgage Loan | 0,00457 | 0,878 |
| Produit Personal Loan | -0,49588 | <2E-16 |
| Âge | 0,03607 | <2E-16 |

(a) Variables, coefficients et p-values obtenus pour le GLM de la couverture maladie

| Nom de la variable explicative | Coefficient associé | p-value |
|---|---|---|
| Produit Car Loan (et intercept) | -5,53819 | <2E-16 |
| Produit Mortgage Loan | -0,18432 | 0,878 |
| Produit Personal Loan | -0,39363 | 9,55E-07 |
| Genre femme | -0,45405 | 3,75E-10 |

(b) Variables, coefficients et p-values obtenus pour le GLM de la couverture accident

TABLE C.1 : Variables, coefficients et p-values obtenus pour les GLMs

Dans les deux GLMs maladie et accident, les variables choisies sont statistiquement significatives sauf la variable du produit Mortage Loan. Mais dans une volonté de garder cette variable pour expliquer la fréquence par produit, elle est gardée dans la modélisation. En outre, les coefficients obtenus sont très proches de ceux obtenus avec une distribution de Poisson (tableau C.1).

## C.2 Anova()

| Variable explicative | Degré de liberté | Déviance | Degré de liberté résiduel | Déviance résiduelle | p-value |
|---|---|---|---|---|---|
| Produit | 2 | 533,45 | 5059523 | 148419 | <2.2E-16 |
| Âge | 1 | 2252,99 | 5059522 | 146166 | <2.2E-16 |

(a) Résultat de la fonction anova() du logiciel R CORE TEAM (2022) avec test du $\chi^2$ pour analyser le pouvoir explicatif de chaque variable dans le GLM maladie

| Variable explicative | Degré de liberté | Déviance | Degré de liberté résiduel | Déviance résiduelle | p-value |
|---|---|---|---|---|---|
| Produit | 2 | 31,471 | 5050501 | 5928,9 | 1,47E-07 |
| Genre | 1 | 36,534 | 5050500 | 5892,4 | 1,50E-09 |

(b) Résultat de la fonction anova() du logiciel R CORE TEAM (2022) avec test du $\chi^2$ pour analyser le pouvoir explicatif de chaque variable dans le GLM accident

TABLE C.2 : Résultat de la fonction anova() du logiciel R CORE TEAM (2022) avec test du $\chi^2$ pour analyser le pouvoir explicatif de chaque variable dans les GLMs

Les résultats obtenus mènent aux mêmes conclusions que les GLMs avec distribution Poisson. Les variables explicatives choisies réduisent suffisamment la déviance pour être significatives dans les GLMs (tableau C.2).

Le graphique de la modélisation de la fréquence par année pour le produit Personal Loan est analysé.

### C.2.1 Résultats par année d'incident, maladie et accident ensemble

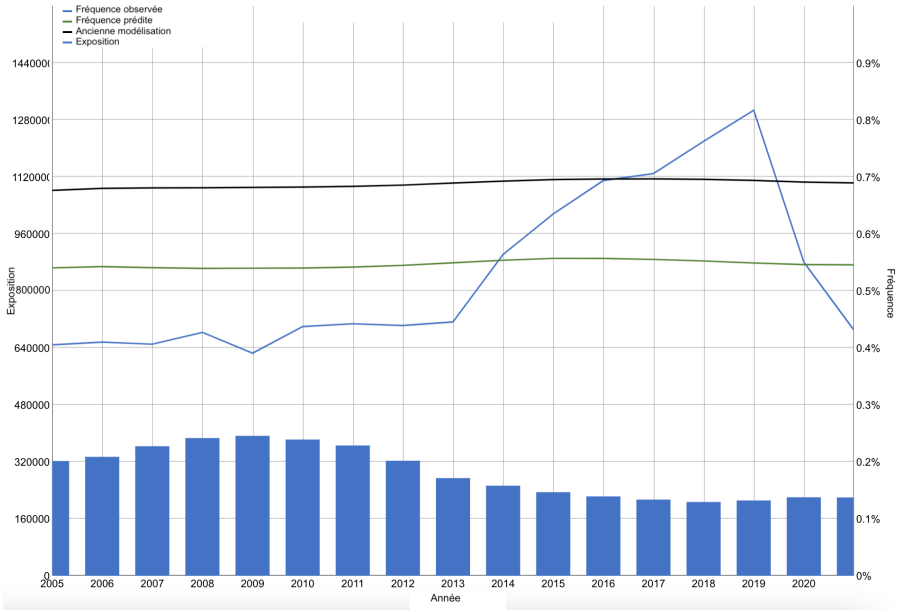



FIGURE C.1 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Personal Loan

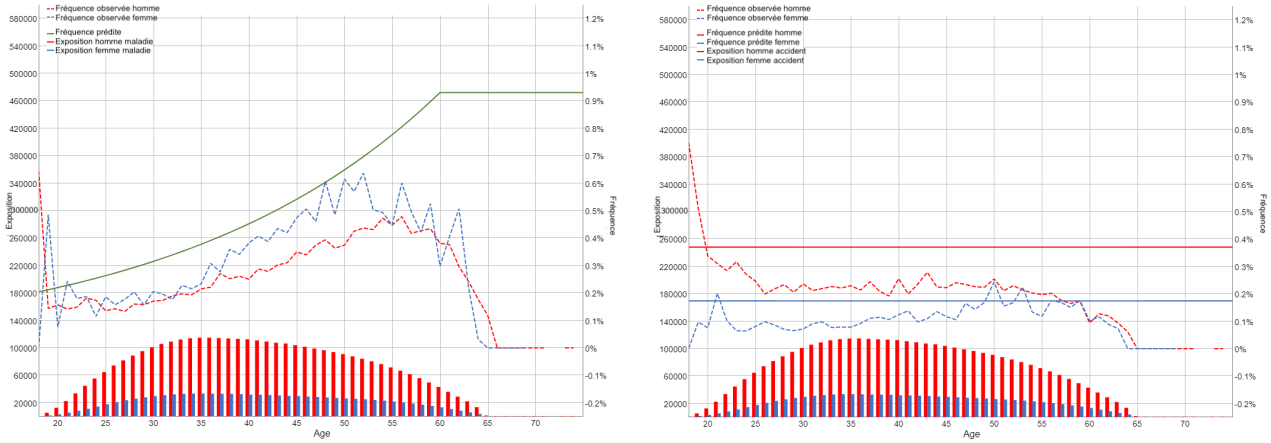
La fréquence obtenue est légèrement supérieure à celle du modèle avec distribution Poisson, mais la stagnation avec les années reste la même (figure C.1).

# Annexe D

# Résultats de la 1re modélisation GLM après ajustement

## D.1 Produit Personal Loan

### D.1.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite par âge pour le produit Personal Loan en couverture maladie après modification du GLM

(b) Fréquences observées vs Fréquences prédites par âge pour le produit Personal Loan en couverture accident après modification du GLM

FIGURE D.1 : Fréquences observées et fréquences prédites par âge après modification du GLM pour le produit Personal Loan

L'augmentation du coefficient de la variable "Personal Loan" se remarque sur ces deux graphiques. Les courbes prédites se retrouvent pour les deux couvertures bien au-dessus des observations. **En voulant prendre en compte la fréquence importante de l'année 2019 que les GLMs n'ont pas modélisée, l'ajustement éloigne le résultat des GLMs de la réalité observée** (figure D.1).

### D.1.2 Résultats par année d'incident, maladie et accident ensemble

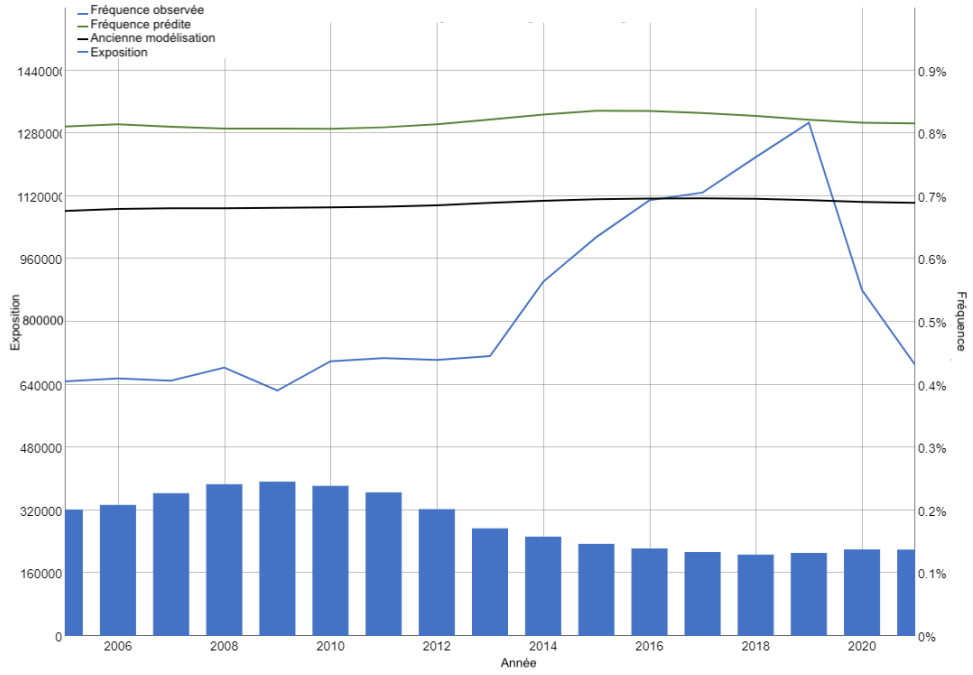
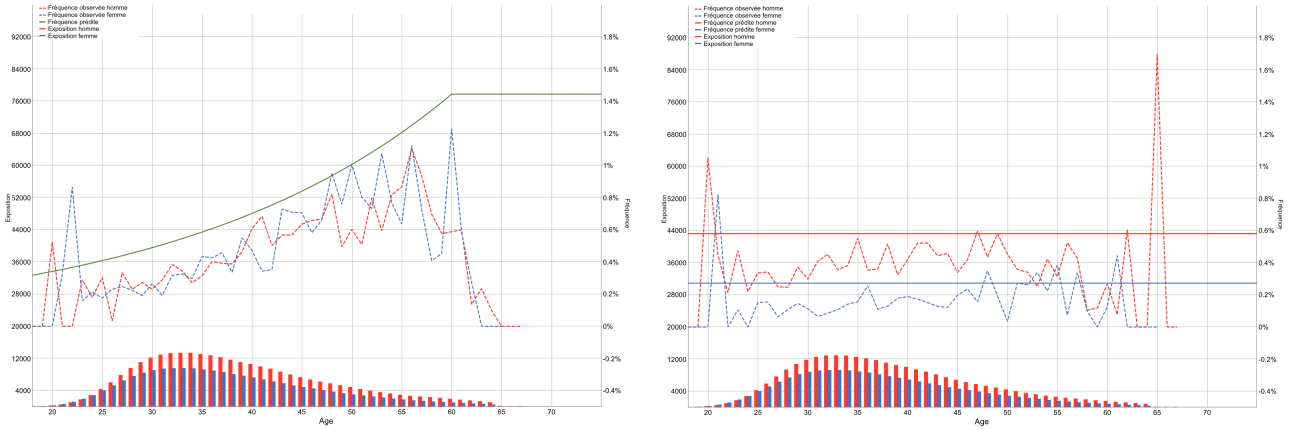



FIGURE D.2 : Fréquence observée, fréquence prédite et ancienne fréquence modélisée par année d'incident tous genres et toutes couvertures confondus pour le produit Personal Loan après modification des GLMs

La fréquence prédite est maintenant alignée avec le pic de 2019, mais reste stagnante avec les années sans prendre en compte la hausse observée (figure D.2).

## D.2 Produit Mortgage Loan

### D.2.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite pour le produit Mortgage Loan en couverture maladie selon l'âge

(b) Fréquences observées vs Fréquences prédites pour le produit Mortgage Loan en couverture accident selon l'âge

FIGURE D.3 : Fréquences observées vs Fréquences prédites par âge pour le produit Mortgage Loan des deux couvertures

Comme pour le Personal Loan, les fréquences prédites se trouvent maintenant au-dessus des observations (figure D.3).

### D.2.2 Résultats par année d'incident, maladie et accident ensemble

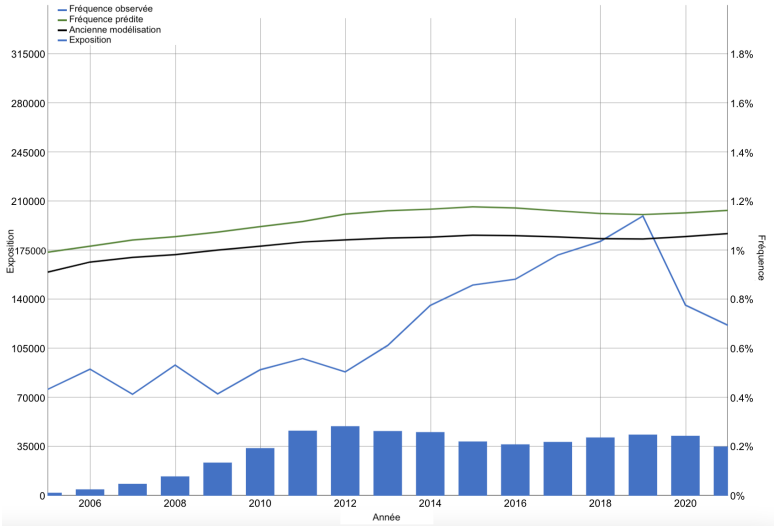
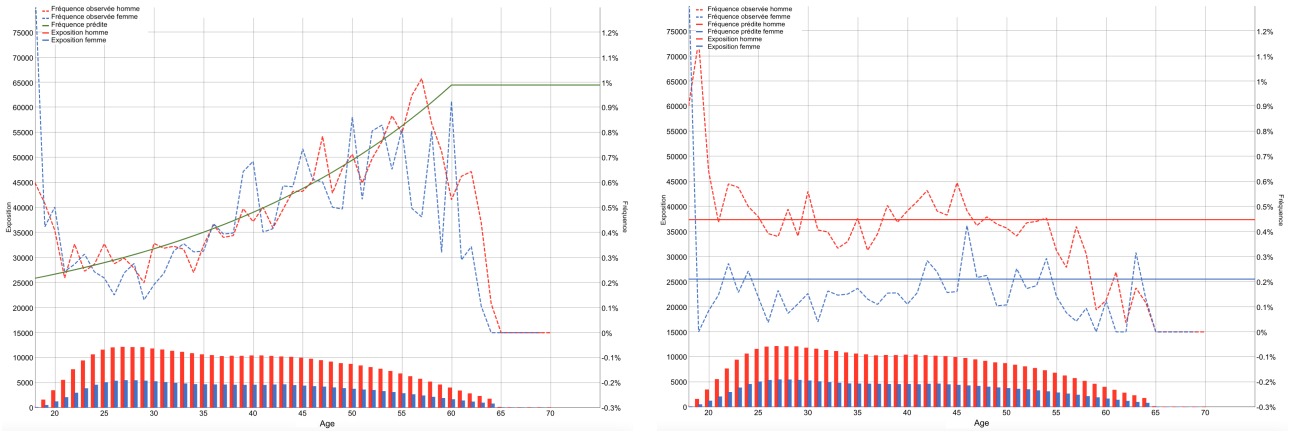



FIGURE D.4 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Mortgage Loan

Comme pour le Personal Loan, la fréquence prédite est maintenant alignée avec le pic de 2019, mais reste stagnante avec les années sans prendre en compte la hausse observée (figure D.4).

## D.3 Produit Car Loan

### D.3.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite pour le produit Car Loan en couverture maladie selon l'âge

(b) Fréquences observées vs Fréquences prédites pour le produit Car Loan en couverture accident selon l'âge

FIGURE D.5 : Fréquences observées vs Fréquences prédites par âge pour le produit Car Loan des deux couvertures

Les prédictions sont alignées avec les observations par âge pour les deux couvertures (figure D.5).

### D.3.2 Résultats par année d'incident, maladie et accident ensemble

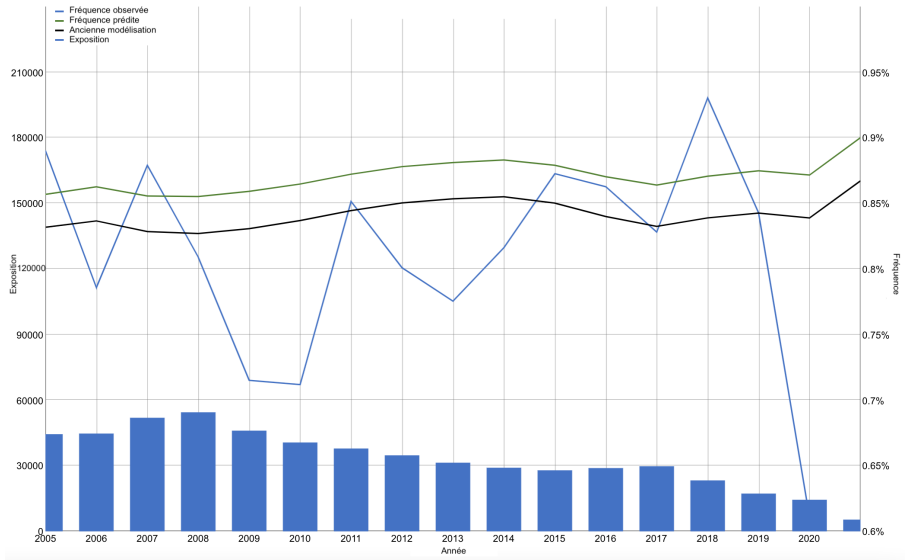



FIGURE D.6 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Car Loan

La fréquence prédite a légèrement été surélevée pour être placée au-dessus de la précédente modélisation

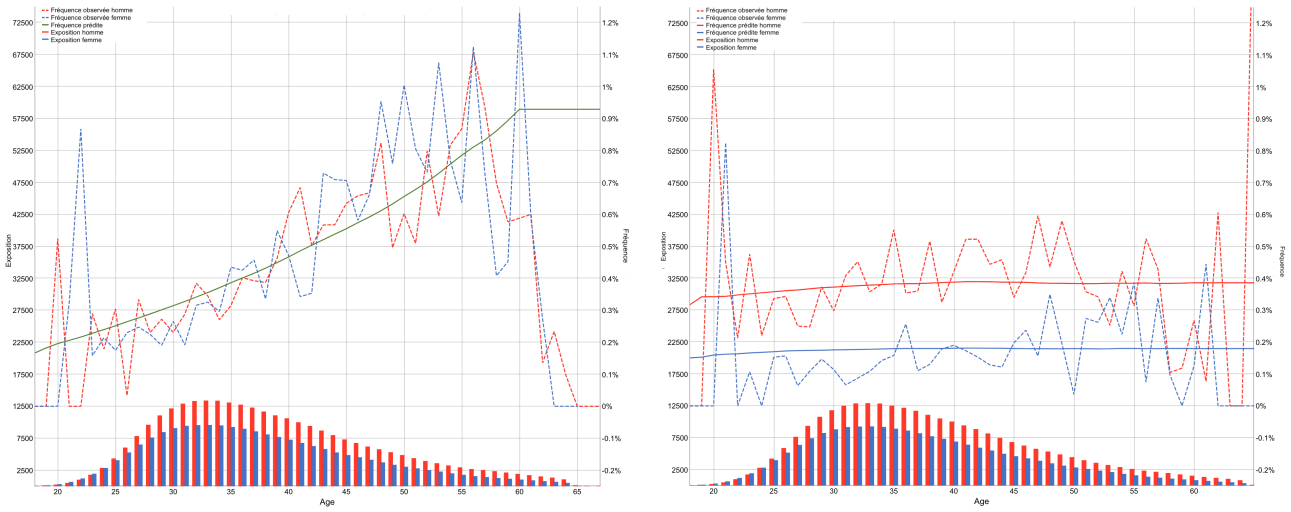
et de la fréquence observée. L'ajustement pour ce produit reste très léger (figure D.6).

# Annexe E

# Résultats de la 2e modélisation GLM pour les autres produits

## E.1 Produit Mortgage Loan

### E.1.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite pour le produit Mortgage Loan en couverture maladie selon l'âge

(b) Fréquences observées vs Fréquences prédites pour le produit Mortgage Loan en couverture accident selon l'âge

FIGURE E.1 : Fréquences observées vs Fréquences prédites par âge pour le produit Mortgage Loan des deux couvertures

Les prédictions sont alignées avec les observations par âge pour les deux couvertures (figure E.1).

### E.1.2 Résultats par année d'incident, maladie et accident ensemble

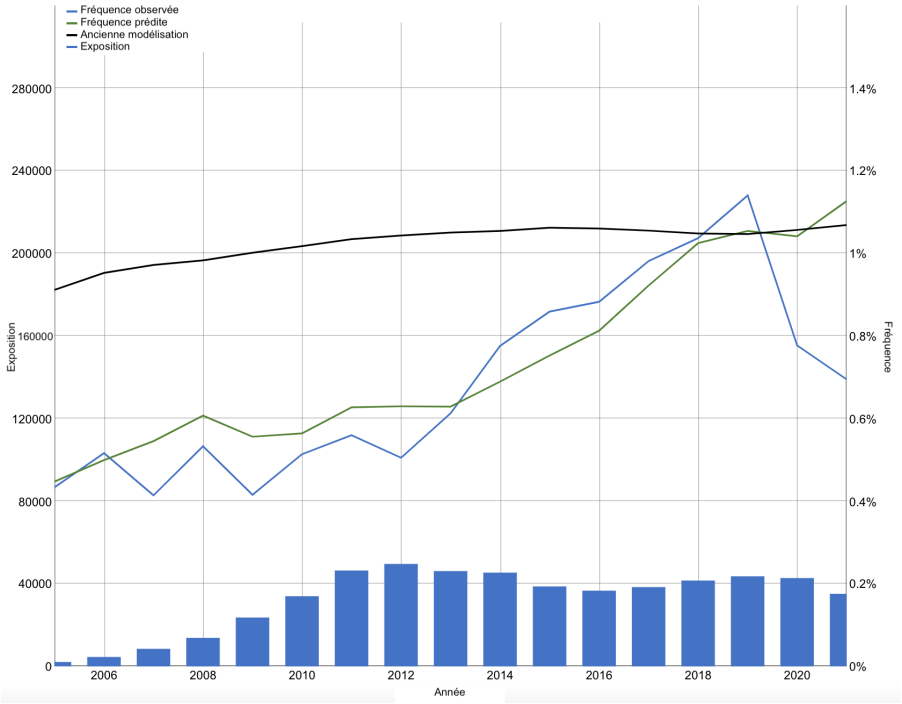
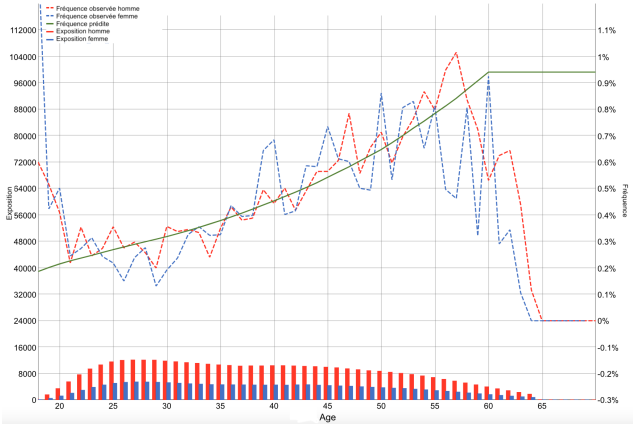



FIGURE E.2 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Mortgage Loan

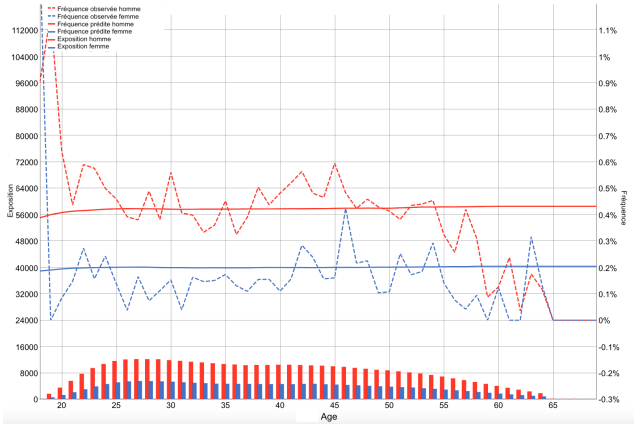
La modélisation GLM modélise bien la croissance de la fréquence avec les années à partir de 2013 (figure E.2).

## E.2 Produit Car Loan

### E.2.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite pour le produit Car Loan en couverture maladie selon l'âge




(b) Fréquences observées vs Fréquences prédites pour le produit Car Loan en couverture accident selon l'âge

FIGURE E.3 : Fréquences observées vs Fréquences prédites par âge pour le produit Car Loan des deux couvertures

Les prédictions sont alignées avec les observations par âge pour les deux couvertures (figure E.3).

### E.2.2 Résultats par année d'incident, maladie et accident ensemble

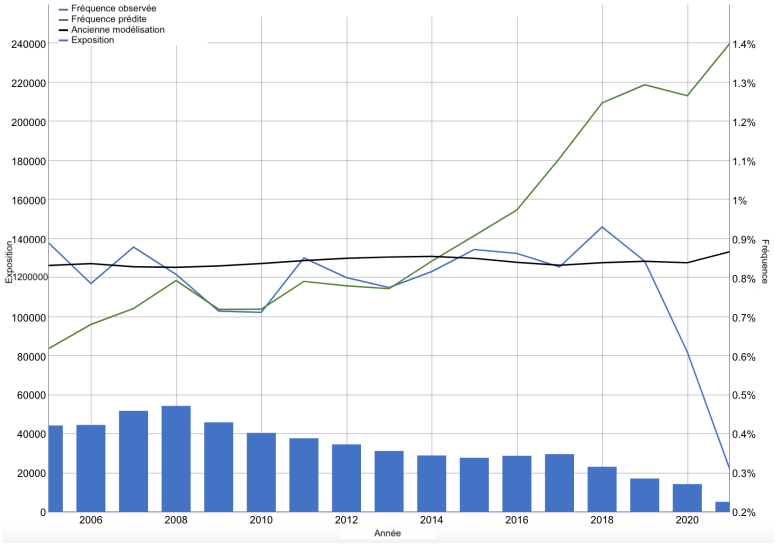



FIGURE E.4 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Car Loan

La fréquence prédite par le GLM est croissante à partir de 2013 et s'éloigne grandement des observations qui restent stagnantes. Le GLM enrichi des variables macroéconomiques modélise donc mal la

fréquence pour le produit Credit Card (figure E.4).

# Annexe F

# Méthodologie d'ajout des produits "Bulk" aux GLMs

Les produits Credit Card, Waiver of Premium et Income Protection ne disposent pas des informations nécessaires à leur modélisation directe via un GLM. Il est regardé ici comment le produit Credit Card est inclus aux équations de fréquence, la méthodologie restant la même pour les deux autres produits.

Le produit Credit Card présente 98% de données "Bulk", donc pas de variables explicatives âge et genre utilisables. Il faut utiliser le peu d'informations à disposition sur ce produit ainsi que toutes les informations disponibles sur les produits "non Bulk" pour en dériver des coefficients dans les deux GLMs.

Après avoir calibré les deux GLMs avec les trois produits "non Bulk", des tendances ont été remarquées :

- le coefficient pour l'âge dans le GLM maladie est le même pour les trois produits Personal Loan, Mortgage Loan et Car Loan. Il y a la même tendance exponentielle avec l'âge de la fréquence ;
- le coefficient pour le genre dans le GLM accident est le même pour les trois produits Personal Loan, Mortgage Loan et Car Loan. Il y a une segmentation similaire par genre.

**Il est fait la supposition que l'impact de l'âge et du genre sera le même pour les produits "Bulk"**. Pour ajouter les coefficients de Credit Card dans les deux GLMs, il convient de déterminer la fréquence maladie et accident séparément des trois produits "non Bulk" sur les périodes les plus pertinentes. Pour Personal Loan et Mortgage Loan, il est seulement regardé les fréquences en 2019 comme expliqué précédemment (2.2.3, il faut procéder au même ajustement sur le pic de fréquence de l'année 2019 pour le produit Mortgage Loan). Pour Car Loan, il est possible de se baser sur une période plus large à savoir 2015-2019.

Dans le GLM maladie, l'âge qui correspond aux fréquences déterminées ci-dessus est regardé et dans le GLM accident, la proportion d'hommes pour chaque produit "non Bulk" est déterminée. Cela donne le tableau suivant (tableau F.1).

Il faut calculer à présent la fréquence maladie du produit Credit Card sur la période 2012-2019 (afin de ne garder que les observations les plus récentes). Cela donne une fréquence de **0.24**%. La même chose est faite pour la fréquence accident et cela donne une fréquence de **0.11**%.

Afin de déterminer les coefficients associés au produit Credit Card dans les deux GLMs maladie et

| | Maladie | | Accident | |
|---|---|---|---|---|
| Produit | Fréquence moyenne | Age correspondant dans le GLM | Fréquence moyenne | Proportion d'hommes |
| Personal Loan (2019) | 0,51% | 43 | 0,31% | 75% |
| Mortgage Loan (2019) | 0,71% | 40 | 0,42% | 56% |
| Car Loan (2012-2019) | 0,54% | 43 | 0,32% | 66% |

TABLE F.1 : Fréquences des produits "non Bulk" de la couverture maladie avec l'âge correspondant dans le GLM maladie ainsi que celles de la couverture accident et de la proportion d'hommes associée

accident, il faut trouver un âge moyen et une distribution d'hommes. Il a été choisi de prendre l'âge le plus bas parmi les trois produits "non Bulk" soit **40 ans (celui de Mortgage Loan).** Il est décidé de prendre la proportion d'hommes la plus faible soit **56% (celui de Mortgage Loan). L'objectif ici est d'être le plus prudent possible en prenant les données âge et genre qui vont donner la fréquence finale la plus élevée dans les GLMs**, et donc la prime pure calculée en conséquent sera la plus élevée.

Il est possible maintenant de résoudre les équations suivantes :

- pour la maladie, $0.24\% = \exp(0.036 \times 40 - 6.780 + \textbf{CoefficientMaladieCreditCard})$,

  $\Longrightarrow \textbf{CoefficientMaladieCreditCard} = -0.684$,

  avec 0.24% la fréquence moyenne pondérée pour le produit Credit Card en couverture maladie, 0.036 le coefficient GLM de l'âge, 40 l'âge moyen choisi prudemment et -6.780 l'intercept du GLM (donc le coefficient du produit Car Loan) après modification (2.2.3) ;

- pour l'accident, $0.11\% = 56\% \times \exp(-5.408 + \textbf{CoefficientAccidentCreditCard}) + 44\% \times \exp(-0.755 - 5.408 + \textbf{CoefficientAccidentCreditCard})$,

  $\Longrightarrow \textbf{CoefficientAccidentCreditCard} = -1.174$,

  avec 0.11% la fréquence moyenne pondérée pour le produit Credit Card en couverture accident, 56% la proportion d'hommes choisie prudemment (et donc 44% la proportion de femmes), -5.408 l'intercept du GLM (donc le coefficient du produit Car Loan) après modification (2.2.3) et -0.755 le coefficient associé au genre féminin dans le GLM accident.

Comme pour les produits "non Bulk", il convient de dessiner les graphiques de la prédiction des GLMs maladie et accident en fonction de l'âge et du genre du produit Credit Card (figure F.1). Avec en haut la fréquence prédite en vert par âge par rapport à la fréquence moyenne en noir pour la couverture maladie, et en bas les fréquences prédites pour les hommes en rouge et les femmes en vert par âge par rapport à la fréquence moyenne en noir.

Il se retrouve bien l'évolution exponentielle de la fréquence avec l'âge pour le GLM maladie, qui croise la fréquence moyenne observée à l'âge 40 ans. Pour le GLM accident, il y a une différentiation entre homme et femme, les hommes étant au-dessus de la moyenne observée et les femmes en-dessous ce qui fait sens par rapport à ce qui a pu être observé pour les produits "non Bulk".

Si l'évolution de la fréquence prédite des deux GLMs ensemble par âge du produit Credit Card est regardée en séparant les hommes et les femmes, il est trouvé comme pour les produits "non Bulk" une croissance exponentielle avec l'âge qui est comparée à la précédente fréquence validée par AXA Partners (figure F.2). Avec en haut la fréquence prédite en vert et la précédente fréquence en noir par âge pour les hommes, et en bas la fréquence prédite en vert et la précédente fréquence en noir par âge pour les femmes.

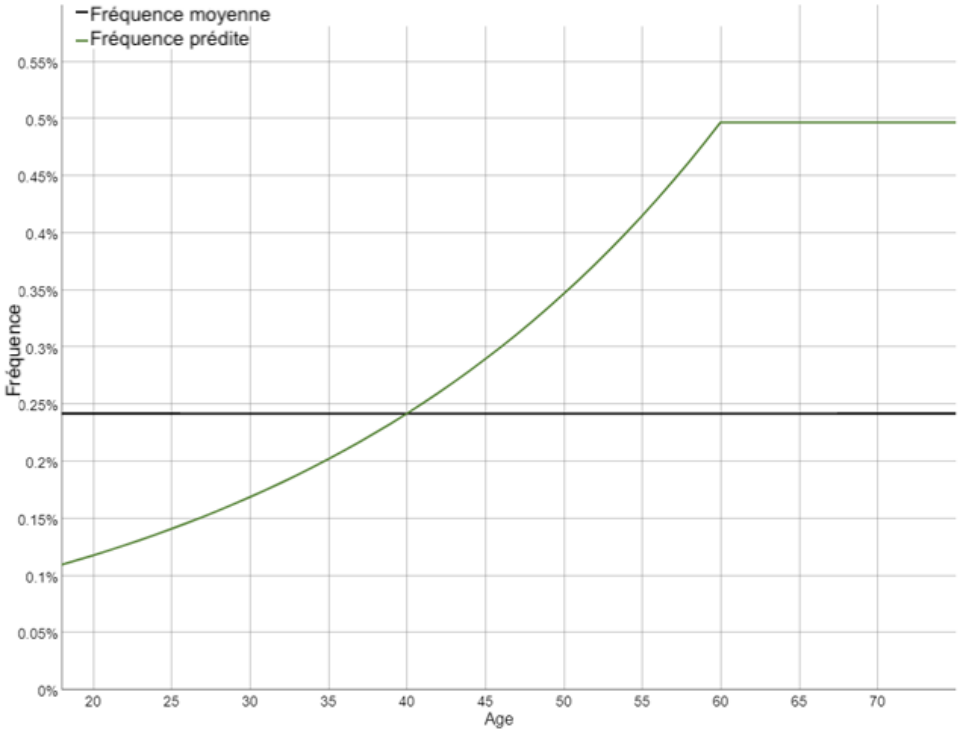
Cette nouvelle modélisation propose une évolution exponentielle de la fréquence avec l'âge pour les deux genres du produit "Bulk" Credit Card, là où la précédente fréquence choisie était identique à tout âge et pour les deux genres (courbes noires).

Les avantages de cette méthodologie :

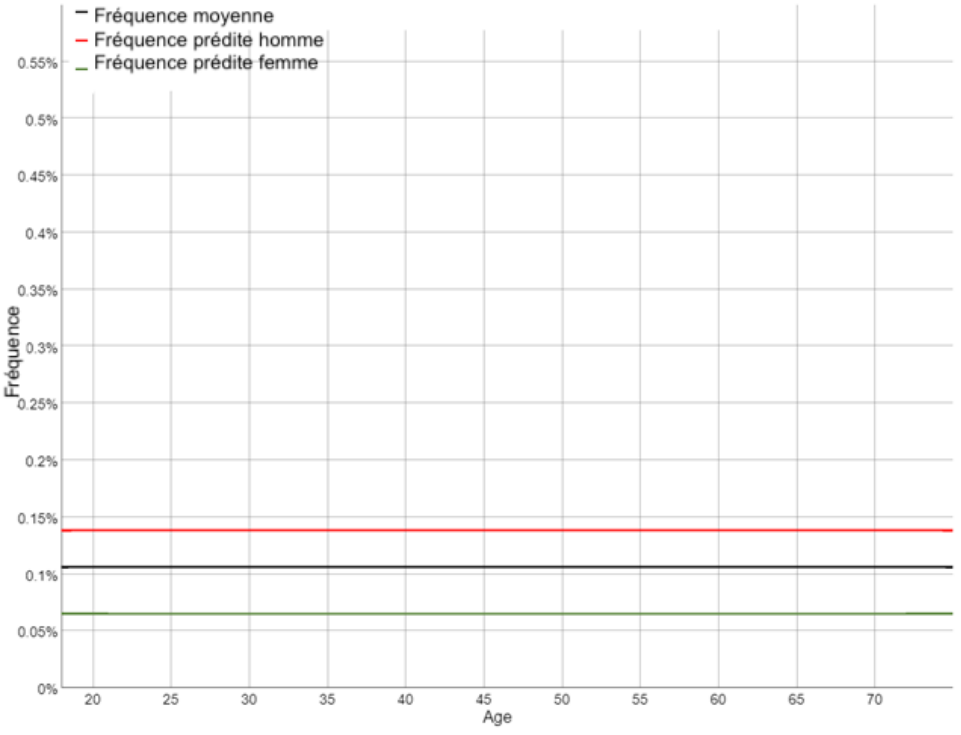
- simplicité car tous les produits sont présents dans les GLMs maladie et accident comme variables explicatives ;
- segmentation car il est gardé la même segmentation par âge et genre pour tous les produits.

Les principaux inconvénients de cette méthodologie :

- utilisation très faible des données à disposition sur ces produits ;
- impossibilité de vérifier la qualité de cette modélisation.

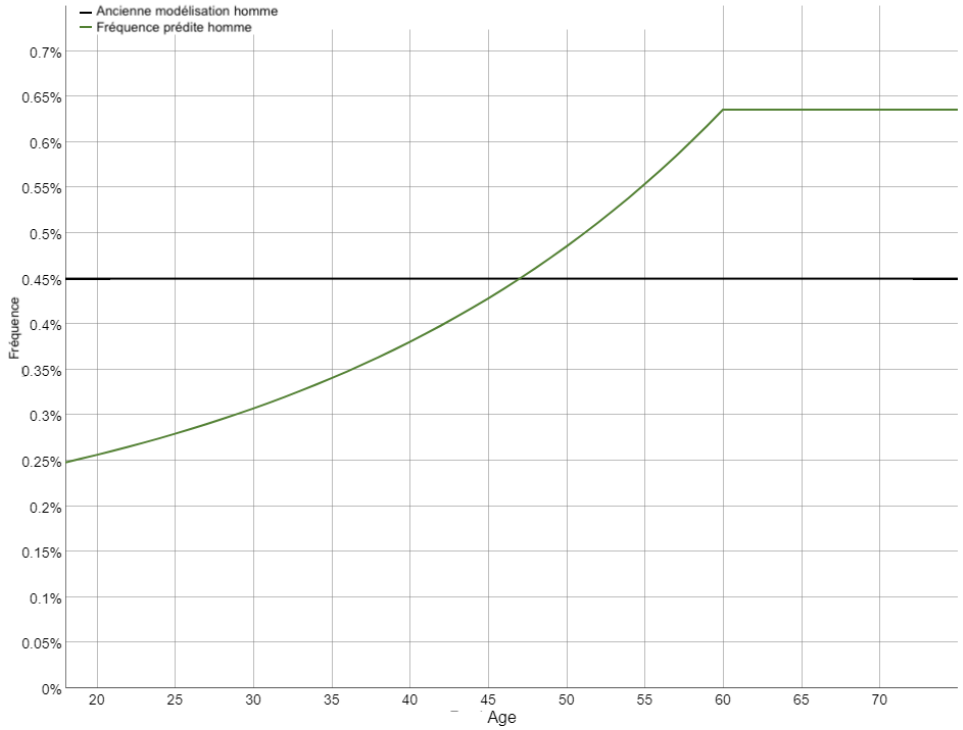



(a) Fréquence prédite du GLM maladie comparée à la fréquence moyenne observée par âge pour le produit Credit Card

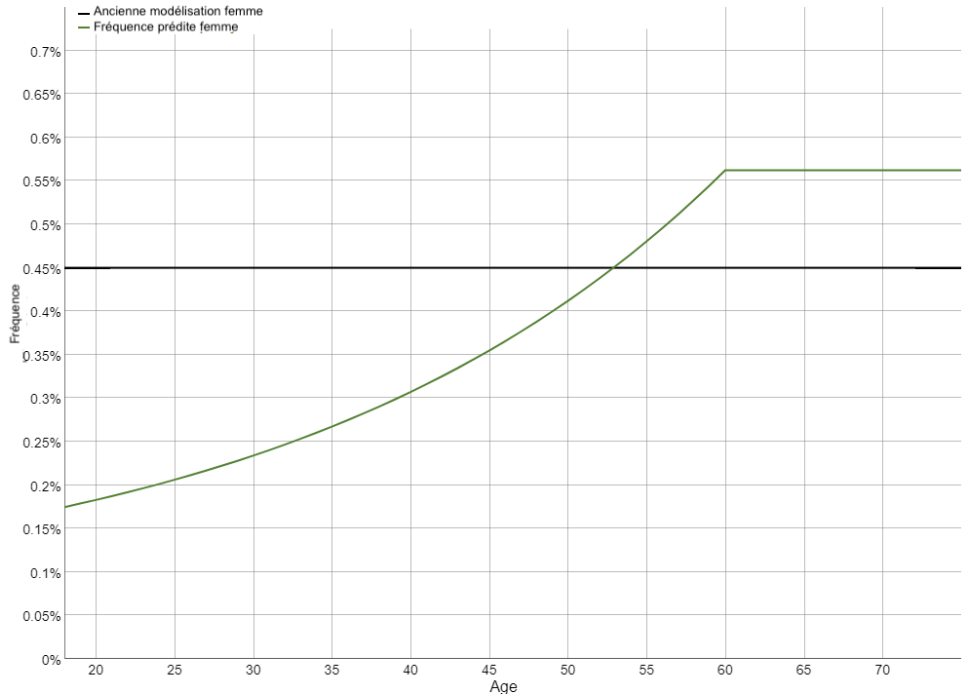



(b) Fréquences prédites par genre du GLM accident comparées à la fréquence moyenne observée par âge pour le produit Credit Card

FIGURE F.1 : Fréquences prédites des GLMs pour le produit Credit Card comparées aux fréquences moyennes par âge et par couverture




(a) Nouvelle fréquence prédite (maladie et accident ensemble) vs précédente fréquence validée par âge pour les hommes




(b) Nouvelle fréquence prédite (maladie et accident ensemble) vs précédente fréquence validée par âge pour les femmes

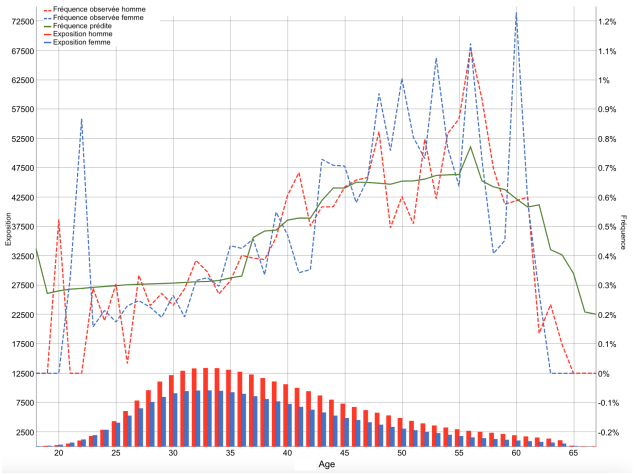
FIGURE F.2 : Nouvelle fréquence prédite (maladie et accident ensemble) vs précédente fréquence validée par âge et par genre

# Annexe G

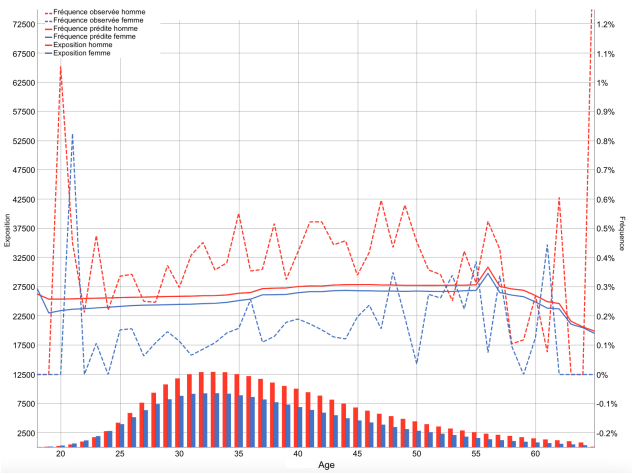
# Résultats de la modélisation *gradient boosting* pour les autres produits

## G.1 Produit Mortgage Loan

### G.1.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite pour le produit Mortgage Loan en couverture maladie selon l'âge




(b) Fréquences observées vs Fréquences prédites pour le produit Mortgage Loan en couverture accident selon l'âge

FIGURE G.1 : Fréquences observées vs Fréquences prédites par âge pour le produit Mortgage Loan des deux couvertures

Les prédictions sont alignées avec les observations par âge pour les deux couvertures (figure G.1).

### G.1.2 Résultats par année d'incident, maladie et accident ensemble

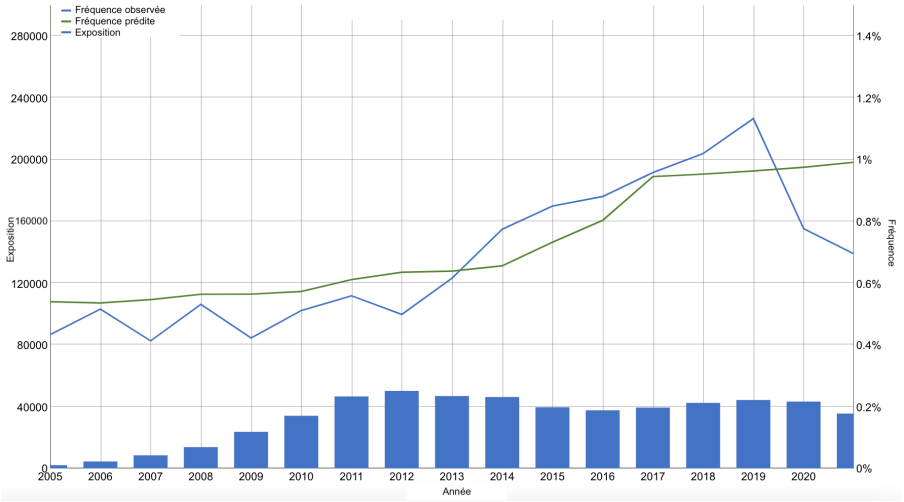
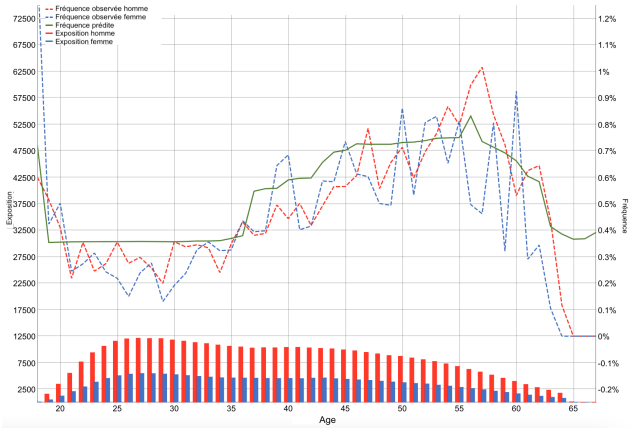



FIGURE G.2 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Mortgage Loan

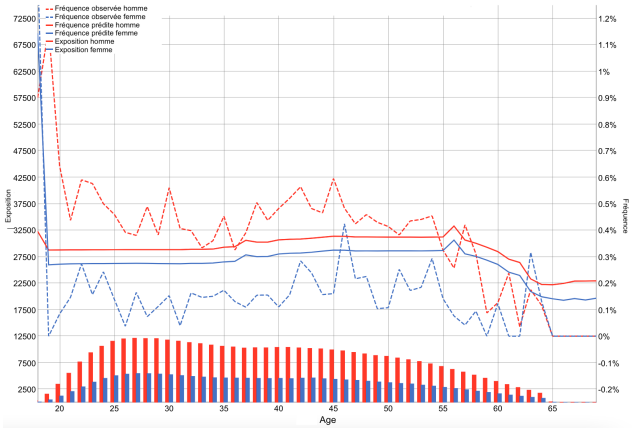
La croissance de la fréquence avec les années est plutôt bien modélisée par le *gradient boosting* pour le produit Mortgage Loan, avec la prédiction qui s'aligne sur les observations (figure G.2).

## G.2 Produit Car Loan

### G.2.1 Résultats par couverture selon l'âge et le genre




(a) Fréquences observées vs Fréquence prédite pour le produit Car Loan en couverture maladie selon l'âge




(b) Fréquences observées vs Fréquences prédites pour le produit Car Loan en couverture accident selon l'âge

FIGURE G.3 : Fréquences observées vs Fréquences prédites par âge pour le produit Car Loan des deux couvertures

Les prédictions sont alignées avec les observations par âge pour les deux couvertures (figure G.3).

### G.2.2 Résultats par année d'incident, maladie et accident ensemble

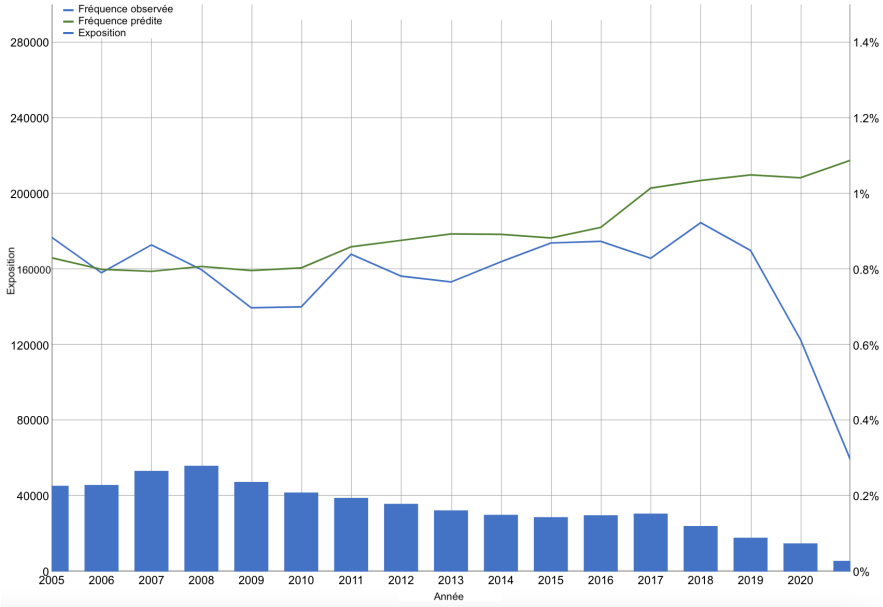



FIGURE G.4 : Fréquence observée et fréquence prédite par année tous genres et toutes couvertures confondus pour le produit Car Loan

La prédiction du *gradient boosting*, malgré une légère croissance, est alignée avec les observations stagnantes par année du produit Car Loan. Le *gradient boosting* est le seul des trois modèles de cette étude qui a su prédire une croissance pour les produits présentant une croissance de la fréquence, et quasiment pas de croissance pour le produit n'en présentant pas (figure G.4).

# Annexe H

# L'influence relative

L'influence relative est une méthode d'analyse des modèles de *machine learning* qui permet de déterminer l'importance des différentes variables explicatives pour la prédiction de la variable réponse. Cette méthode est particulièrement utile pour comprendre comment les différentes variables du modèle affectent la prédiction, et pour optimiser les performances du modèle. L'une des mesures couramment utilisées pour évaluer l'importance des variables explicatives est le Mean Squared Error (MSE).

Le MSE est une mesure de la qualité de l'ajustement d'un modèle aux données d'entraînement. Il est défini comme la moyenne des carrés des écarts entre les valeurs prédites par le modèle et les valeurs réelles. Plus précisément, le MSE mesure la variance de l'erreur de prédiction du modèle. Un MSE faible indique une bonne adéquation entre le modèle et les données d'entraînement.

L'influence relative est déterminée en mesurant la variation du MSE lorsque chaque variable explicative est retirée du modèle. Pour chaque variable, le MSE est calculé pour le modèle complet et pour le modèle sans cette variable. La différence entre ces deux valeurs de MSE est ensuite divisée par la valeur du MSE du modèle complet. Le résultat est une mesure de l'influence relative de chaque variable, exprimée en pourcentage ou en score.

Cependant, il est important de noter que l'influence relative n'est pas une mesure absolue de l'importance d'une variable. Elle peut varier en fonction du modèle et des données d'entrée. De plus, l'influence relative ne doit pas être utilisée comme un critère absolu pour déterminer quelles variables doivent être retirées du modèle. En effet, retirer une variable peut avoir un impact sur la qualité de l'ajustement et les performances du modèle.

# Glossaire des abréviations

Voici le liste des abréviations utilisées dans cette étude :

- GLM : Generalized Linear Model, en français Modèle Linéaire Généralisé ;
- IARD : Incendies, Accidents et Risques Divers ;
- ACPR : Autorité de Contrôle Prudentiel et de Résolution ;
- TTD : Total and Temporary Disability, en français incapacité ;
- IU : Involuntary Unemployment ;
- H : Hospitalisation ;
- LE : Life Events ;
- CLP : Credit & Lifestyle Protection, une des unités d'AXA Partners ;
- EM : Excess Monthly ;
- RM : Retro Monthly ;
- RD : Retro Daily ;
- PMP : Portfolio & Monitoring Management ;
- AS : Actuarial Services ;
- ZIP : Zero-Inflated Model ;
- AIC : Akaike's Information Criterion ;
- BIC : Bayesian Information Criterion ;
- EDR : Explained Deviance Ratio ;
- MSE : Mean Squared Error ;
- IPC : Indice des Prix à la Consommation.