Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : JULIEN Nathan

Titre : **Confrontation de modèles prédictifs dans le but de quantifier les impacts de la crise sanitaire sur l'incidence du risque incapacité.**

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires

. . . . . .

. . . . . .

. . . . . .

Membres présents du jury de l'ISFA

. . . . . .

. . . . . .

. . . . . .

Nom entreprise : AG2R La Mondiale
Signature :

Directeur de mémoire en entreprise :
Nom : Auryane HOAREAU

Signature :
Invité :
Nom :
Signature :

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

Signature du responsable entreprise

Signature du candidat

Résumé

En 2020, l'apparition de la crise sanitaire, liée au virus de la Covid-19, est venue bouleverser la vie des entreprises et de leurs salariés. Elles ont dû faire face à des défis économiques et sociaux émergents. En particulier, cette pandémie a engendré une hausse brutale de l'absentéisme. Néanmoins, cette année de crise semble quelque peu biaisée en matière d'arrêts de travail. Les conditions d'octroi de certains arrêts, définis comme dérogatoires, ne sont pas représentatives d'une année ordinaire. Cet absentéisme nécessite donc une étude approfondie. De plus, les conséquences pour les organismes assureurs vont être multiples. Les prestations versées et les provisions à constituer au titre du risque incapacité vont drastiquement augmenter. L'idée est donc d'envisager une nouvelle norme tarifaire, spécifique à cette année de crise.

Dans la méthode classique « *coût x fréquence* », le processus de tarification se base sur deux lois : une loi d'incidence et une loi de maintien en arrêt. Après une analyse de suivi du portefeuille menée sur différents indicateurs de sinistralité, seule la première loi nous intéresse. Par conséquent, il est question d'utiliser des modèles d'incidence pertinents afin de capter les sensibilités de ce risque dans le portefeuille.

Auparavant, en prévoyance collective, les systèmes d'informations contenaient uniquement les individus sinistrés dont la durée de l'arrêt dépassait la franchise. L'apport du processus de Déclaration Sociale Nominative (DSN) vient résoudre les limites de ce manque de données. Désormais, l'information disponible est précise, régulière et comprend l'ensemble des assurés du groupe. Les travaux sous-tendant ce mémoire débutent par la construction d'une base de données fiable et complète. Cette base servira de point de départ pour toutes les modélisations futures. Ensuite, des statistiques descriptives globales puis spécifiques à chaque variable sont présentées. Cela permet de donner un premier constat des facteurs qui influent sur la fréquence d'arrêts.

Pour mener à bien ce projet, différents modèles prédictifs sont opposés. D'abord, un modèle de régression paramétrique est simulé. Il est construit à l'aide de modèles linéaires généralisés. En assurance, ce sont des modèles classiques de tarification. Dans un second temps, diverses méthodes de Machine Learning sont testées pour prédire la fréquence d'arrêts. Seuls les modèles d'apprentissage XGBoost sont présentés dans ce mémoire. Enfin, ces deux modèles sont comparés à un outil nommé Akur8. Son fonctionnement repose sur des modèles additifs généralisés automatisés.

À travers différents critères, ces modèles sont confrontés pour ne retenir que le plus performant. Une fois que cette étape est réalisée, l'objectif est donc de comparer les résultats obtenus entre, d'une part les modèles construits sur les données en période de crise sanitaire et de l'autre, ceux sur les années précédentes, sans choc majeur sur la sinistralité. Basée sur différents indicateurs de performance, une analyse exhaustive de ces deux modèles est alors effectuée.

Finalement, les résultats permettent de comprendre et de mieux appréhender les conséquences de cette pandémie sur le risque incapacité. En effet, la fréquence d'arrêts est marquée par une hausse générale de l'ensemble du portefeuille. Néanmoins, l'évolution des profils types montre que certaines catégories de population ont été plus exposées que d'autres.

Mots clés : Prévoyance collective, Déclaration Sociale Nominative, absentéisme, crise sanitaire, incapacité, arrêt de travail, suivi de portefeuille, indicateurs, dérive, sinistralité, modélisation, incidence, GLM, XGBoost, Akur8, GAM

Abstract

In 2020, the emergence of the health crisis, linked to the Covid-19 virus, turned the lives of companies and their employees upside down. They had to face emerging economic and social challenges. In particular, the pandemic led to a sharp increase in absenteeism. Nevertheless, this year of crisis seems somewhat biased in terms of work stoppages. Indeed, the conditions for granting certain work stoppages, defined as derogatory, are not representative of an ordinary year. This absenteeism therefore requires an in-depth study. More, the consequences for insurance companies will be multiple. The benefits paid and the provisions to be set aside for the risk of inability will increase drastically. The idea is therefore to consider a new tariff standard, specific to this year of crisis.

In the classic "cost x frequency" method, the pricing process is based on two laws : a law of incidence and a law of maintenance during work stoppage. After a portfolio monitoring analysis conducted on different loss indicators, only the first law interests us. Therefore, the use of relevant incidence models to capture the sensitivities of the inability risk and the insured portfolio is discussed.

Previously, in the group benefit plan, the information systems contained only those claimants whose downtime exceeded the deductible. The contribution of the « Déclaration Sociale Nominative » process solves the limitations of this lack of data. From now on, the information available is precise, regular and includes all of the group's insureds. The work underlying this work begins with the construction of a reliable and complete database. This database will serve as a starting point for all future modeling. Then, global and specific descriptive statistics for each variable are presented. This allows to give a first observation of the factors that influence the frequency of work stoppage.

To carry out this project, different predictive models are opposed. First, a parametric regression model is simulated. It is built using generalized linear models. In insurance, these are classical pricing models. In a second step, various Machine Learning methods are tested to predict the incidence of work stoppage. Only the XGBoost learning models are presented in this thesis. Finally, these two models are compared to a pricing tool named Akur8. Its operation is based on automated generalized additive models.

Through different criteria, these models are compared to retain only the best performing one. Once this step has been completed, the objective is to compare the results obtained between the models built on the data during a health crisis and the models built on the data from previous years, without any major shock on the claims experience. Based on different performance indicators, an exhaustive analysis of these two models is then performed.

Finally, the results allow us to understand and better apprehend the consequences of this pandemic on inability risk. Indeed, the incidence of work stoppage is marked by a general increase in the entire portfolio. Nevertheless, the evolution of the typical profiles shows that certain categories of population were more exposed than others.

Keywords : Group benefit, Nominative Social Declaration, absenteeism, health crisis, inability, labour disruption, indicators, data monitoring, claims rate drift, modeling, incidence, GLM, XGBoost, Akur8, GAM

Note de synthèse

Contexte général

Connaître et maitriser son risque est essentiel pour toute activité d'assurance. En effet, le cycle de production est inversé. L'assureur reçoit des cotisations, fixées en avance, pour des prestations dont les versements sont conditionnés à la réalisation des risques assurés. C'est pourquoi la tarification occupe une place centrale, en particulier pour des risques complexes comme la prévoyance.

L'étude réalisée au cours de ce mémoire porte sur la prévoyance collective et spécifiquement le risque d'incapacité. C'est un risque multiple qui peut s'étendre sur plusieurs années. Il est encadré par une réglementation rigoureuse. L'enjeu est donc de trouver une segmentation tarifaire capable de capter la complexité de ce risque.

Avoir une base de données robuste est donc primordiale pour analyser ce risque. Cependant, les systèmes d'informations utilisés en prévoyance collective ne permettent pas d'avoir une vision globale de notre portefeuille. En effet, seuls les assurés victimes d'arrêts de travail, dont la durée dépasse le nombre de jours de franchise, apparaissent dans ces systèmes d'information. Tous les individus qui n'ont pas d'arrêts ou qui ont des arrêts courts sont donc exclus de ces données. Cela génère donc un biais. Pour pallier ce manque d'information, les données issues de la Déclaration Sociale Nominative (DSN) sont exploitées au cours de cette étude.

Aujourd'hui, la DSN est obligatoire pour toutes les entreprises, qu'elle soit du secteur privé ou public. C'est un processus qui fait suite à la paie. Elle remplace toutes les formalités déclaratives, liées à l'emploi et à la vie des salariés. La particularité de la DSN réside dans le fait qu'elle soit unique, dématérialisée et mensuelle. Elle permet de fiabiliser et de sécuriser les informations et les droits des salariés. L'apport de la DSN comble les limites des anciens systèmes d'informations. Désormais, tous les assurés, qu'ils soient sinistrés ou non, sont connus et présents dans les données utilisées. Néanmoins, la construction d'une base fiable et adaptée à notre besoin nécessite de nombreux retraitements.

Une étude préalable de suivi du portefeuille a permis d'affiner les objectifs de ce mémoire. Elle se base sur différents indicateurs de sinistralité (prestations payées, nombre de sinistres, durée moyenne, IJ, âge à la survenance ...) présentés par année comptable et par année de survenance. En effet, le but est de quantifier les impacts de la crise sanitaire. Les prestations versées aux assurés ont explosé au cours de cette année 2020. Cette forte hausse est principalement due à l'augmentation significative du nombre d'arrêts. C'est pourquoi seule la modélisation de l'incidence, c'est-à-dire la fréquence de tomber en arrêt, sera étudiée. Au cours de ce mémoire, les modèles sont construits en retenant une franchise égale à 0 jour. Cependant, ces travaux sont généralisés sur diverses franchises récurrentes allant de 3 à 365 jours.

En accord avec la direction d'AG2R La Mondiale qui voulait une étude complète et précise de l'absentéisme 2020, l'incidence de cette sinistralité bien spécifique est alors analysée. De plus, l'ensemble de ces travaux va servir de point de départ pour la refonte des normes tarifaires en prévoyance collective. Néanmoins, il semble que la période de crise sanitaire n'était pas la plus représentative en matière d'arrêts de travail. En effet, les conditions d'octroi des certains arrêts, définis comme « dérogatoires », laisse penser que le nombre de sinistres déclarés soit quelque peu biaisé.

Compréhension des données utilisées

Maintenant que le contexte et les notions essentielles sont définis, il est temps de présenter les étapes de construction pour aboutir à la base finale. Au préalable, il est nécessaire de définir le périmètre d'étude. Ensuite, comme les données proviennent de la DSN, donc d'un nouveau processus d'informations, elles doivent être adaptées à nos besoins. En effet, les données brutes fournies nécessitent de nombreux retraitements.

En outre, des règles de gestion, notamment sur les rechutes, sont appliquées pour mieux appréhender le risque incapacité. Par la suite, il faut choisir les variables à retenir, et éventuellement les transformer pour pouvoir les intégrer dans nos modélisations. Enfin, des statistiques globales sur la population sont réalisées, ainsi que des statistiques descriptives sur l'ensemble des variables retenues. Cela permet donc de bénéficier d'un premier constat des facteurs qui influent sur la fréquence d'arrêts. En effet, cette analyse descriptive a permis de mettre en évidence quelques points principaux :

- Malgré un pic d'assurés au cours de l'année 2019, la population reste relativement stable entre 2018 et 2020. De même, l'exposition annuelle moyenne est stable. Elle gravite autour de 75%, soit 9 mois de présence par an.

- L'incidence moyenne du portefeuille accuse une forte hausse. En effet, le taux moyen d'incidence était de 43% durant les années antérieures à la crise sanitaire. Désormais, en 2020, le taux moyen atteint les 55%, soit une augmentation générale de 28%.

- L'ensemble de la population subit cette pandémie et les conséquences sur l'absentéisme sont importantes. Cependant, certaines catégories de salariés semblent plus exposées que d'autres.

Présentation des modèles appliqués

Une fois que la base de données est construite, il est important de présenter les différents modèles employés. En effet, trois algorithmes sont comparés.

Dans un premier temps, les Modèles Linéaires Généralisés (GLM) sont mis en place. En effet, ce modèle est très répandu pour la tarification en assurance. Les GLM sont des processus relativement simples qui permettent de prédire une variable cible à l'aide de diverses variables explicatives. Après avoir vérifié les hypothèses sous-jacentes des GLM, les résultats obtenus permettent d'identifier les variables significatives. Ce sont elles qui vont influencer la fréquence des arrêts de travail. De plus, un coefficient est associé à chaque variable, si bien que l'importance de chacune d'entre elles est interprétable.

Dans un second temps, divers modèles de Machine Learning sont étudiés. D'abord, l'algorithme de *CART*, puis de *Random Forest* et enfin XGBoost. Le modèle XGB s'avère être le plus pertinent et sera le seul à être présenté dans ce mémoire. L'*Extreme Gradient Boosting* est un modèle d'apprentissage qui repose sur la méthode du boosting. Cela consiste à assembler un grand nombre de « *weak leaner* » pour former un « *strong learner* » plus robuste. Il repose sur une longue liste d'hyperparamètres, choisis ou non par l'utilisateur. À chaque itération, cet algorithme apprend de ses erreurs, pour ainsi réduire les écarts de prédiction.

Enfin, un outil de tarification nommé Akur8 est également utilisé pour modéliser l'incidence des arrêts de travail. Il s'appuie sur des Modèles Additifs Généralisés (GAM) automatisés. C'est une solution flexible qui propose de faire varier les modélisations en fonction des besoins. De plus, il dispose de nombreux graphiques pour évaluer les performances et analyser les résultats. Néanmoins, la place de l'utilisateur reste centrale. Il définit, contrôle et ajuste la construction des modèles selon ses objectifs et ses contraintes.

En outre, pour aboutir à des modèles plus robustes, une technique de validation croisée est appliquée pour chacune de ces trois solutions. La méthode des *K*-folds est employée. Elle permet de résoudre des problèmes de sous ou sur-apprentissage. De plus, un rééchantillonnage des données est opéré. Les valeurs rares sont exploitées avec une probabilité faible, tandis que les valeurs intermédiaires avec une probabilité plus importante. Ainsi, le sous-échantillon construit est plus représentatif et plus harmonieux. En définitive, après cross-validation, les modèles sont plus robustes et moins biaisés.

Confrontation des trois modèles

Les trois modèles ainsi conçus sont opposés pour ne retenir que le plus performant et ensuite analyser les impacts de la crise sanitaire.

Pour mener à bien cette comparaison, des indicateurs de performance communs aux trois modèles doivent être étudiés. Ces métriques sont construites pour analyser la qualité d'ajustement des modèles et estimer les écarts de prédiction. Ils sont résumés dans le tableau suivant.

	GLM	XGBost	Akur8
Gini	27,44%	28,99%	26,67%
R^2	4,5%	5,61%	7,99%
RMSE	1,437	1,422	1,136
MSE	2,211	2,023	1,29
MAE	0,876	0,833	0,738

Les métriques sont relativement proches, ce qui rassure quant à la pertinence des trois modèles. Néanmoins, l'outil Akur8 semble se dégager des autres grâce à son meilleur coefficient R^2 et aux mesures d'écarts d'estimation (RMSE, MSE et MAE) plus faibles. L'algorithme XGBoost concurrence fortement Akur8 puisque c'est lui qui a le meilleur coefficient de Gini, c'est à dire qu'il segmente mieux la population sous risque.

L'analyse de ces métriques n'est pas suffisante. C'est pourquoi l'étude des courbes de Lift et de Lorenz ainsi qu'une analyse de l'incidence prédite, variable par variable, vient compléter cette comparaison. À titre d'illustration, prenons un exemple concernant l'évolution du taux d'incidence en fonction de l'âge des salariés.

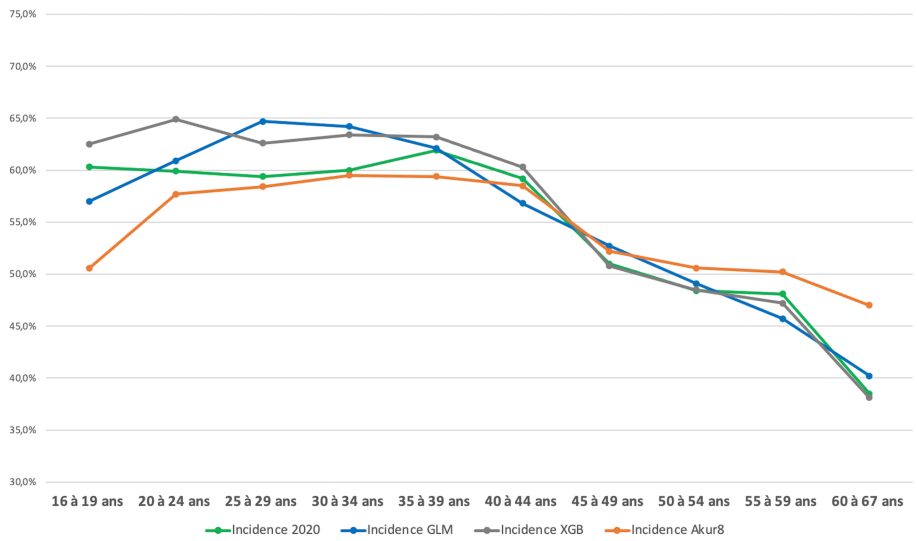


FIGURE 1 – Étude de l'incidence en fonction de l'âge

Le taux d'incidence réel observé en 2020, en fonction de l'âge des salariés, se manifeste par la courbe verte. Ensuite, les incidences des trois autres modèles viennent se superposer. Les GLM, XGBoost et Akur8 captent la tendance générale de stagnation jusqu'à l'âge de 35 ans puis de décroissance passée ce seuil. Néanmoins, le modèle XGBoost s'ajuste mieux à la population. Il présente les écarts d'estimation les plus faibles.

Finalement, après cette triple comparaison, c'est le modèle XGBoost qui paraît être le meilleur. Cependant, cet algorithme de Machine Learning présente deux défauts récurrents. Il est sensible aux problèmes « *d'overfitting* » et reste encore trop « *black box* ». En effet, même si le principe de construction est plutôt clair, le fonctionnement précis de cet algorithme s'avère bien trop opaque. Pour rappel, la prévoyance collective est un risque très réglementé. En cas de contrôle, il faut pouvoir identifier chacune des étapes et expliciter tous les résultats. Malgré ces très bonnes performances, ce modèle ne peut donc être mis en place dans le processus de construction de loi d'incidence. En revanche, l'outil Akur8 présente des résultats un peu moins performants mais bénéficie de méthodes certifiées et auditables. Le choix final se tourne donc vers ce modèle.

Étude comparative pour quantifier les impacts de la crise sanitaire

Pour rappel, le but de ce mémoire est de quantifier les impacts de la crise sanitaire, principalement sur la fréquence d'arrêt. Pour ce faire, un premier modèle prédictif est mis en place, sur les années antérieures à la Covid-19, à savoir 2018 et 2019. Puis, un second qui repose sur des données de l'année 2020. C'est donc à l'aide de l'outil Akur8 que sont réalisées ces deux modélisations.

D'abord, il est important de constater que pour ces deux bases distinctes, ce sont les mêmes variables explicatives qui entrent en jeu. Malgré le changement de comportement des assurés et la hausse de l'absentéisme, les deux modèles prédisent une incidence à l'aide des cinq mêmes variables discriminantes. Ces variables sont : le salaire annuel, le nombre de salariés par établissement, la CSP, l'âge et enfin le genre des individus. Or, certaines catégories de la population, qu'elles soient plus ou moins touchées par cette pandémie, auraient pu faire varier les informations à inclure dans les modèles. Les deux variables les plus importantes pour prédire l'incidence en incapacité sont le salaire annuel et la taille de l'entreprise.

Comme effectué précédemment, les cinq métriques de performance sont présentées pour comparer les modèles avant et pendant crise sanitaire.

	2018/2019 avec 5 var.	2020 avec 5 var.
Gini	27,59%	26,67%
R^2	7,12%	7,99%
RMSE	1,042	1,136
MSE	1,086	1,29
MAE	0,633	0,738

Les résultats apportés par le modèle antérieur à la crise sanitaire présentent majoritairement de meilleurs indicateurs. En effet, ce modèle cumule la sinistralité de deux années consécutives. Par conséquent, cela permet de gagner en robustesse et d'avoir de meilleures prédictions.

Ensuite, les courbes de Lift et de Lorenz associées à ces deux modélisations sont étudiées. Ces deux courbes sont de bons indicateurs pour illustrer la répartition et la segmentation des prédictions.

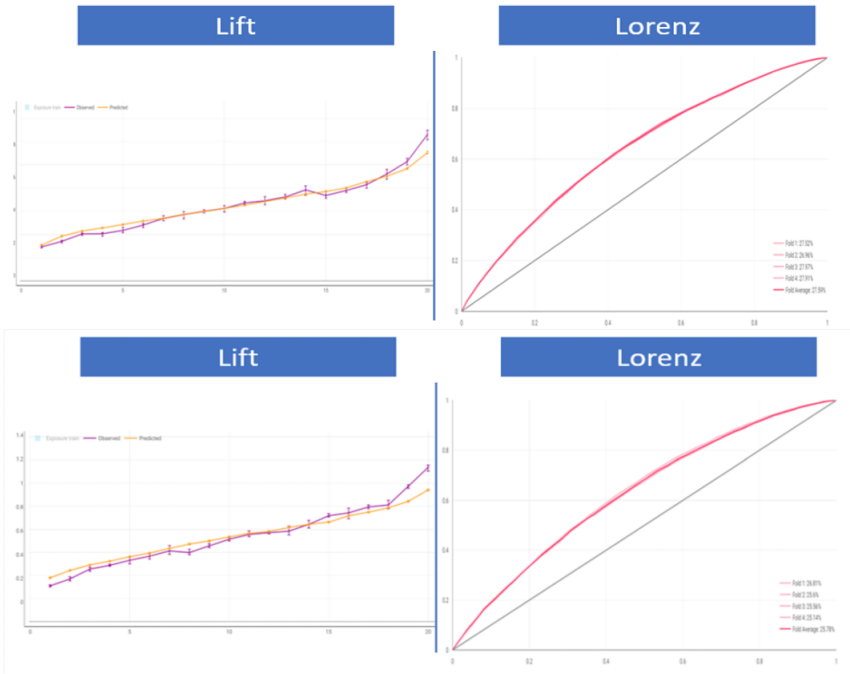


FIGURE 2 – Synthèse des courbes de Lift et de Lorenz (base 2018/2019 en haut et 2020 en bas)

Elles sont quasiment identiques. La courbe de Lift, associée aux années 2018 et 2019, affiche des prédictions (en jaune) légèrement plus proches des observations (en violet). De plus, pour les deux modèles, les prédictions sont légèrement surévaluées pour les faibles incidences, à gauche de la courbe de Lift, et sous-évaluées pour les fortes incidences, à droite.

En outre, les courbes de Lorenz sont également similaires. Pour chacun des « *folds* » considérés, une courbe de Lorenz est exposée. Elles rendent compte de modèles stables, puisqu'elles se superposent quasiment en tout point, pour former la courbe de Lorenz globale, très lisse.

Enfin, tous les résultats des modèles sont étudiés et en particulier le comportement des coefficients associés aux variables significatives. Ce sont ces coefficients qui vont servir à la construction de lois d'incidence.

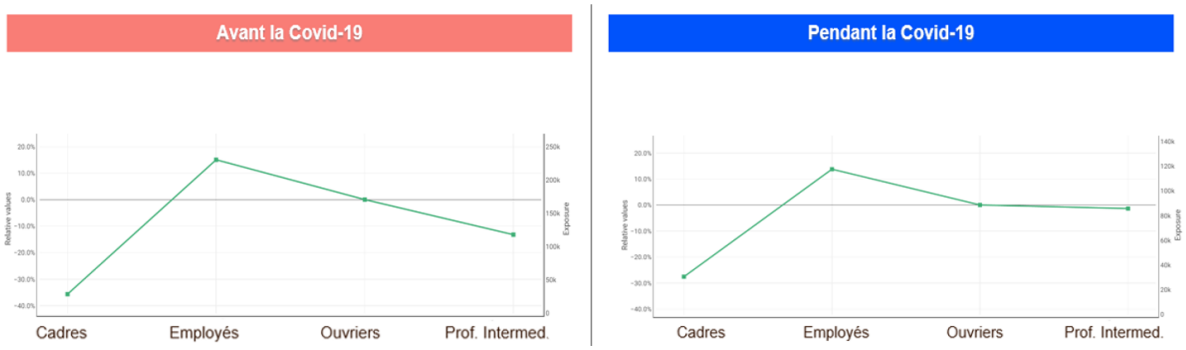


FIGURE 3 – Variation des coefficients associés à la CSP

À travers cet exemple sur la catégorie socio-professionnelle des salariés, deux points majeurs se dégagent. Premièrement, malgré une tendance générale à la hausse, l'effet des cadres, des employés et des ouvriers sur la variable cible est similaire pour les deux modèles. Deuxièmement, les professions intermédiaires sont plus impactées que le reste de la population. En effet, cette catégorie est principalement composée de professeurs des écoles et de personnel travaillant dans la santé ou dans le social. Toutes ces personnes sont en contact permanent avec de nombreux individus. Ils sont donc plus enclins à recevoir et transmettre des virus.

Conclusion

À la suite de l'apparition de la crise sanitaire en France, une démarche basée sur la modélisation de l'incidence en incapacité a été menée. L'objectif est de comprendre quelles ont été les conséquences de ce virus, en estimant les évolutions par rapport à des années sans choc de sinistralité et en identifiant les profils les plus impactés.

Finalement, la forte hausse de la fréquence d'arrêt touche l'intégralité des assurés de ce portefeuille. En effet, l'incidence a augmenté de quasiment 28% au cours de cette année 2020. Toutefois, les salariés avec un revenu compris entre 15 000€ et 48 000€ annuel, les professions intermédiaires ainsi que ceux dont l'âge est compris entre 22 et 42 ans ont subi cette crise dans des proportions plus importantes. En revanche, les cadres, les salariés d'entreprises de plus de 250 personnes et les assurés âgés de plus de 48 ans ont été plus

En vue d'approfondir l'étude réalisée lors de ce mémoire, les perspectives et axes d'amélioration suivants peuvent être envisagés :

- Mener une étude sur la durée de maintien en arrêt pour aboutir à une refonte des normes tarifaires. Cependant, la sinistralité relative à cette année s'avère bien trop atypique et temporaire. Par conséquent, les normes tarifaires ainsi obtenues ne pourraient être prises en compte pour l'indexation des contrats de prévoyance. Toutefois, aboutir à une étude complète durant la période de crise sanitaire est important pour le groupe, désireux de pouvoir quantifier les conséquences de ce virus.

- Ces travaux ont servi de point de départ à la refonte des normes tarifaires du portefeuille d'AG2R La Mondiale. Inchangée depuis plusieurs années, les modélisations vont désormais se baser sur les années post-covid. Depuis cette crise, la sinistralité a considérablement évolué. C'est pourquoi, de nouvelles normes plus adaptées doivent être mises en place.

Synthesis

General context

Knowing and controlling your risk is essential for any insurance business. Indeed, the production cycle is reversed. The insurer receives contributions, fixed in advance, for benefits whose payment is conditional on the realization of the insured risks. This is why pricing plays a central role, especially for complex risks such as group benefit.

The study carried out in the course of this dissertation concerns group insurance and specifically the risk of inability. This is a multiple risk that can extend over several years. It is governed by rigorous regulations. The challenge is therefore to find a tariff segmentation capable of capturing the complexity of this risk.

Having a robust database is therefore essential to analyze this risk. However, the information systems used in group insurance do not provide a global view of our portfolio. In fact, only policyholders who have been off work stoppage for more than the deductible number of days appear in these information systems. All individuals who have no work stoppage or who have short work stoppages are therefore excluded from these data. This generates a bias. To compensate for this lack of information, data from the « Déclaration Sociale Nominative » are used in this study.

Today, the DSN is mandatory for all companies, whether in the private or public sector. It is a process that follows the payroll. It replaces all declarative formalities, linked to the employment and life of employees. The particularity of the DSN lies in the fact that it is unique, dematerialized and monthly. It makes it possible to ensure the reliability and security of employee information and rights. The DSN's contribution overcomes the limitations of the old information systems. From now on, all insured persons, whether or not they have made a claim, are known and present in the data used. Nevertheless, the construction of a reliable database adapted to our needs requires numerous adjustments.

A preliminary study of the portfolio monitoring has allowed us to refine the objectives of this thesis. It is based on different loss indicators (benefits paid, number of claims, average duration, daily allowances, age at occurrence, etc.) presented by accounting year and by year of occurrence. Indeed, the aim is to quantify the impact of the health crisis. The benefits paid to the insured have exploded during this year 2020. This sharp increase is mainly due to the significant increase in the number of work stoppages. For this reason, only the modeling of incidence, i.e., the frequency of falling into an work stoppage, will be studied. During this thesis, the models are built with a deductible equal to 0 days. However, this work is generalized to various recurring deductibles ranging from 3 to 365 days.

In agreement with the management of AG2R La Mondiale, who wanted a complete and precise study of absenteeism in 2020, the impact of this very specific claims experience is then analyzed. In addition, all of this work will serve as a starting point for the overhaul of group insurance rate standards. Nevertheless, it seems that the health crisis period was not the most representative in terms of work stoppages. Indeed, the conditions for granting certain stoppages, defined as "derogatory", suggest that the number of claims reported is somewhat biased.

Understanding the database

Now that the context and the essential notions are defined, it is time to present the construction steps to arrive at the final base. First, it is necessary to define the scope of the study. Then, as the data comes from the DSN, thus from a new information process, it must be adapted to our needs. Indeed, the raw data provided requires numerous restatements.

In addition, management rules, particularly on relapses, are applied to better understand the risk of inability. Then, we have to choose the variables to be retained, and eventually transform them in order to integrate them in our models. Finally, global statistics on the population are carried out, as well as descriptive statistics on all the variables retained. This allows us to have a first observation of the factors that influence the frequency of work stoppages. Indeed, this descriptive analysis made it possible to highlight some main points :

- Despite a peak in insureds during 2019, the population remains relatively stable between 2018 and 2020. Similarly, the average annual exposure is stable. It hovers around 75%, or 9 months of presence per year.
- The average incidence of the portfolio shows a strong increase. Indeed, the average incidence rate was 43% in the years before the health crisis. Now, in 2020, the average rate reaches 55%, an overall increase of 28%.
- The entire population is affected by this pandemic and the consequences on absenteeism are significant. However, certain categories of employees seem more exposed than others.

Introduction to Applied Models

Once the database is built, it's important to present the different models used. Indeed, three algorithms are compared.

First, the Generalized Linear Models (GLM) are implemented. Indeed, this model is very common for insurance pricing. GLMs are relatively simple processes that predict a target variable using various explanatory variables. After verifying the underlying assumptions of GLMs, the results obtained make it possible to identify the significant variables. These are the variables that will influence the frequency of work stoppages. Moreover, a coefficient is associated with each variable, so that the importance of each of them can be interpreted.

In a second step, various Machine Learning models are studied. First, the *CART* algorithm, then *Random Forest* and finally XGBoost. The XGB model turns out to be the most relevant and will be the only one presented in this thesis. The Extreme Gradient Boosting is a learning model based on the boosting method. It consists in assembling a large number of "*weak leaner*" to form a more robust "*strong learner*". It relies on a long list of hyperparameters, chosen or not by the user. At each iteration, this algorithm learns from its mistakes, thus reducing the prediction gaps.

Finally, a pricing tool named Akur8 is also used to model the incidence of work stoppages. It is based on automated Generalized Additive Models (GAM). It is a flexible solution that allows you to vary the models according to your needs. In addition, it has many graphs to evaluate the performance and analyze the results.

Nevertheless, the user's role remains central. He defines, controls and adjusts the construction of the models according to his objectives and constraints.

In addition, to obtain more robust models, a cross-validation technique is applied for each of these three solutions. The method of K-folds is used. It allows to solve under- or over-fitting problems. Moreover, a resampling of the data is performed. Rare values are exploited with a low probability, while intermediate values with a higher probability. Thus, the constructed subsample is more representative and harmonious. Finally, after cross-validation, the models are more robust and less biased.

Confrontations of the three models

The three models thus designed are pitted against each other in order to select only the best performing model and then analyze the impacts of the health crisis.

To carry out this comparison, performance indicators common to the three models must be studied. These metrics are constructed to analyze the goodness of fit of the models and to estimate the prediction errors. They are summarized in the following table.

	GLM	XGBost	Akur8
Gini	27,44%	28,99%	26,67%
R^2	4,5%	5,61%	7,99%
RMSE	1,437	1,422	1,136
MSE	2,211	2,023	1,29
MAE	0,876	0,833	0,738

The metrics are relatively close to each other, which reassures us about the relevance of the three models. Nevertheless, the Akur8 tool seems to stand out from the others thanks to its better R^2 coefficient and lower measures of estimation errors (RMSE, MSE and MAE). The XGBoost algorithm competes strongly with Akur8 since it has the best Gini coefficient, i.e. it segments the population at risk better.

The analysis of these metrics is not sufficient. This is why the study of Lift and Lorenz curves as well as an analysis of the predicted incidence, variable by variable, completes this comparison. To illustrate, let us take an example concerning the evolution of the incidence rate according to the age of the employees.

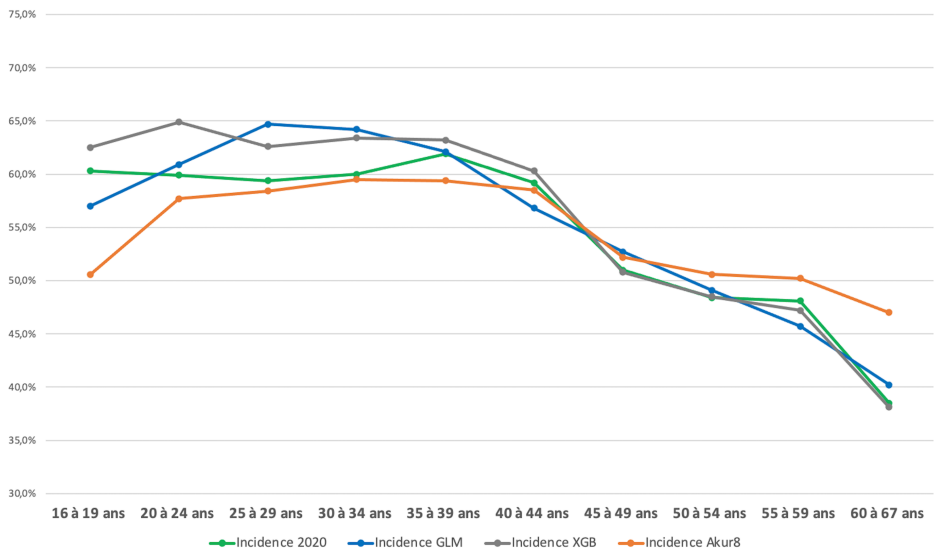


FIGURE 4 – Study of the incidence according to age

The actual incidence rate observed in 2020, depending on the age of the employees, is shown in the green curve. Then the impacts of the other three models are superimposed. The GLM, XGBoost and Akur8 models capture the general trend of stagnation up to age 35 and then decline after this threshold. Nevertheless, the XGBoost model fits the population better. It has the smallest estimation differences.

Finally, after this triple comparison, the XGBoost model appears to be the best. However, this Machine Learning algorithm has two recurring flaws. It is sensitive to "overfitting" problems and is still too "black box". Indeed, even if the construction principle is rather clear, the precise functioning of this algorithm is far too opaque. As a reminder, group benefit is a highly regulated risk. In the event of an audit, it is necessary to be able to identify each of the steps and explain all the results. Despite its very good performance, this model cannot be used in the process of constructing an impact law. On the other hand, the Akur8 tool presents slightly less efficient results but benefits from certified and auditable methods. The final choice is therefore made for this model.

Comparative study to quantify the impacts of the health crisis

As a reminder, the goal of this thesis is to quantify the impacts of the health crisis, mainly on the frequency of work stoppage. To do so, a first predictive model is implemented, based on the years prior to Covid-19, namely 2018 and 2019. Then, a second one based on data from the year 2020. It is therefore with the help of the Akur8 tool that these two models are carried out.

First, it is important to note that for these two distinct bases, the same explanatory variables come into play. Despite the change in policyholder behavior and the increase in absenteeism, both models predict an impact using the same five discriminating variables. These variables are : annual salary, number of employees per establishment, socio-professional category, age and gender.

However, certain categories of the population, whether more or less affected by this pandemic, could have varied the information to be included in the models. The two most important variables for predicting inability incidence are annual salary and company size.

As done previously, the five performance metrics are presented to compare the models before and during the health crisis.

	2018/2019 with 5 var.	2020 with 5 var.
Gini	27,59%	26,67%
R^2	7,12%	7,99%
RMSE	1,042	1,136
MSE	1,086	1,29
MAE	0,633	0,738

The results provided by the model prior to the health crisis show mostly better indicators. Indeed, this model accumulates the loss experience of two consecutive years. Consequently, this allows for greater robustness and better predictions.

Next, the Lift and Lorenz curves associated with these two models are studied. These two curves are good indicators to illustrate the distribution and segmentation of predictions.

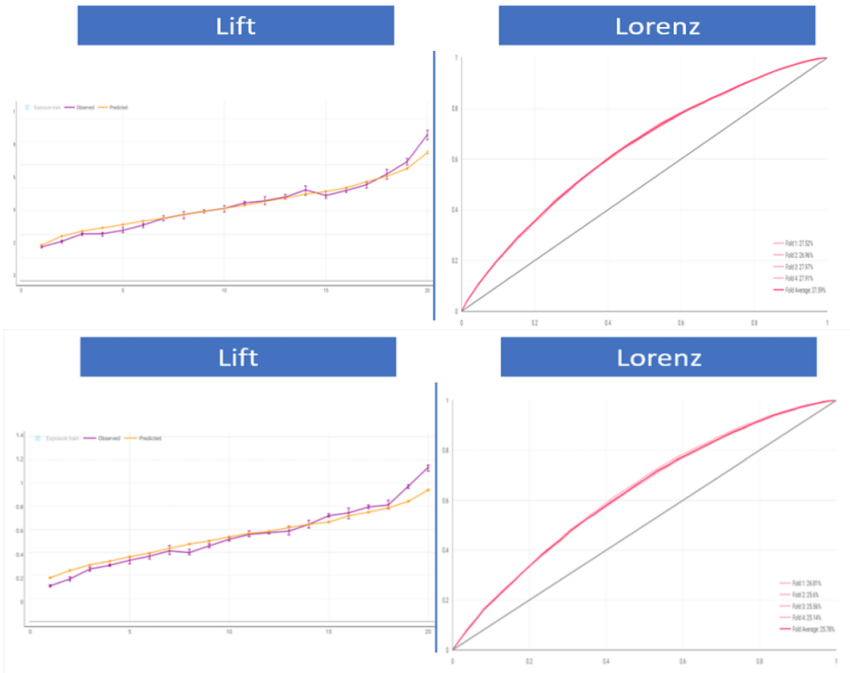


FIGURE 5 – Synthesis of Lift and Lorenz curves (base 2018/2019 on top and 2020 on bottom)

They are almost identical. The Lift curve, associated with the years 2018 and 2019, shows predictions (in yellow) slightly closer to the observations (in purple). Moreover, for both models, the predictions are slightly overestimated for low impacts, on the left of the Lift curve, and underestimated for high impacts, on the right.

In addition, the Lorenz curves are also similar. For each of the considered "*folds*", a Lorenz curve is exposed. They give an account of stable models, since they overlap almost at any point, to form the global Lorenz curve, very smooth.

Finally, all the results of the models are studied and in particular the behavior of the coefficients associated with the significant variables. These coefficients will be used to construct the incidence laws.

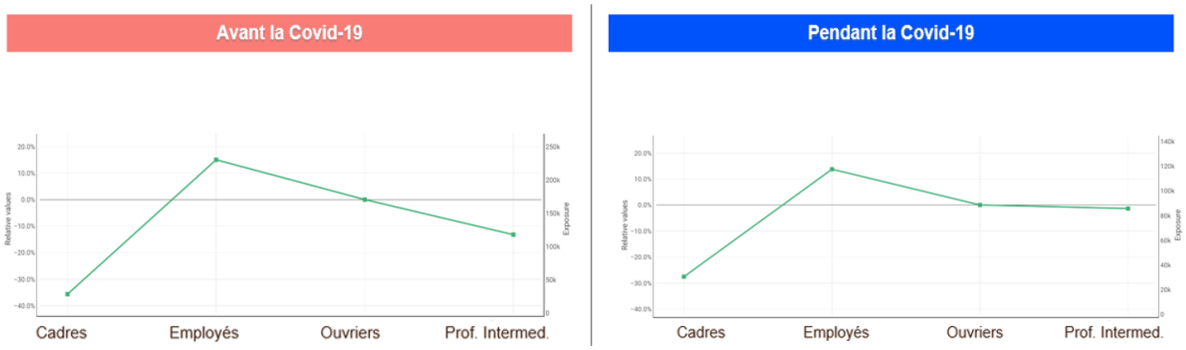


FIGURE 6 – Variation of coefficients associated with the socio-professional category

Two major points emerge from this example on the socio-professional category of employees. First, despite a general upward trend, the effect of executives, employees and workers on the target variable is similar for both models. Second, the intermediate professions are more affected than the rest of the population. Indeed, this category is mainly composed of school teachers and personnel working in health or social services. All these people are in constant contact with many people. They are therefore more likely to receive and transmit viruses.

Conclusion

Following the outbreak of the health crisis in France, an approach based on modeling the incidence of inability was conducted. The objective is to understand the consequences of this virus, by estimating the changes compared to years without a claims shock and by identifying the most affected profiles.

Finally, the strong increase in the incidence of work stoppages affects all the insureds in this portfolio. Indeed, the incidence increased by almost 28% in the year 2020. However, employees with an annual income of between €15,000 and €48,000, intermediate professions and those aged between 22 and 42 have suffered more from this crisis. On the other hand, executives, employees of companies with more than 250 employees and policyholders over the age of 48 have been more affected by the crisis.

In order to deepen the study carried out during this thesis, the following perspectives and axes of improvement can be considered :

- Conduct a study on the duration of the downtime in order to recast the tariff standards. However, the claims experience for this year is far too atypical and temporary. As a result, the rate standards thus obtained could not be taken into account for the indexation of provident contracts. However, it is important for the group to carry out a complete study during the health crisis period, in order to quantify the consequences of this virus.
- This work served as a starting point for the overhaul of the pricing standards of AG2R La Mondiale's portfolio. Unchanged for several years, the modelling will now be based on the post-covid years. Since the crisis, the claims experience has changed considerably. This is why new, more adapted tariffv standards must be put in place.

Remerciements

Tout d'abord, mes remerciements s'adressent à toute l'équipe du pôle Prévoyance de la Direction Actuariat et Acceptations Médicales d'AG2R La Mondiale. L'accueil et le cadre de travail offerts m'ont permis d'être dans les meilleures conditions pour mener à bien ce mémoire. J'ai été très heureux de faire partie de cette équipe, et j'en garde d'agréables souvenirs.

Je tiens particulièrement à remercier Michael JACCAZ, responsable de l'équipe, pour la confiance qu'il a su porter à mon travail. Son accompagnement et ses remarques constructives m'ont été d'une grande aide. Ce fut un réel plaisir de travailler avec lui.

Un immense merci à Auryane HOAREAU, ma tutrice chez AG2R La Mondiale. Elle a su m'apporter son soutien et m'accorder sa confiance tout au long des travaux que j'ai pu effectuer, le tout dans une atmosphère très agréable. Je la remercie aussi pour son implication et ses nombreuses relectures dans des périodes pas toujours favorables.

Je remercie également Lysanne RAIDOT, Fabiola DEWULF et Frédérique CHAUVIREY pour l'ensemble des conseils qu'elles m'ont partagé.

Mes pensées vont également au corps professoral de l'ISFA qui, par sa qualité d'enseignement, a su rendre ces trois dernières années très enrichissantes.

Je tiens à exprimer toute ma gratitude aux amis et aux collègues qui ont participé, de près ou de loin, à la réflexion ou à la relecture de ce mémoire. Chacun et chacune d'entre vous se reconnaîtra et je vous en suis très reconnaissant.

Enfin, je remercie ma famille pour son soutien infaillible durant toutes ces années. Une attention particulière à ma maman ainsi qu'à Lou pour leurs encouragements et leurs relectures.

Avant-propos

Pour des raisons de confidentialité, les résultats chiffrés présents dans ce mémoire ont été déformés, sans perte de généralisation. Ce choix a été fait en accord avec la direction du groupe d'AG2R La Mondiale. L'objectif de ces déformations est de toujours pouvoir analyser et interpréter les résultats, sans divulguer d'informations confidentielles. Cette décision ne vient donc pas altérer ni le sens des résultats, ni les messages de conclusion.

Toutes les modélisations et les calculs présents dans ce mémoire ont été effectués à l'aide du langage Python.

Table des matières

Introduction

Depuis toujours, l'Homme cherche à se protéger contre les risques auxquels il s'expose. C'est pourquoi, dès 1260, les premières traces de la notion de protection sociale apparaissent avec la création de l'hospice des Quinze-Vingt, sous l'impulsion des organismes religieux. Néanmoins, ce n'est qu'à partir de l'édit de 1670 que Colbert, alors ministre de la Marine sous Louis XIV, développe la notion de protection du corps marin. Certaines couvertures étaient octroyées aux marins en cas d'infirmité.

C'est ensuite, durant la III^{ème} République, que les premières vraies lois sociales se développent. En effet, cette période est marquée, sous le gouvernement de De Gaulle, par la loi d'avril 1930 sur les assurances sociales, mais surtout par l'institution de la Sécurité Sociale en 1945. De ces créations en découle le fondement du système social public.

Aujourd'hui, d'après la loi Évin de 1989, la prévoyance est définie comme l'ensemble des opérations ayant pour but de couvrir le salarié contre les risques sociaux auxquels il peut être exposé. Les risques sociaux englobent la maternité et la paternité, tous les risques qui portent atteinte à l'intégrité physique de la personne, les risques incapacité ou invalidité de travail, le risque chômage et enfin le risque décès.

Connaître et maîtriser son risque est essentiel pour toute activité d'assurance. En effet, le cycle de production est inversé. L'assureur reçoit des cotisations, fixées en avance, pour des prestations dont les versements sont conditionnés à la réalisation des risques assurés. C'est pourquoi la tarification occupe une place centrale. De plus, la prévoyance couvre un ensemble de risques particuliers qui peuvent s'étendre sur de nombreuses années.

Ensuite, la mise en place du processus de Déclaration Sociale Nominative a apporté de nombreux avantages pour les sociétés d'assurance. La DSN est obligatoire depuis janvier 2017 pour les entreprises du secteur privé. D'abord, la DSN permet d'avoir une connaissance exacte de la population sous risque. Ensuite, grâce à son fonctionnement mensuel, elle fournit des informations précises et quasi-instantanées sur l'absentéisme des assurés. Le processus de la DSN vient donc corriger les limites du système d'information en prévoyance collective. Auparavant, les organismes assureurs n'avaient connaissance que des salariés indemnisés au titre de leur contrat de prévoyance. Désormais, tous les assurés sont connus, qu'ils soient sinistrés ou non, et quelle que soit la durée de leur arrêt.

En France, en 2020, l'apparition du virus de la Covid-19 et la crise sanitaire qui en a suivi renforcent l'enjeu autour des couvertures de prévoyance. Cette pandémie est venue perturber la vie des entreprises et de leurs salariés. Elles ont dû faire face à des défis économiques et sociaux émergents. En particulier, cette pandémie a engendré une hausse brutale de l'absentéisme.

L'apparition d'une crise sanitaire sans précédent couplée à la mise en place d'un nouveau processus d'informations nous ont motivés à revoir les normes tarifaires sur le risque incapacité. En effet, le groupe AG2R La Mondiale voulait une étude complète de la sinistralité 2020. Malgré la persistance de ce virus après l'année 2020, ce mémoire se focalise uniquement sur cette année pour décrire les effets de la pandémie. En effet, le but ici est d'avoir un recul suffisant sur les données et donc sur les arrêts observés. À travers différents modèles prédictifs, l'objectif de ce mémoire est donc de comprendre et de mieux appréhender l'évolution de la sinistralité.

Au regard d'une étude sur des indicateurs de sinistralité, l'explosion du nombre de sinistres semble être la raison principale de la hausse des prestations payées. Par conséquent, seule la construction

de modèles d'incidence va être étudiée. Une confrontation des modélisations est réalisée avec d'un côté, une base de données en période de crise (en 2020) et de l'autre une base dite normale, faisant référence aux années antérieures à la Covid-19.

Dans une première partie, le contexte de cette étude est exposé. Pour commencer, le principe de la protection sociale en France et le fonctionnement de la prévoyance sont définis. Ensuite, l'apparition du processus de la DSN, son fonctionnement et ses avantages sont détaillés. Pour continuer, une analyse approfondie est réalisée sur l'absentéisme en France, spécifiquement durant la période de crise sanitaire. Enfin, cette partie se termine sur une étude de suivi du portefeuille, nécessaire pour définir les enjeux de ce mémoire. Cette analyse repose sur différents indicateurs de sinistralité, indispensables pour les calculs de provisions.

Puis dans une seconde partie, les données utilisées pour nos modélisations sont présentées. Pour débuter, le périmètre des données et les caractéristiques retenues sont définis. Ensuite, les différentes étapes de construction pour aboutir à la base de données finale sont énumérées. De plus, les différents retraitements réalisés sont présentés. Enfin, dans le but d'avoir une meilleure connaissance de la population sous risque, cette partie se termine sur l'étude et l'analyse de diverses statistiques descriptives.

Une troisième partie vient introduire les différents modèles étudiés au cours de ce mémoire. D'abord, une formulation de la théorie, de fonctionnement et des avantages des modèles linéaires généralisés est effectuée. Ensuite, l'étude bascule sur le fonctionnement de l'algorithme de Boosting, principalement dans le cas d'une utilisation avec les modèles Extreme Gradient Boosting. Enfin, cette partie sur les modélisations se clôture par la présentation de l'outil de tarification Akur8.

Pour terminer, une quatrième et dernière partie présente les modélisations effectuées ainsi que les conclusions faites en rapport à la problématique de ce mémoire. D'abord, les trois modèles mis en place (GLM, XGBoost et Akur8) sont comparés, à hypothèses identiques, afin de ne retenir que le plus performant. Ensuite, une analyse est réalisée en parallèle entre, d'un côté le modèle retenu pour la base antérieure à la crise sanitaire et de l'autre, le modèle sur la base 2020. Enfin, les évolutions et les transformations causées par la crise sanitaire sont détaillées.

1 Généralités et contexte de l'étude

Dans cette première partie, le cadre réglementaire français de la protection sociale sera défini. De plus, la notion de prévoyance, les risques qui la composent et ses mécanismes de remboursement seront introduits.

Par la suite, le processus de la Déclaration Sociale Nominative (DSN) sera prédéfinie ainsi que les différentes étapes de sa mise en œuvre. Après, le fonctionnement et la plus-value de son utilisation dans ce projet seront spécifiés.

Ensuite, la notion d'absentéisme et ses évolutions au cours des dernières années seront développées en s'appuyant sur une étude spécifique. Une attention particulière sera apportée à l'impact du virus de la Covid-19 et de ses conséquences sur les organismes de prévoyance.

Enfin, pour définir et fixer les objectifs de ce mémoire, une étude de suivi du portefeuille sera menée. Elle se basera sur l'analyse de différents indicateurs de sinistralité.

1.1 Protection sociale et régime de prévoyance en France

1.1.1 La protection sociale en France

En France, la notion de protection sociale est introduite par le conseil de l'Europe. C'est un processus réunissant tous les mécanismes permettant aux individus de faire face aux risques sociaux. Ces risques étant susceptibles d'affecter leur stabilité financière, sociale ou économique, en provoquant une baisse des ressources ou une hausse des dépenses. Parmi les risques sociaux, il y a la maternité, la maladie, le chômage, les accidents de travail, mais également la vieillesse et le décès. En résumé, la protection sociale, à travers la solidarité des français et les cotisations sociales, va permettre de compenser cette perte.

La protection sociale française s'articule autour de deux mécanismes :

- Les prestations de services sociaux : c'est l'accès à des services à tarifs réduits voire gratuits, comme par exemple la prise en charge dans les hôpitaux publics ou le financement des crèches.
- Les prestations monétaires : elles sont prises en charge par les différents organismes de protection sociale et versées directement aux ménages, en numéraires (pensions de retraite, allocations familiales ...) ou en nature (remboursement de soins de santé par exemple).

Le financement de la protection sociale est pris en charge majoritairement par les cotisations sociales. Ces cotisations sont obligatoires et versées par les employeurs ainsi que par les salariés. On peut noter que depuis 1980 les Impôts et Taxes Affectés [1] représentent une part de plus en plus importante dans le financement de la protection sociale. Parmi la cinquantaine d'ITAF existants, la Contribution Sociale Généralisée (CSG), créée en 1991, est la principale.
Enfin, les ressources de la protection sociale sont complétées par des contributions publiques provenant de l'Etat ou des collectivités locales. Elles permettent notamment de financer les dépenses de solidarité (Revenu de Solidarité Active, Fond de Solidarité Vieillesse) ou une partie des exonérations de cotisations employeurs pour les bas salaires, par exemple.

Désormais, le graphique suivant décrit l'évolution de ces ressources au cours du temps. [2]

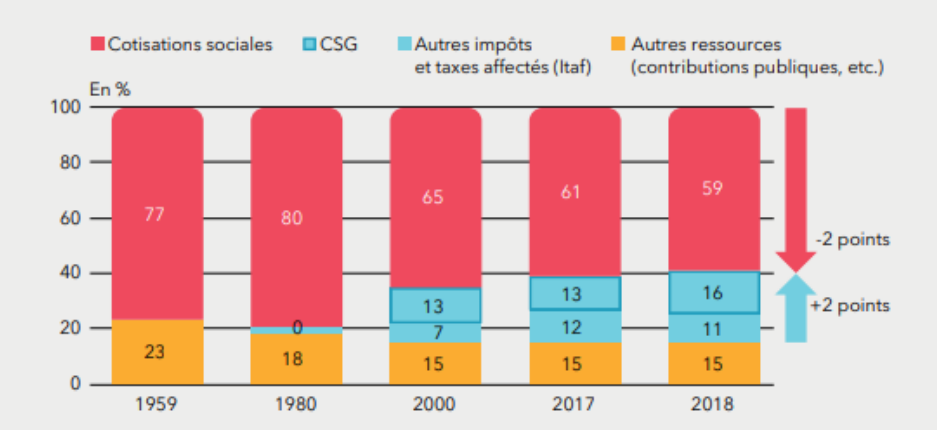


FIGURE 1.1 – Ressources de la protection sociale

1. On notera par la suite ITAF pour désigner ces impôts
2. Source : DREES, les comptes de la protection sociale en France et en Europe en 2018

La protection sociale française s'articule autour de quatre acteurs :

- La Sécurité Sociale :

  Créée en 1945, elle a pour objectif de soutenir les personnes qui résident en France face à la survenance de certains risques sociaux. Elle garantit aux salariés et à leur famille une couverture minimale et un accès aux soins essentiels de santé. Elle est un fondement du système social public et de l'économie française. Elle représente l'assurance maladie de base.

- L'assurance maladie complémentaire :

  Au côté de la sécurité sociale, elle joue aujourd'hui un rôle important en France. Elle permet aux résidents français de bénéficier d'une assurance maladie de qualité. L'assurance complémentaire couvre une partie de la dépense de soins et de biens médicaux, non remboursée par l'assurance maladie de base. En effet, les prestations versées viennent s'ajouter à celles de la sécurité sociale, et ainsi réduire le reste à charge pour les assurés.
  L'assurance maladie complémentaire se compose de différents régimes de protection sociale complémentaires. Certains d'entre eux sont obligatoires, comme par exemple les régimes complémentaires de retraite des salariés.

- L'assurance chômage :

  Depuis 1958, l'assurance chômage protège tous les salariés du secteur privé et certains du secteur public, lorsqu'ils perdent leur emploi de manière involontaire. Pour bénéficier de cette assurance, il faut valider certains critères. Si tel est le cas, cette assurance verse une allocation et favorise le retour à l'emploi grâce à diverses aides. C'est une assurance obligatoire à laquelle cotisent tous les employeurs du secteur privé et certains du secteur public.

  Le fonctionnement du service public de l'emploi est organisé avec l'aide de nombreux partenaires sociaux, autour de deux structures :

  - UNEDIC : Union Nationale Interprofessionnelle pour l'Emploi dans l'Industrie et le Commerce. Elle est indépendante de la sécurité sociale.
  - Pôle Emploi.

- Les aides sociales de l'Etat et des départements, venant en aide aux plus démunis.

Le bon fonctionnement de la sécurité sociale est basé sur deux grands principes. D'abord, l'*Universalité* de la protection sociale qui est un droit acquis par tous les individus qui résident en France. Et ensuite, la *Répartition* / *Solidarité* fait que chacun participe au financement à hauteur de ses moyens et en bénéficie selon ses besoins. Les contributions viennent alimenter « un pot commun » et sont proportionnelles aux revenus des salariés.

La sécurité sociale a donc pour vocation de couvrir l'ensemble des personnes résidant en France. Cependant, tous les individus ne sont pas couverts par le même régime. En effet, il existe trois grands régimes :

- Le régime général :

  Il est destiné aux salariés du secteur privé mais également à leurs ayants-droits. Depuis 2018, il inclut les travailleurs indépendants, non soumis à un régime spécifique (artisans, commerçants, industriels, ...), les étudiants et les bénéficiaires de certaines prestations. Aujourd'hui, ce régime couvre plus de 80% de la population française.

– Le régime agricole :

  Il assure la protection sociale des exploitants, des salariés agricoles ainsi que des entreprises agricoles. Ce régime est plus connu sous le nom de Mutualité Sociale Agricole (MSA). C'est le seul régime de la sécurité sociale à ne pas dépendre du ministère chargé des affaires sociales, mais de celui de l'agriculture.

– Les régimes spéciaux :

  Ils fonctionnent sur la base d'une solidarité restreinte à une profession ou une entreprise. Dans ces régimes, sont couverts les salariés qui ne dépendent pas du régime général.
  Il existe notamment un régime spécial pour les fonctionnaires, pour les agents de la SNCF, pour ceux d'EDF-GDF, pour les clercs de notaires, pour les mineurs, ou encore pour les professions liées au culte par exemple. Actuellement, il y a 27 régimes spéciaux différents. Certains d'entre eux sont regroupés, depuis 1994, au sein de l'UNRS.

Par ailleurs, Il existe d'autres régimes légaux, comme par exemple les régimes autonomes de vieillesse, le régime chômage ou encore le régime de retraite complémentaire AGIRC-ARRCO.

Le régime général de la sécurité sociale est organisé en cinq branches qui sont les suivantes [3].

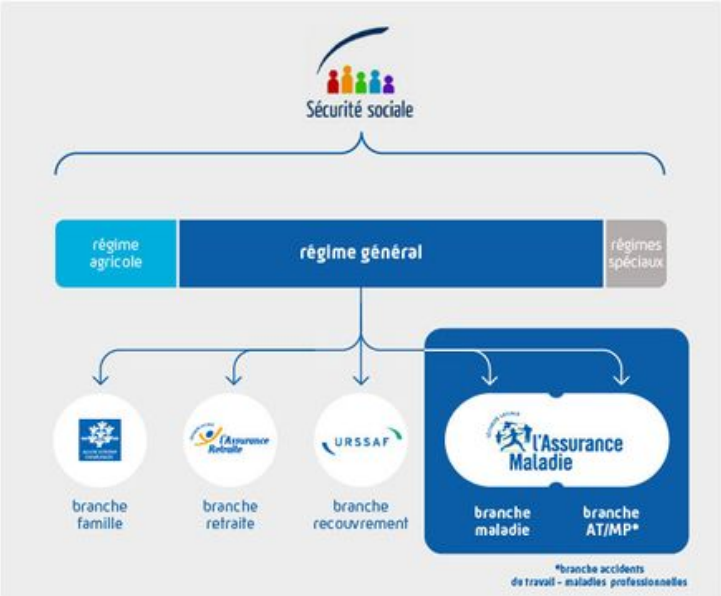


FIGURE 1.2 – Les régimes de la protection sociale

Ces cinq branches correspondent aux différents risques sociaux. Elles se distinguent par leur organisation dans la prise en charge des différents risques couverts :

– Branche famille :

  La mission de cette branche est d'atténuer les inégalités de niveaux de vie, en fonction du nombre d'enfants par ménage. Elle gère les prestations familiales, s'occupe de l'accompagnement des familles dans leur quotidien, de l'accueil des enfants, de l'accès au logement et de la lutte contre la précarité ou le handicap. Elle est gérée par la CNAF et par les différentes CAF départementales pour les actions sociales locales.

3. Source : l'assurance maladie, Ameli : `https://assurance-maladie.ameli.fr/qui-sommes-nous/organisation/securite-sociale`

- Branche retraite :
  Elle est gérée par l'assurance retraite et plus particulièrement la CNAV. A partir des données de la DSN, elle prend en charge l'inscription des revenus sur le compte vieillesse de chaque salarié, tout au long de leur vie active. Par la suite, elle calcule le montant des retraites et verse les pensions ou les allocations veuvage dues. Elle est actuellement au cœur de multiples réformes.
- Branche recouvrement :
  Elle est chargée de collecter l'ensemble des cotisations et de les redistribuer aux différentes caisses de la sécurité sociale. Cette branche gère également la trésorerie de la sécurité sociale. A l'échelle nationale, cette mission est gérée par les URSSAF, qui sont des organismes privés.
- Branche maladie :
  Elle assure la prise en charge des dépenses de santé des assurés et garantit l'accès aux soins. Elle couvre les risques maladie, maternité, invalidité ou encore décès.
- Branche AT/MP (Accident de Travail et Maladies Professionnelles) :
  Elle s'occupe des risques professionnels auxquels sont confrontés les travailleurs, gère leur protection et mène des actions de prévention.
  Ces deux dernières branches sont principalement gérées par la CNAM.

Au cours des vingt dernières années, la sécurité sociale, par l'intermédiaire du versement des retraites et de la prise en charge des soins, a permis l'augmentation de l'espérance de vie. Cependant, ces progrès impliquent des difficultés de financement. En effet, l'allongement de la durée de vie entraîne des retraites plus longues et des dépenses de santé plus importantes, qui déséquilibrent le système. Par conséquent, de nouvelles mesures, notamment concernant la durée des cotisations versées, doivent être prises.

1.1.2 La prévoyance complémentaire

Comme indiqué précédemment, la sécurité sociale couvre une grande majorité de la population française. Cependant, elle ne permet pas de garantir l'intégralité de la perte de revenu ou de la hausse des dépenses. Afin de combler ce manque et pour compléter la part de la sécurité sociale, l'entreprise, en tant qu'employeur, peut souhaiter, ou être contrainte d'offrir à ses salariés, un régime de protection sociale complémentaire en santé, en retraite ou en prévoyance.

Les différents régimes complémentaires de prévoyance peuvent être proposés par des sociétés d'assurance, des mutuelles ou encore des institutions de prévoyance.

Selon la loi n°89-1009 du 31 décembre 1989, dite loi EVIN, la prévoyance regroupe : « les opérations ayant pour objet la prévention et la couverture du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail ou d'invalidité ou du risque chômage ». En résumé, la prévoyance est une manière de se protéger contre les aléas de la vie.

C'est donc en contrepartie de cotisations versées par les assurés, que les régimes complémentaires de prévoyance vont pouvoir indemniser les salariés et leurs ayants-droits. Ces prestations viennent s'ajouter à celles de la sécurité sociale.
En fonction des garanties souscrites, la prévoyance permet de faciliter l'accès au soin, d'assurer un maintien partiel ou total des revenus, ou encore de percevoir un capital ou une rente.

1.1.2.1 Les régimes obligatoires

La mise en place d'un régime complémentaire de prévoyance est en général facultative. Cependant, elle peut devenir obligatoire dans certaines situations :

- Prévoyance décès obligatoire pour les cadres :

  Depuis la convention collective nationale de retraite et de prévoyance des cadres de 1947, les entreprises du secteur privé doivent proposer une couverture de prévoyance, couvrant leurs cadres tant qu'ils font partie de leurs effectifs et jusqu'à leur retraite.
  Cette couverture doit être affectée en priorité au risque décès, et doit au minimum couvrir le décès du salarié adhérent. En cas de non-respect de cette obligation et du décès d'un de ses cadres, l'entreprise s'expose à verser à ses ayants droits un capital égal à trois fois le Plafond Annuel de la Sécurité Sociale.
  Les contrats de prévoyance proposés aux salariés cadres et assimilés-cadres peuvent tout à fait être mis en place pour les autres salariés de l'entreprise.

- Les conventions collectives ou accords de branches :

  Les accords de branches proviennent de négociations entre les syndicats de salariés et les organisations patronales. Tout ce qui est signé dans ces accords entrainent des obligations conventionnelles, qui sont consultables dans les conventions collectives. Les entreprises qui dépendent de ces accords doivent les appliquer au même titre que la loi. Ces obligations peuvent porter sur des garanties, en cas d'arrêt de travail, d'invalidité ou encore de décès.
  Il existe plus de 300 conventions collectives nationales (CCN) applicables aux salariés en France. Selon une étude menée en 2018 par le CTIP [4], 215 branches avaient signé un accord de prévoyance.

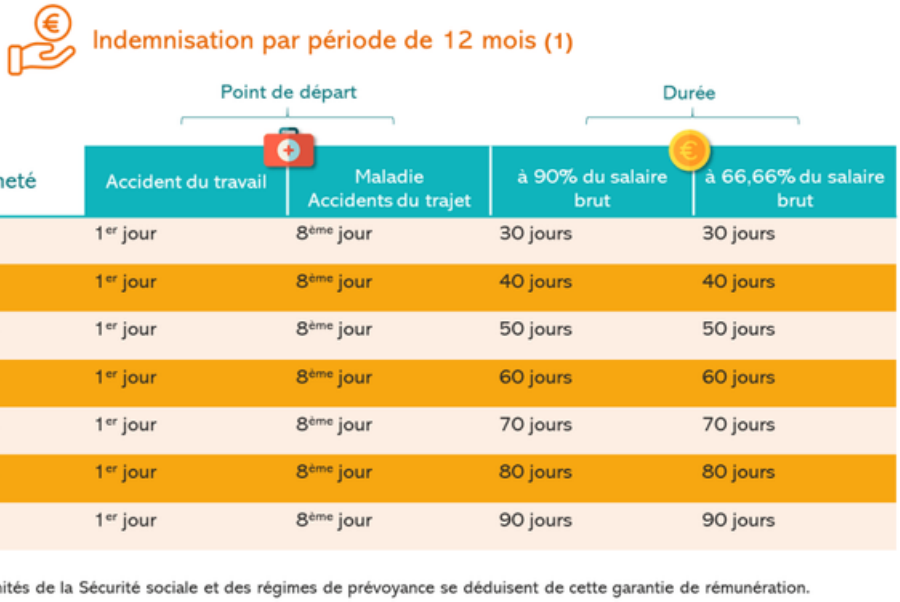
  En revanche, lorsqu'un accord collectif ne prévoit pas l'adhésion à une prévoyance, celle-ci peut être décidée par décision unilatérale de l'employeur. Cependant, seuls les salariés embauchés après cette décision d'adhésion seront concernés par cette obligation.

- La loi de mensualisation :

  Cette loi de mensualisation intervient dans le cadre de la protection financière contre la baisse de revenu. Elle fait suite à l'accord national interprofessionnel du 10 décembre 1977 et est inscrite dans la loi en janvier 1978. Elle va par la suite être modifiée en juin 2008.
  Cette loi impose aux employeurs de maintenir, en tout ou partie, la rémunération des salariés absents pour cause de maladie ou accident de travail, s'ils remplissent certaines conditions. En effet, pour en bénéficier, il faut qu'au premier jour d'arrêt, le salarié puisse justifier un minimum d'un an d'ancienneté dans l'entreprise, que l'arrêt soit médicalement constaté et indemnisé par la sécurité sociale.
  Un délai de carence de 7 jours s'applique en cas d'accident ou de maladie non professionnels, ce qui signifie que la couverture se déclenche qu'à partir du 8ème jour d'arrêt. En revanche, il n'y a aucun délai de carence en cas d'AT/MP.

  La durée de maintien du salaire varie selon l'ancienneté du salarié dans l'entreprise. Ceci est résumé dans le tableau ci-dessous [5].

4. Centre Technique des Institutions de Prévoyance
5. Source : étude sur le maintien de salaire, assurance et garantie incapacité en arrêt de travail, réalisée par Apicil : https://pro.apicil.com/actualites/la-mensualisation-expliquee/



Indemnisation par période de 12 mois (1)

Ancienneté	Point de départ		Durée	
	Accident du travail	Maladie Accidents du trajet	à 90% du salaire brut	à 66,66% du salaire brut
1 à 6 ans	1er jour	8ème jour	30 jours	30 jours
6 à 11 ans	1er jour	8ème jour	40 jours	40 jours
11 à 16 ans	1er jour	8ème jour	50 jours	50 jours
16 à 21 ans	1er jour	8ème jour	60 jours	60 jours
21 à 26 ans	1er jour	8ème jour	70 jours	70 jours
26 à 31 ans	1er jour	8ème jour	80 jours	80 jours
+31 ans	1er jour	8ème jour	90 jours	90 jours

(1) Les indemnités de la Sécurité sociale et des régimes de prévoyance se déduisent de cette garantie de rémunération.

FIGURE 1.3 – loi de mensualisation, maintien de salaire

1.1.2.2 Les régimes facultatifs

Dans la majorité des cas, les régimes de prévoyance complémentaires sont facultatifs. Il existe différentes bonnes raisons pour un employeur de mettre en place un système de prévoyance. Premièrement, cela permet à l'entreprise de gagner en attractivité. Elle va bénéficier d'une politique sociale forte qui va être un atout pour l'embauche de nouveaux candidats. Deuxièmement, cela va lui permettre de fidéliser ses effectifs. En effet, elle va prendre en charge une partie de leurs cotisations et ainsi leur faire faire des économies. Enfin, cela lui permet de bénéficier d'un régime social et fiscal avantageux. Effectivement, sous certaines conditions, les cotisations de prévoyance sont déductibles de l'impôt sur les sociétés et sont exonérées de charges sociales.

Il existe deux grands types de contrats en prévoyance :

- Le contrat individuel :
  C'est un contrat qui provient d'une démarche volontaire et individuelle du salarié. Il peut donc souscrire seul des garanties de prévoyance et être couvert en cas d'accident de la vie. Il bénéficie ainsi de prestations financières compensatoires. Le montant des cotisations, des prestations et des conditions de versement sont définis au moment de la souscription du contrat. Les sommes perçues sont versées directement au salarié ou aux personnes désignées dans le contrat (conjoints, enfants, ...). Elles ont pour vocation de maintenir le niveau de vie du salarié ainsi que celui de ses proches.
- Le contrat collectif :
  La prévoyance collective est un contrat « groupe » souscrit par une entreprise au profit de ses salariés. En général, cette souscription n'est donc pas obligatoire. Cependant si l'entreprise a décidé d'y souscrire, il y a certaines conditions à respecter. En effet, le contrat doit soit concerner l'ensemble des salariés, soit un ensemble de salariés qui n'est pas distinguable selon des critères de discrimination (âge, poste occupé, genre, handicap, ancienneté, rémunération, état de santé, etc).

1.1.2.3 Les risques couverts

Les garanties de prévoyance collective s'articulent autour de trois risques principaux :

– L' Incapacité :

L' incapacité désigne une inaptitude physique ou psychologique à exercer une activité professionnelle de façon temporaire ou permanente, totale ou partielle.

Il existe deux types d' incapacité :

- o L' incapacité temporaire de travail : durant une période de temps évolutive définie lors de la survenance de l'arrêt, le salarié ne peut plus travailler. Cette incapacité peut être partielle (ITP) ou totale (ITT).
- o L' incapacité permanente de travail : de manière définitive, le salarié ne peut plus travailler, ou il ne peut plus exercer certains postes particuliers. La aussi, l'incapacité peut être partielle (IPP) ou totale (IPT).

Afin de compenser la baisse de revenu liée à l' incapacité, la sécurité sociale verse des indemnités journalières à la personne en arrêt. Il est à noter qu'un délai de franchise, variable en fonction du risque, est toujours prévu. De plus, dans le cadre de la loi de mensualisation, cette personne bénéficie d'une indemnisation de la part de son employeur. Cependant, ces aides sont en général insuffisantes. C'est pourquoi les salariés peuvent souscrire un contrat de prévoyance complémentaire, avec des garanties qui s'additionnent avec les indemnités de la sécurité sociale.

L' incapacité dure au maximum trois ans (36 mois). Après ces trois années, si l'assuré n'est toujours pas en mesure de reprendre son activité, il devient invalide.

– L' invalidité :

L' invalidité concerne un état physique et/ou psychique irréversible. Un assuré est reconnu invalide si sa capacité de travail ou de revenu est réduite d'au moins deux tiers, à la suite d'une maladie ou d'un accident de la vie courante (hors cause professionnelle).

Selon le code de la sécurité sociale, il existe trois catégories d' invalidité :

- ◦ Catégorie 1 :

  L'assuré a perdu deux tiers de ses capacités de travail mais peut néanmoins exercer une activité professionnelle rémunérée. Il n'est donc pas totalement inapte au travail mais peut l'être à certains postes. Une personne de cette catégorie bénéficie alors d'un aménagement de travail au niveau de ses horaires et ou de ses missions.

- ◦ Catégorie 2 :

  Le salarié a perdu deux tiers de ses capacités et ne peut normalement pas exercer un travail, quel qu'il soit.

- ◦ Catégorie 3 :

  Dans cette classe, le salarié est incapable de travailler. De plus, il est assisté d'une tierce personne pour les actes de la vie quotidienne.

Une fois que la sécurité sociale reconnait le statut d'invalide, le salarié perçoit une pension d'invalidité, en compensation de la perte de salaire provoquée.

La rente est revalorisée chaque année afin de suivre l'évolution du coût de la vie. Elle peut également être majorée lorsque l'assuré a des enfants à charge. Au plus tard, les prestations sont versées jusqu'à l'âge de la retraite. Néanmoins, les prestations peuvent s'arrêter si l'individu retrouve tout ou partie de ses capacités.

– Le décès :
  Le décès est toujours couvert, qu'il provienne des suites d'un accident ou d'une maladie. En revanche, certains cas peuvent être exclus. Ils doivent être clairement mentionnés dans les clauses du contrat. Cette garantie est destinée à compenser la perte de ressources subie par la famille, à la suite du décès de l'assuré. Le contrat doit désigner un ou des bénéficiaires qui se verront recevoir un certain capital en cas de décès de l'assuré.
  En plus de ce capital, les bénéficiaires peuvent recevoir d'autres prestations facultatives qui seront alors précisées dans le contrat. Il peut s'agir par exemple d'une rente pour le conjoint survivant ou d'une rente éducation pour les enfants du défunt. Le contrat peut également comporter une allocation pour frais d'obsèques.

Dans la suite de ce mémoire, seul les risques de mensualisation et d'incapacité seront étudiés.

Maintenant que le cadre de la protection sociale française est posé et que la notion de prévoyance est définie, nous allons introduire le processus de la Déclaration Sociale Nominative.

1.2 La Déclaration Sociale Nominative

1.2.1 L'avant DSN

Les entreprises font partie du tissu économique d'un pays. Dans ce cadre, elles participent indirectement au budget de l'État à travers différentes contributions, comme la TVA, le salaire des employés, etc. Cependant, afin d'informer l'administration française et ainsi assurer le financement de la protection sociale (maladie, chômage, retraite, ...) les entreprises doivent fournir un certain nombre de déclarations sociales auprès de différents organismes sociaux.

Ce sont des tâches administratives obligatoires pour tous les employeurs, qui peuvent être périodiques ou événementielles. En effet, il existe des déclarations mensuelles ou trimestrielles, comme par exemple la Déclaration Unifiée des Cotisations Sociales par Echange de Données Informatisées (DUCS EDI). D'autres déclarations ont une périodicité annuelle : c'est le cas de la Déclaration Automatisée des Données Sociales Unifiées (DADSU). Enfin, certaines déclarations sont obligatoires mais à n'effectuer que lors de la réalisation d'événements : c'est le cas pour les attestations employeurs à destination de Pôle Emploi (DNAE AED) ou les attestations de salaire pour le paiement des indemnités journalières (IJSS).

Avant l'instauration de la DSN[6], il existait deux actes distincts :

– La paie : utilisation des données fournies par le logiciel et le bulletin de paie. Cet acte est régi par le code du travail.

6. Dans la suite de ce mémoire on remplacera l'expression Déclaration Sociale Nominative par son acronyme : DSN

– Les déclarations : une logique déclarative, orientée sur la protection sociale. Envoi des déclarations vers les différents organismes sociaux.

Ces deux actes reposent sur des données communes mais sont fondés sur des codifications et des besoins différents.

Avant la mise en place de la DSN, l'employeur devait s'adresser à de nombreux organismes, dont les principaux sont l'URSSAF, Pôle emploi, la CNAV, le Centre des impôts ou encore les différentes caisses des régimes spéciaux. Ce processus de transmission d'informations était long et fastidieux pour les entreprises mais également pour les tiers déclarants. En effet, les entreprises remplissaient chaque déclaration indépendamment les unes des autres.

1.2.2 Les phases de mise en place de la DSN

La loi « Warsmann II » relative à la simplification du droit et à l'allègement des démarches administratives, qui a été adoptée en 2012, a permis de simplifier toutes ces lourdeurs administratives, en assouplissant leur fonctionnement. En effet, c'est au cours de cette loi que fut instauré la Déclaration Sociale Nominative. Par la suite, elle a été mise en place de manière progressive et renforcée au cours de trois phases distinctes.

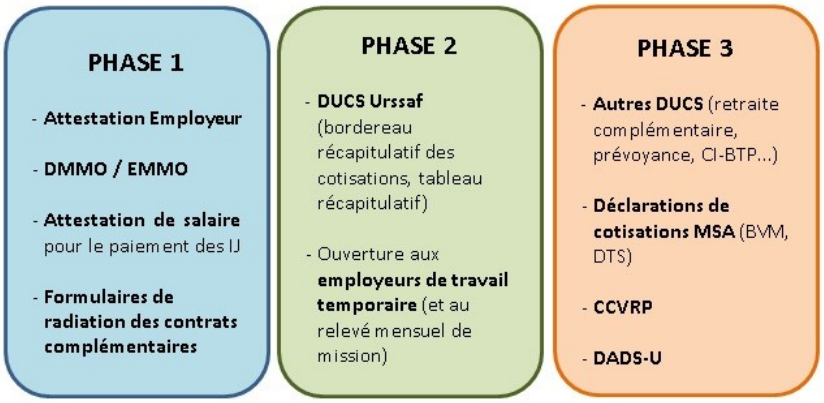


FIGURE 1.4 – Les différentes phases de mise en place de la DSN

– **DSN phase 1 :** depuis le 1er avril 2013, les entreprises volontaires peuvent transmettre leur DSN en remplacement de :

  - La déclaration mensuelle des mouvements de main d'œuvre (DMMO).
  - L'attestation employeur pour pôle emploi
  - L'attestation de salaire pour le versement des IJ auprès de la CNAM et la MSA [7] en cas de maladie, maternité ou paternité.
  - Les formulaires de radiation pour les institutions de prévoyance, les mutuelles ou les sociétés d'assurance.

  Néanmoins, il ne s'agit encore que d'une phase de test.

7. CNAM MSA : Caisse Nationale d'Assurance Maladie et Mutualité Sociale Agricole

– **DSN phase 2 :** depuis le 1er avril 2015, elle devient obligatoire pour certains employeurs qui, sur l'année 2013, remplissent les conditions suivantes :

  - Déclarer plus de 2 millions d'impôts, de cotisations ou de contributions sociales.
  - Déclarer plus d'un million d'impôts, tout en ayant eu recours à un tiers-déclarant (en général un expert-comptable), qui a lui-même déclaré plus de 10 millions d'euros pour l'ensemble de son portefeuille clients.

  Au cours de cette phase, de nouvelles déclarations disparaissent et sont désormais regroupées au sein de la DSN. C'est le cas notamment de la déclaration unifiée des cotisations sociales (DUCS) et des attestations d'arrêt de travail AT/MP.

– **DSN phase 3 :** depuis le 26 septembre 2016, la DSN remplace de nouvelles formalités déclaratives, comme par exemple la régularisation des cotisations sociales de l'année, le tableau récapitulatif ou le relevé mensuel des contrats de travail temporaires. À partir du 1er janvier 2017, elle devient obligatoire pour toutes les entreprises relevant du régime général.

Depuis janvier 2019, la DSN remplace définitivement la déclaration annuelle des données sociales unifiées (DADS-U) et permet de déclarer le prélèvement à la source. Elle alimente également le compte personnel de formation et le compte pénibilité. À ce jour, elle remplace plus d'une trentaine de déclarations périodiques ou évènementielles.

1.2.3 Présentation et fonctionnement de la DSN

Aujourd'hui, la DSN est obligatoire pour tous les employeurs rémunérant des salariés affiliés au régime de la sécurité sociale. C'est un fichier produit à partir de la paie. Il permet à tout employeur de déclarer auprès des organismes et administrations concernés, un ensemble d'informations nécessaires à la gestion de la protection sociale du salarié. La particularité de cette déclaration réside dans le fait qu'elle soit unique, dématérialisée et mensuelle.

De plus, la DSN constitue aujourd'hui le principal canal de transmission entre, d'un côté les données sociales d'une entreprise, et de l'autre les organismes de protection sociale. Le principe fondamental de la DSN est donc de remplacer toutes les formalités déclaratives, liées à l'emploi et la vie des salariés, par une seule et unique déclaration. Pour rappel, toute ces formalités administratives servent principalement à déclarer des événements du type arrêt de travail, fiche de paye ou attestation employeur destinée à pôle emploi. En outre, cela permet de fiabiliser et de sécuriser les informations transmises.

La DSN fonctionne avec le numéro de SIRET. Elle doit donc être émise pour chaque établissement d'affectation et inclure tous les salariés y étant rattachés. De plus, afin que la déclaration soit validée, il faut obligatoirement qu'elle comporte certaines informations. Par exemple, elle doit contenir le numéro de sécurité sociale [8] du salarié, l'ensemble des salaires versés ainsi que les cotisations payées aux différents organismes. Le numéro de SIRET de l'établissement ou encore les numéros des contrats de mutuelle ou de prévoyance doivent également apparaitre.

Cette déclaration peut être remplie de différentes manières. Elle peut être gérée, soit en interne par l'employeur, le service des ressources humaines ou les équipes de gestion qui procèdent eux-mêmes

8. le numéro de sécurité sociale est le numéro d'identification des assurés sociaux. On l'appelle également le Numéro d'Inscription au Répertoire

à la déclaration mensuelle, soit l'entreprise peut confier cette tâche à un tiers déclarant tel qu'un expert-comptable par exemple. Toutefois, l'employeur et le tiers déclarant peuvent interagir afin que chacun ne remplisse qu'une partie de la DSN. [9]

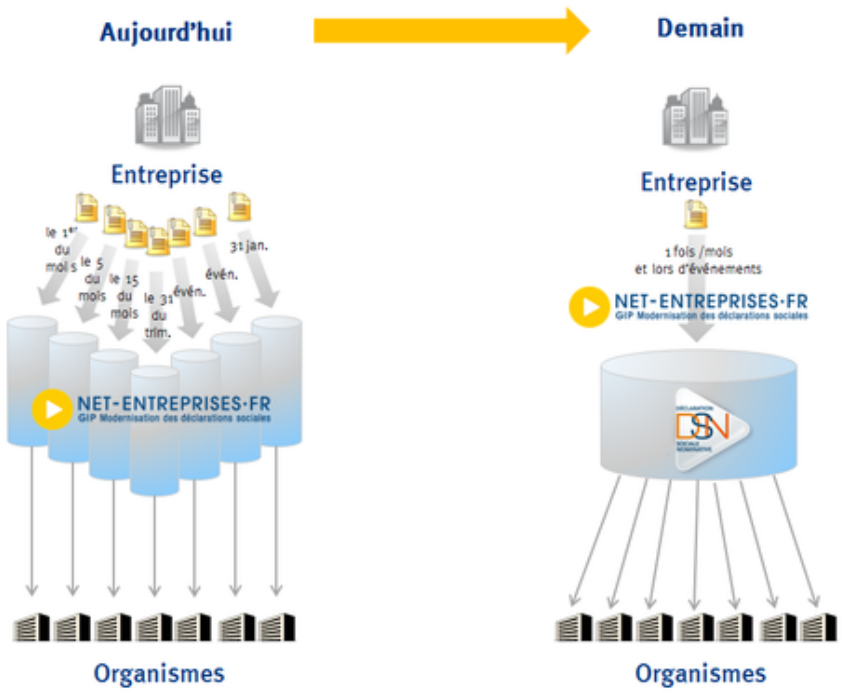


FIGURE 1.5 – Mise en place de la DSN

La transmission de la DSN est un acte règlementé qui doit être réalisé une fois par mois et qui respecte donc un calendirer spécifique, avec des dates butoires à ne pas dépasser. Toutefois, la DSN permet également de prendre en compte les événements ponctuels, comme par exemple un arrêt de travail (maladie, maternité, paternité...), une rupture de contrat de travail ou un accident de travail. Ces événements sont dits exceptionnels et doivent être déclarés dans les 5 jours suivant leur connaissance, grâce à *« la DSN signalement d'événement »* qui est distincte de la DSN mensuelle.

À la suite de son dépôt, la DSN est soumise à de nombreux contrôles. En fonction des résultats, le déclarant doit éventuellement apporter des modifications. L'ensemble des retours et des corrections apportées doivent être faits au plus tôt.

De plus, en cas de non-respect de ces règles, en cas de défaut de production ou d'omission, les entreprises s'exposent à de nombreuses sanctions. Par exemple, une pénalité d'une valeur de 1,5% du plafond de la sécurité sociale par salarié et par mois de retard peut être infligée à l'employeur. Les pénalités sont calculées en fonction des effectifs connus ou transmis lors de la dernière déclaration de l'employeur.

1.2.4 Avantages et apports de la DSN

La DSN apporte de nombreux aspects positifs pour toutes les parties prenantes d'un processus administratif :

9. Source du schéma : `https://www.net-entreprises.fr/tableau-de-bord-dsn/`

1) Les entreprises :

Tout d'abord, cela représente un net bénéfice pour les entreprises. Effectivement, la DSN constitue une simplification radicale de leurs démarches déclaratives. Les employeurs ne réalisent désormais qu'une unique déclaration, qui fait suite à la paie.

Des contrôles sont effectués mensuellement, ce qui permet de sécuriser et de fiabiliser les données, de réduire la charge de travail et de minimiser les risques d'erreurs, d'oublis ou de pénalités. De plus, ces contrôles sont clairs et complets, ce qui augmente l'assimilation par les entreprises.

De manière générale, la DSN équivaut à une meilleure maitrise des données. En bref, cela améliore la performance et la productivité au sein des entreprises.

2) Les salariés :

La DSN simplifie les démarches et permet un traitement plus rapide de leur dossier. De plus, leurs informations personnelles et confidentielles sont de plus en plus sécurisées et robustes. En bref, elle leur offre une situation actualisée plus rapidement.

3) La collectivité :

De plus, pour la collectivité, la DSN permet un meilleur suivi des données relatives aux entreprises. En effet, ces dernières se verront recevoir, chaque mois, des données sur l'évolution des entreprises et de leurs emplois. Cela va leur permettre de lutter contre la fraude et améliorer le suivi des politiques publiques. Afin de faire en sorte de diminuer les coûts de gestion, les acteurs institutionnels optimisent le processus de la DSN dans leurs projets.

4) Les tiers déclarants :

Les tiers déclarants bénéficient des mêmes avantages que les entreprises en matière d'organisation de travail, de rationalisation et de sécurisation des transmissions.

En outre, grâce au processus de la DSN, les entreprises peuvent recueillir une grande diversité d'informations et avoir des précisions beaucoup plus fines.

Ceci constitue donc une révolution dans le monde de l'entreprise, notamment pour les sociétés d'assurances. En effet, la mensualisation du processus permet de disposer d'informations fiables et actualisées sur les salariés mais également sur les sinistres (ce n'était pas le cas auparavant pour les instituts de prévoyance). En effet, aujourd'hui, les arrêts de travail sont vus dans la DSN avant d'être présents dans leurs bases de données, à cause du délai de franchise et de gestion.

Très utile en assurance collective, cette déclaration va permettre aux assureurs d'avoir plus de détails sur leur portefeuille d'assurés. Désormais, les assureurs ont une vision bien plus large sur l'ensemble des salariés. Cela concerne tous les types de contrats, pour des salariés sinistrés ou non, de durée inférieure ou non à la franchise. Tout cela va permettre de construire des modèles plus justes, plus précis et ainsi avoir des tarifs mieux adaptés à la population sous risque.

En revanche, l'utilisation de la DSN présente quelques inconvénients.

Premièrement, le fait de disposer d'une telle quantité d'information peut être un challenge pour les assureurs de personnes. En effet, rendre exploitable et accessible toutes ces données n'est pas chose facile.

Deuxièmement, la DSN est répartie en différents blocs. Par conséquent, pour mener à bien un projet de data sciences ou d'actuaires, il est souvent indispensable de faire des correspondances entre toutes ces tables.

Enfin, comme c'est un processus relativement récent, les entreprises ont besoin d'un temps d'adaptation. Il peut être très variable d'une structure à l'autre. Par manque de maîtrise de la DSN, certaines entreprises commettent des erreurs lors des déclarations. Les contrôles sont effectués dans le but de réduire ces erreurs au maximum.

Dans l'optique de construire des modèles prédictifs de l'incidence en arrêt de travail, l'utilisation de la DSN est un point incontournable de ce mémoire. En effet, les données utilisées sont extraites de la DSN, durant une période comprise entre 2018 et 2020. Dans la suite, nous allons définir l'absentéisme en entreprise et détailler les conséquences de la pandémie du Covid-19 sur ce phénomène.

1.3 La notion d'absentéisme en période de crise sanitaire

Dans cette section, le phénomène de pandémie est défini ainsi que les raisons pour lesquelles le coronavirus en fait partie. Par la suite, les différentes évolutions de l'absentéisme sont détaillées, en particulier en période de pandémie mondiale. Après, les enjeux de l'absentéisme pour une société d'assurance sont énumérés. Enfin, les conséquences des différentes contraintes réglementaires mises en place au cours de cette pandémie sont analysées.

1.3.1 Le terme de pandémie

Le terme pandémie provient du Grec ancien. C'est la concaténation du terme « pãn », qui veut dire tous et de « dẽmos » qui désigne le peuple. L'OMS apporte quelques précisions à cette définition : le mot pandémie est employée en cas de propagation mondiale d'une nouvelle maladie.

La Covid-19 est une maladie induite par un microbe invisible à l'œil nu. Au microscope, ces virus semblent entourés d'une petite couronne. C'est pourquoi le nom de coronavirus leur est attribué.

Néanmoins, ces virus ont la particularité d'être très contagieux. Ils se propagent par voie aérienne ou simplement au contact d'objets contaminés. Les personnes infectées par ces virus peuvent réagir de différentes manières.

Les coronavirus ne sont pas nouveaux ni inconnus. En effet, il y a eu différentes épidémies de coronavirus, notamment en 2002-2003 avec le SARS-CoV et en 2012 avec un virus nommé MERS-CoV. Cependant, c'est en novembre 2019 qu'un nouveau virus fait son apparition sous le nom de SARS-CoV-2. Bien que l'origine exacte de ce virus semble encore floue, les premiers cas sont recensés

dans la ville de *Wuhan* en Chine. Les personnes infectées vont rapidement transmettre la maladie dans toute la ville, puis dans tout le pays. En quelques semaines, ce n'est pas moins de 20 000 personnes qui sont contaminées. Trois mois plus tard, le virus se propage sur la planète entière et aujourd'hui quasiment tous les pays du monde ont été touchés par ce virus.

Cependant, il faut attendre mars 2020 pour que l'Organisation Mondiale de la Santé (OMS) qualifie la Covid-19 de pandémie, en précisant que c'est la première fois qu'une pandémie est causée par un coronavirus. L'OMS définit six niveaux d'alerte qui se basent sur le mode de transmission et sur l'ampleur de propagation du phénomène.

1.3.2 La notion de l'absentéisme

Il est clair que l'augmentation du nombre de malades va avoir un impact sur le nombre et la durée des arrêts de travail. Avant d'étudier cet impact, nous allons définir la notion d'absentéisme au travail.

L'absentéisme représente un indicateur important pour la gestion des ressources humaines et doit devenir une préoccupation, voir un signal d'alerte, pour tous les acteurs de l'entreprise. En effet, ce phénomène tient compte principalement des conditions de travail, de son organisation et de son management.

Néanmoins, les causes de l'absentéisme peuvent être très diverses d'une entreprise à l'autre. Elles sont spécifiques à un secteur d'activité, une branche ou une manière de fonctionner. L'absentéisme représente un enjeu majeur pour les entreprises. Elles ont pour but de réduire au maximum cet indicateur en essayant de trouver les solutions les plus appropriées.

Malgré des origines variées, l'absentéisme se caractérise souvent par l'apparition de certains facteurs récurrents. Dans un premier temps, l'absence peut provenir de motifs professionnels, comme par exemple les accidents de travail, les maladies professionnelles ou encore de mauvaises conditions de travail. Les mauvaises conditions de travail peuvent être la conséquence d'une dégradation de l'ambiance générale, d'un niveau de stress trop important mais aussi d'horaires de travail inadaptés. Dans un second temps, les causes d'absence des salariés découlent régulièrement de motifs personnel. En effet, l'influence de l'environnement familial, les problèmes de santé non liés au travail ou les absences dites de conforts, c'est-à-dire sans aucun motif valable, sont responsables de nombreux arrêts.

Le taux d'absentéisme se calcule en faisant le rapport entre le nombre de jours calendaires d'absence, divisé par le nombre de jours calendaires de présence sur une année. Ce taux est donné en pourcentage.

1.3.3 L'évolution de l'absentéisme

La crise de la Covid-19 est venue bouleverser la vie des entreprises et celle de leurs salariés. Ces entreprises ont donc dû faire face à des défis économiques et sociaux sans précédent. Cette crise va donc avoir des conséquences majeures sur l'absentéisme, principalement sur les années 2020 et 2021.

Grace à une étude intitulée *« le 13ème baromètre de l'absentéisme et de l'engagement »* menée par *Ayming* en partenariat avec AG2R La Mondiale, nous allons pouvoir présenter quelques chiffres clés.

Cette étude rassemble :

- D'un côté une enquête quantitative réalisée en France sur l'année 2020. Elle concerne 49 227 entreprises du secteur privé employant pas moins de 5 008 478 salariés en CDI. Les données utilisées proviennent de la DSN.
- De l'autre, une étude plutôt qualitative, réalisée en collaboration avec Kantar TNS, auprès de 1 000 salariés, en CDI et du secteur privé.

Toutefois, il est important de souligner qu'il existe divers baromètres sur l'absentéisme et qu'ils présentent des résultats qui peuvent diverger les uns des autres.

Cette étude va nous permettre de faire un comparatif de différents indicateurs de l'absentéisme 2020, par rapport aux années antérieures. Elle est réalisée dans le respect des exigences liées au RGPD (Règlement Général de la Protection des Données).

Ces dernières années en France, le taux d'absentéisme se dégrade. En 2020, le taux d'absentéisme annuel moyen s'élève à 6,87% [10], alors qu'il n'était que de 5,54% en 2019. La hausse constatée entre ces deux années est de 24%.

En moyenne, la hausse de ce taux représente l'équivalent de 3,3 jours d'absence par salarié et par an. En effet, en 2019, le nombre de jours d'absence moyen est de 14 jours par an et par salarié. En 2020, ce chiffre dépasse les 17 jours.

De plus, un fait est alarmant. Parmi la population active interrogée, la part des salariés absents au moins un jour dans l'année atteint 41% pour l'année 2020. Cela veut dire qu'environ 4 personnes sur 10 ont été absentes au moins un jour au cours de l'année 2020. C'est 17% de plus que l'année précédente.

En outre, 25% des salariés questionnés indiquent s'être absentés pour un arrêt en lien avec la Covid-19. Dans ce type d'arrêt sont regroupées les personnes atteintes du virus mais également les arrêts dérogatoires pour garde d'enfants, pour personnes vulnérables ou cas contact.

Désormais, concentrons-nous sur l'évolution de l'absentéisme par secteur d'activité. Entre 2016 et 2018 ce taux ne cesse de s'intensifier en passant de 4,59% en 2016 à 5,1% en 2018 [11], tous secteurs d'activité confondus. En termes de taux d'absence, l'année 2019 s'est révélée relativement proche de l'année 2018. C'est pourquoi 2019 est considérée comme une année de référence et sera par la suite comparée à 2020 pour quantifier les impacts de cette crise sanitaire.

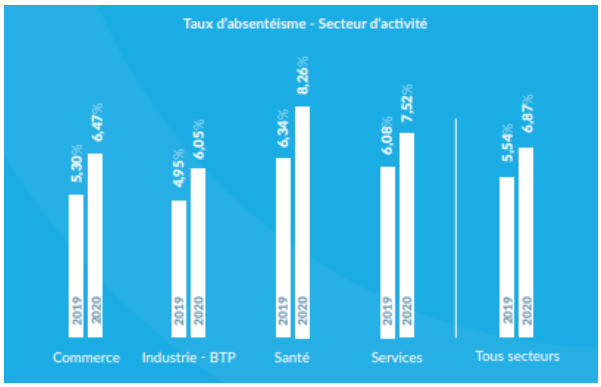


FIGURE 1.6 – Évolution de l'absentéisme par secteurs d'activité

10. Source : l'étude Ayming, le 13ème baromètre de l'absentéisme (étude 2021).
11. Source : l'étude Ayming, le 12ème baromètre de l'absentéisme (étude 2020)

Entre ces deux années et pour tous les secteurs d'activité, le taux d'absentéisme a tendance à augmenter. Les légères différences entre secteurs d'activité s'expliquent par des conditions et des modalités de travail différentes, mais également par des expositions à la Covid-19 différentes. Effectivement, le secteur qui accuse la plus grosse hausse est celui de la santé. Son taux d'absence passe de 6,34% à 8,2%, soit une augmentation de +30% en 2020. Dans ce secteur, les emplois sont plus à risques. En effet, les travailleurs sont directement en contact avec des individus malades et infectés par le virus. De plus, ils ne peuvent pas bénéficier de conditions de télétravail contrairement à d'autres secteurs par exemple.

Il semble également nécessaire de porter un point d'attention sur le taux d'absentéisme en fonction de l'évolution des durées des absences, tous secteurs d'activité confondus.

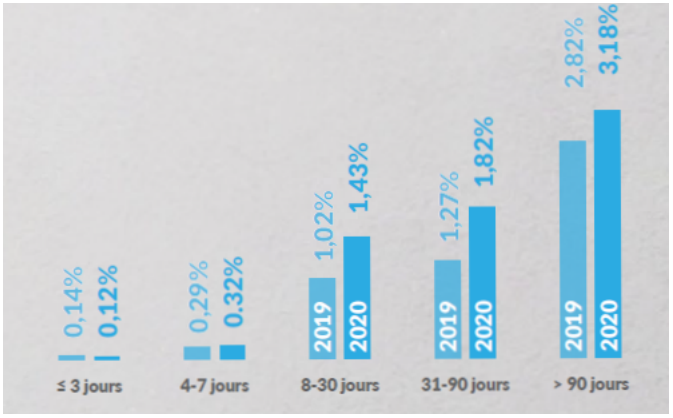


FIGURE 1.7 – Taux d'absentéisme par durée d'absence

Pour toutes les durées d'arrêt supérieures à quatre jours, l'absentéisme en 2020 subit un accroissement significatif. L'augmentation la plus importante correspond aux arrêts compris entre 8 et 90 jours. En effet, cette évolution est de 40% par rapport à l'année 2019. Cette tendance s'explique principalement à cause des arrêts liés à la Covid-19. Cependant, pour le micro-absentéisme, c'est-à-dire les arrêts d'une durée inférieure à 3 jours, l'évolution reste relativement stable.

Cette enquête réalisée sur la population active française, en 2020, attire notre attention sur différents indicateurs :

- La répartition des absences selon les tranches d'âges est en augmentation. Effectivement, le taux d'absentéisme évolue à la hausse, d'environ 24% pour chaque tranche d'âge. Cependant, l'évolution est plus importante pour la population active, agée de 31 à 40 ans. Il semble y avoir un lien de corrélation non négligeable avec le virus de la Covid-19 et la situation personnelle des salariés : garde d'enfant principalement.
- L'évolution de l'absentéisme par année d'ancienneté montre là aussi une certaine tendance à la hausse. En effet, pour toutes les catégories d'ancienneté le taux d'absentéisme a augmenté de plus de 1%, en comparant 2020 avec 2019.
- Dans la continuité des observations des années précédentes, l'écart entre les hommes et les femmes continue de s'accroitre. Effectivement, la hausse du taux d'absentéisme est de 24% chez les femmes et de 22% chez les hommes. Au total, il y a 45% des femmes et 37% des hommes qui sont absents au moins un jour au cours de l'année 2020.

– Il existe une différence significative entre les cadres, qui ont un taux d'absentéisme de 3,18% en 2020 et les non-cadres avec un taux égal à 7,72%. De plus, l'accroissement est plus brutal pour les non-cadres avec une hausse de 26%, comparé à l'année précédente. Dans une moindre mesure, l'élan haussier chez les cadres n'est que de 16%.

De plus, le baromètre annuel sur l'absentéisme de 2021, menée par Malakoff Humanis, détaille les différents motifs d'arrêts prescrits par les médécins. Cette étude croise l'avis des salariés avec ceux de leurs dirigeants. Les arrêts liés à la Covid-19 sont en forte hausse et représentent une part importante de l'ensemble des arrêts. Voici un classement des différents motifs d'arrêt prescrits, en 2020 puis en 2021.

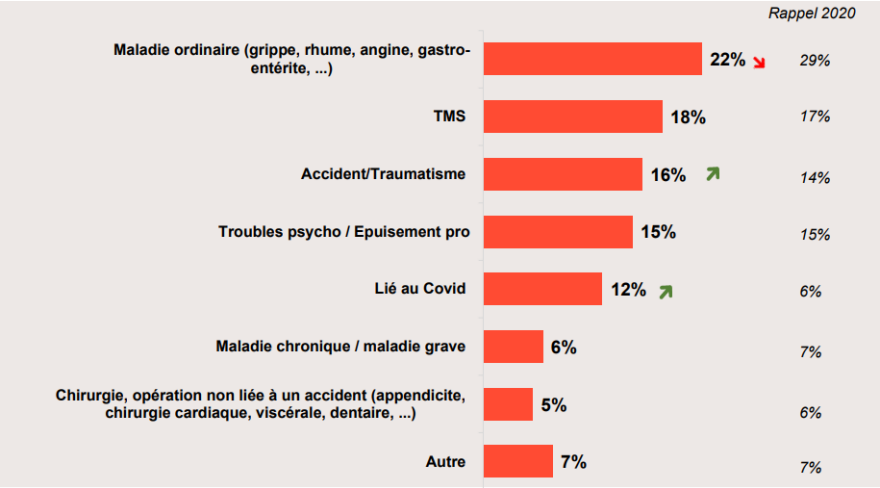


FIGURE 1.8 – Motifs d'arrêts de travail prescrits par le médecin

Dès 2020, les arrêts liés à la Covid-19 représentent environ 6% de tous les arrêts (colonne de droite « Rappel 2020 »). Ce motif regroupe les personnes infectées par la maladie, les cas contacts et les arrêts pour garde d'enfants. Cependant, il faut noter que ce sont les motifs d'arrêts prescrits par les médecins. En effet, au début de la pandémie, de nombreux arrêts en lien avec la Covid-19 étaient déclarés comme de simples arrêts, comptabilisés en maladie ordinaire, ou grave s'il y avait des complications. C'est pourquoi les 6% d'arrêts liés à la Covid en 2020 sont sûrement sous-estimés.

En observant le chiffre de 2021, la proportion d'arrêts liés à la Covid-19 a doublé et vaut désormais 12%. Dans un même temps, cette hausse est principalement compensée par la baisse des arrêts pour maladie ordinaire. En effet, le virus est présent en France depuis presque un an et la connaissance de celui-ci ne fait que de se perfectionner.

1.3.4 L'impact des mesures gouvernementales

Depuis le début de l'année 2020, la crise sanitaire est venue perturber le quotidien de tous les français, mais également de la population mondiale. Chaque pays a été contraint de modifier ses comportements et ses habitudes afin d'appréhender au mieux cette crise. En France, le gouvernement a dû s'adapter en fonction de la propagation de la pandémie.

Afin d'avoir un premier aperçu des impacts de cette crise sur la sinistralité, des études ont été menées au sein de notre direction. Les deux graphiques suivants résument cette analyse. Le premier

concerne uniquement les prestations versées au titre du risque de mensualisation, tandis que le second contient les prestations d'incapacité.

Les prestations sont exprimées en pourcentage, par comparaison avec l'année 2019. Elle est considérée comme une année normale, c'est à dire sans qu'aucun choc majeur ni identifiable ne soit venu perturber la sinistralité. Elle est donc prise pour base de référence. Les prestations versées au cours des années suivantes sont calculées en fonction de celles de 2019.

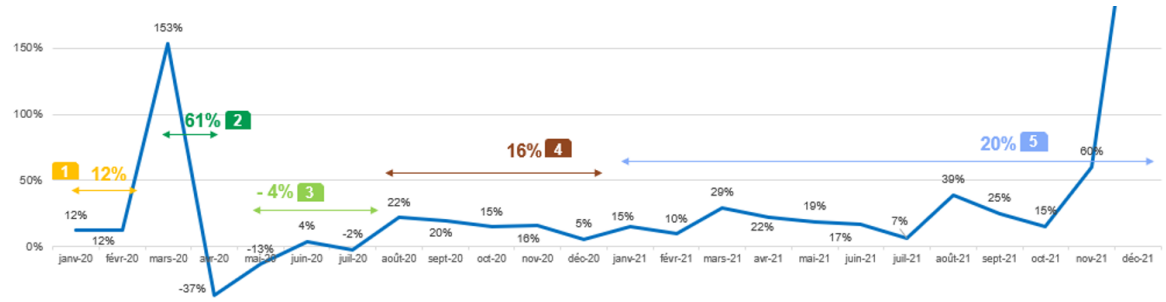


FIGURE 1.9 – Prestations versées pour le risque de mensualisation pour les années 2020 et 2021, en comparaison avec 2019 définie en base 100

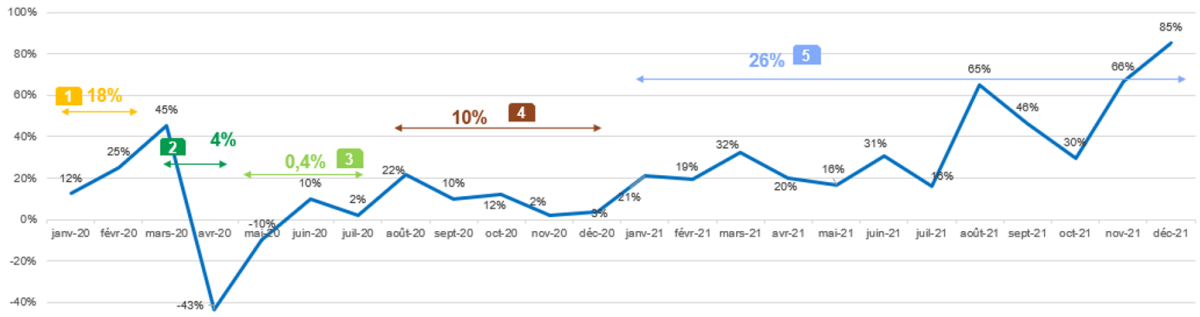


FIGURE 1.10 – Prestations versées pour le risque incapacité pour les années 2020 et 2021, en comparaison avec 2019 définie en base 100

La méthode de comparaison des prestations retenue est dite « en base 100 ». La base représente les prestations mensuelles de l'année 2019, considérée comme la référence pour la comparaison. Par la suite, l'analyse se construit autour de l'évolution des prestations versées en 2020 et en 2021, mois par mois, par rapport à cette base. De plus, toutes les comparaisons ont été effectuées avec un recul équivalent.

Prenons un exemple, afin de faciliter la compréhension : concernant le graphique des prestations de mensualisation (celui du haut), pour le mois de janvier 2020, une hausse de 12% est constatée. Par conséquent, les prestations versées en janvier 2020 sont 12% supérieures par rapport aux prestations de janvier 2019.

A travers ces deux graphiques, la temporalité de la crise sanitaire semble se diviser en cinq périodes distinctes :

1) L'avant crise :

Pour rappel, la crise sanitaire a impacté la France à partir de fin février, début mars de l'année 2020. En revanche, dès le mois de janvier, une hausse des prestations versées est perceptible.

Chaque année depuis 2016, l'absentéisme augmente de plusieurs dixièmes de pourcents. Cependant, pour l'année 2020, l'augmentation est bien plus importante. Pour les deux mois précédents la crise, la hausse des prestations pour le risque de mensualisation se stabilise autour de 12%, alors que celle de l'incapacité se situe en moyenne à 18%. Une étude plus approfondie aura pour but d'expliquer les raisons de ce phénomène.

2) Le 1er confinement (du 17/03/2020 au 11/05/2020) :

La mise en place de ce confinement général a pour but de réduire au maximum la circulation du virus. De nombreuses mesures très strictes ont été mises en place afin de diminuer les mouvements de la population. En effet, tous les commerces jugés non essentiels sont fermés. De plus, les écoles, collèges, lycées et universités ferment également. La population est forcée de rester au maximum à son domicile et seul les trajets nécessaires sont autorisés.

Ce confinement très rude a donc eu des conséquences sans précédent sur l'absentéisme. En effet, de nombreux salariés ont vu leur activité s'arrêter du jour au lendemain. D'autres, ont dû poser des arrêts dérogatoires, afin de s'organiser pour la garde de leurs enfants ou l'accompagnement de personnes vulnérables. D'autres encore, ont dû s'adapter a de nouvelles conditions de travail.

Les prestations payées au titre du mois de mars 2020 sont sans précédent. En effet, des augmentations brutales sont constatées. Pour le risque de mensualisation, une hausse de 153% est observée, par rapport à l'année précédente. Cette hausse est moins extrême pour le risque d'incapacité, mais avoisine quand même les 45%.

De plus, les mois d'avril et de mai 2020 présentent des évolutions relativement faibles (écarts négatifs). Effectivement, la totalité de la population française est confinée depuis mi-mars, ce qui entraine une forte réduction des risques sociaux, des accidents du quotidien et des transmissions de microbes. Par conséquent et durant la deuxième moitié du confinement, le nombre de nouveaux sinistres déclarés diminue.

3) Période estivale : l'entre deux vagues/confinements (de fin mai à mi-août) :

Durant cette période et pour les deux risques, les écarts de prestations restent stables entre 2020 et 2019. Effectivement, durant la période suivant le déconfinement, les contaminations sont au plus bas.

4) Fin d'année 2020 : (de fin août à fin décembre)

Dans un premier temps, les contaminations et les prestations versées sont de nouveau en forte hausse à partir de mi-août. Les hypothèses gouvernementales sur le ralentissement du virus en période estivale n'étaient pas spécialement fausses. Cependant, comme toutes les frontières étaient fermées, les français ont du se déplacer uniquement sur le teritoire. De nombreux endroits très touristiques ont vu leur population s'accroitre drastiquement. Cela a donc favorisé la transmission du virus.

C'est pourquoi à la fin de la période des vacances, le virus est de nouveau très menaçant sur le territoire français. En effet, les prestations versées sont en augmentation par rapport à 2019. En moyenne, une hausse d'environ 17% est constatée pour les deux risques cumulés.

Dans un second temps, à partir de mi-octobre, les mesures gouvernementales se sont durcies. D'abord un couvre-feu est mis en place, puis un second confinement. Néanmoins, ce confinement apparait comme plus allégé par rapport au premier. Les déplacements sont limités, les commerces non essentiels sont fermés, les réunions privées et les rassemblements publics sont également interdits. De plus, le télétravail est fortement recommandé et les frontières extérieures à l'Union Européenne sont fermées. En revanche, les crèches, écoles, collèges, lycées et universités restent ouverts. Des protocoles sanitaires stricts sont mis en place.

Finalement, l'évolution des prestations n'est que d'environ 10% pour la mensualisation et de seulement 3% pour l'incapacité.

5) Sinistralité 2021 :

Il est important de souligner qu'au moment de la rédaction de cette partie, les derniers mois de l'année sont très instables. Le recul nécessaire sur les données n'est pas suffisant.

Sans tenir compte des trois derniers mois de l'année 2021, une hausse des prestations est observable, d'environ 20% pour la mensualisation et de 26% pour l'incapacité. De plus, durant le mois d'août 2021, il y a eu un événement marquant, représenté par un pic de prestations versées. La période de temps semble correspondre avec la 4ème vague de covid.

1.4 Suivi du portefeuille

Une analyse de suivi de portefeuille a été menée sur différents indicateurs de sinistralité. Ces indicateurs sont nécessaires pour le calcul des provisions. Finalement, cette étude préliminaire dresse un premier constat des conséquences de la crise sanitaire sur la dérive du portefeuille.

Initialement, l'objectif de ce mémoire était de coupler la construction d'une loi d'incidence avec celle d'une loi de maintien en arrêt. Néanmoins, le suivi de ces indicateurs a permis de montrer que la construction de modèles d'incidence serait suffisante pour nos besoins. En effet, la dérive du portfeuille se concentre principalement autour de l'augmentation du nombre de sinistres.

Le taux d'incidence est défini par le nombre d'arrêts de travail sur l'ensemble des salariés du portefeuille, par période d'observation. Grâce à la DSN, les organismes assureurs ont connaissance des arrêts seulement quelques temps après leur survenance.

Pour avoir des résultats plus robustes, l'année de sinistralité 2021 n'est pas retenue, et seule l'année 2020 est prise en compte dans les modèles de construction. En effet, la connaissance de cette année est trop peu maîtrisée. L'ensemble de la population mais aussi tous les sinistres survenus en 2021 ne sont encore pas connus.

Par la suite, une analyse sur les cinq indicateurs de sinistralités suivants est réalisée :

a) Les prestations versées, exprimées en millions d'euros

b) Le nombre d'arrêts observé

c) La durée moyenne des arrêts, exprimée en jours

d) Les indemnités journalières moyennes (IJ), exprimées en euros

e) L'âge moyen à la survenance de l'arrêt

Les travaux présentés par la suite datent du mois de septembre 2021. Durant cette période, l'analyse des indicateurs n'est pas faite avec l'apport de la DSN mais à l'aide des données internes du groupe AG2R La Mondiale. Par conséquent, cette différence de source d'information va entraîner un certain décalage avec les modèles réalisés, dans les parties suivantes.

Pour suivre l'évolution de l'absentéisme au cours du temps, cette étude est réalisée tous les mois au sein de notre direction. Elle se présente sous forme de différents triangles de développement. Chaque ligne correspond à une année de survenance. Les années comprises entre 2015 et 2021 sont représentées. Chaque colonne correspond à l'année de développement, en fonction du mois d'observation.

Par exemple, dans le triangle ci-dessous, le premier montant de prestations versées, en haut à gauche, d'une valeur égale à 9,32 millions d'euros représente : la somme des prestations payées par l'organisme assureur, pour les sinistres survenus en 2015 et observés au 30 septembre de l'année N, soit en 2015. Dans la cellule de droite, les prestations versées au 30 septembre 2016, pour les sinistres survenus au cours de l'année 2015, sont égales à 455,62 millions d'euros.

L'évolution des prestations versées est détaillée ci-dessous.

Prestations	fin Sept N	fin Sept N+1	fin Sept N+2	fin Sept N+3	fin Sept N+4	fin Sept N+5	fin Sept N+6
2015	9,32M€	455,62M€	766,17M€	875,83M€	908,15M€	916,48M€	918,82M€
2016	10,74M€	456,92M€	747,30M€	867,71M€	903,25M€	914,37M€	
2017	11,69M€	439,46M€	738,40M€	859,54M€	892,74M€		
2018	10,20M€	361,81M€	708,55M€	817,77M€			
2019	10,44M€	459,99M€	747,44M€				
2020	12,31M€	567,83M€					
2021	20,57M€						

FIGURE 1.11 – Triangle de développement des prestations versées

Fin septembre 2015, seulement 9,32 millions d'euros de prestations ont été versés au titre des sinistres survenus au cours de cette même année. Six ans plus tard, le montant versé se stabilise autour de 918 millions d'euros.

Pour les années de survenance suivantes, la proportion des prestations connues la première année semble similaire à celle de 2015. En effet, avec une vision à fin septembre de l'année en cours, environ 1% du montant total des prestations payées est connu.

Les prestations versées en 2020 sont clairement plus importantes que les années antérieures. Dans la deuxième colonne, pour une vision à septembre N+1, les prestations versées, au titre des survenances 2020, atteignent 567,83 millions d'euros. Cela représente 206 millions d'euros de plus par rapport aux survenances 2018, soit plus de 57% d'augmentation, et 108 millions d'euros supplémentaires (soit +23,6%) par rapport à 2019.

Enfin, les prestations versées pour les sinistres survenus au cours l'année 2021 représentent plus du double des prestations associées aux survenances 2019 et quasiment 70% de plus par rapport

à 2020. Néanmoins, cette hausse extrême est à nuancer puisque le processus de gestion a fortement évolué depuis 2020. En effet, la crise sanitaire a permis d'accélérer les processus de gestion, notamment grâce à l'apport de la DSN. Désormais, les sinistres sont connus nettement plus tôt et par conséquent, les prestations correspondantes sont payées en avance, par rapport aux années précédentes

Pour conclure, la crise sanitaire semble avoir fortement impacté le risque incapacité, pour les arrêts de travail survenus au cours des années 2020 et 2021. Le mode de gestion et de règlement de ces prestations a lui aussi été amélioré.

La progession du nombre de sinistres est représentée ci-dessous.

Nombre Sinistres	fin Sept N	fin Sept N+1	fin Sept N+2	fin Sept N+3	fin Sept N+4	fin Sept N+5	fin Sept N+6
2015	38 758	648 416	746 928	758 288	760 668	761 473	761 670
2016	44 849	629 176	711 092	721 368	724 520	725 259	
2017	52 252	602 713	686 402	699 535	702 096		
2018	42 977	459 374	642 374	653 242			
2019	36 821	570 564	632 000				
2020	32 454	702 461					
2021	62 742						

FIGURE 1.12 – Triangle de développement du nombre de sinistres observés

De manière similaire aux prestations versées, une remarque est flagrante. Durant la première année d'observation, la connaissance du nombre de sinistres ne constitue que infime partie de la sinistralité totale. Effectivement, pour l'année de survenance 2018, le nombre de sinistres connus au 30 septembre 2018 n'est que de 42 997, alors qu'il se stabilise autour de 653 242 au 30 septembre 2021, soit uniquement 6,6% la première année.

Dans la suite, pour avoir une analyse plus représentative de la réalité, les comparaisons seront effectuées par rapport à la deuxième colonne, avec une vision au 30 septembre de l'année N+1.

En ce qui concerne les années touchées par la pandémie, le nombre d'arrêts survenus en 2020 a explosé. En effet, au cours des années avant Covid, principalement 2018 et 2019, le nombre moyen de sinistres se situe aux alentours de 515 000. Pour l'année 2020, plus de 700 000 arrêts sont comptabilisés, soit une augmentation d'environ 36%.

Finalement, l'augmentation du nombre de sinistres semble se produire dans des proportions plus importantes que celle des prestations versées.

En effet, en comparant la sinistralité de 2020 avec celle de 2018 et 2019, une hausse du nombre de sinistres de 36% est constatée, contre une hausse de 24% pour les prestations versées. Par conséquent, la construction de ces indicateurs prouve que la hausse du nombre de sinistres s'avère être la principale conséquence de cette crise sanitaire.

Par la suite, la durée moyenne des arrêts est analysée.

Durée	fin Sept N	fin Sept N+1	fin Sept N+2	fin Sept N+3	fin Sept N+4	fin Sept N+5	fin Sept N+6
2015	14,8	48,4	71,9	80,9	83,5	84,1	84,3
2016	15,7	50,5	74,7	85,6	88,6	89,5	
2017	14,6	50,2	75,7	86,4	89,4		
2018	15,2	53,5	77,0	87,7			
2019	17,8	53,0	79,9				
2020	21,3	51,2					
2021	19,2						

FIGURE 1.13 – Triangle de développement de la durée moyenne d'arrêt (en jours)

Contrairement aux deux indicateurs précédents, l'année 2020 n'est pas marquée par une variation significative de la durée moyenne. Si là encore, la comparaison s'effectue au 30 septembre de l'année N+1, la durée moyenne des arrêts survenus en 2020 est inférieure par rapport aux années 2018 et 2019. Les arrêts durent en moyenne deux jours de moins, soit une diminution d'environ 4%.

De plus, une analyse des indemnités journalières (IJ), versées en euros, et de l'âge moyen des assurés à la survenance est réalisée. En revanche, les conclusions sont similaires à la durée moyenne d'arrêt. Aucun choc caractéristique et propre à l'année de la Covid-19 ne semble notable [12].

L'âge moyen à la survenance semble constant, de 2015 jusqu'à aujourd'hui, puisqu'il se situe entre 39 et 40 ans.

En ce qui concerne les indemnités journalières, une légère évolution se profile depuis 2015. Effectivement, selon les chiffres de l'INSEE, les salaires ont augmenté en moyenne de 1,4% par an depuis 2015, en euros constants.

En résumé, cette étude a permis de se rendre compte que l'indicateur le plus touché par cette pandémie est le nombre d'arrêts. De plus, grâce à cette analyse, les objectifs de ce mémoire ont été révisés. En effet, au vu des faibles impacts sur la durée moyenne des arrêts, le choix final s'est concentré sur la modélisation d'une loi d'incidence. Par conséquent, une analyse approfondie de la fréquence d'arrêt au cours de l'année, basée autour de trois modèles de prédictions, sera mise en place.

12. Se référer aux triangles de développement en annexe : 4.3.5

2 Présentation des données

Après avoir mis en contexte les notions essentielles, défini la prévoyance, la DSN, l'évolution de l'absentéisme et enfin les objectifs de ce mémoire, nous allons explorer les données utilisées pour la construction de nos modélisations.

Dans un premier temps, le périmètre des données retenu pour la construction des modèles futurs est détaillé. Il se caractérise par des variables spécifiques comme les années de survenance, le segment du contrat ou encore les motifs d'arrêts.

Dans un second temps, pour aboutir à notre base finale, les différentes étapes de construction des données sont énumérées. Un zoom est réalisé sur les principales variables retenues.

Ensuite, les retraitements et les règles de gestion qui ont le plus impacté la conception de cette base sont listés et présentés.

Enfin, des statistiques descriptives viendront affiner la compréhension du portefeuille étudié, d'abord dans sa globalité, puis variable par variable. Le but est de mettre en évidence certains phénomènes pouvant impacter les variables cibles de notre étude.

2.1 Présentation du périmètre

Rappelons que l'objectif de l'étude est de mettre en évidence les impacts causés par la pandémie de la Covid-19. Pour ce faire, le nombre d'arrêts par année d'observation est modélisé. Dans la suite, le périmètre conservé est façonné autour d'un objectif commun : l'analyse des résultats concernant le taux d'incidence du portefeuille. Ce périmètre se base donc sur les trois délimitations suivantes.

2.1.1 Années de survenance

Tout d'abord, il est indispensable de définir le cadre temporel de l'étude. Pour mener à bien ce projet, séparer les différentes bases de données en fonction de l'année de survenance des sinistres est important. Par la suite, deux modèles prédictifs seront entraînés. Un premier qui repose sur la base avec les années de survenance antérieures à la crise sanitaire. Un second modèle, établit à l'aide des données de l'année de survenance 2020.

2.1.1.1 Base de données antérieure à la Covid

Premièrement, les données brutes récupérées proviennent de la DSN. Ce processus a été instauré de manière progressive. C'est seulement à partir de 2017 que la DSN devient obligatoire. Cependant, comme c'est un processus nouveau pour les entreprises, elles ont du faire face à une période d'adaptation. Durant les premiers mois de cette déclaration, les données récoltées ne semblent pas assez robustes pour être utilisées dans cette étude. C'est pour cette raison qu'avant 2018, les données issues de ce processus sont considérées comme imprécises voire incomplètes.

De plus, le virus de la Covid-19 s'est développé en France lors du premier trimestre de 2020. Si bien que, l'année 2019 est supposée indépendante de cette crise et donc n'en subir aucune conséquence.

Par conséquent, deux années de survenance, 2018 ou 2019, sont finalement gardées. Une analyse de ces données a permis de conclure qu'elles étaient globalement similaires. En outre, 2018 et 2019 sont des années de sinistralités sans perturbation ni choc majeur. Le fait d'avoir ces deux années permet de bénéficier d'une base plus conséquente donc plus robuste lors des modélisations futures.

2.1.1.2 Base de données en situation de pandémie

Dans un second temps et pour répondre à l'objectif de ce mémoire, une base avec les données observées durant la période impactée par la Covid-19 est construite. Ce virus est apparu en France au cours du premier trimestre 2020 et s'est très vite répandu sur tout le territoire.

Pour des raisons techniques et pratiques expliquées précédemment, la base de données en situation de pandémie n'est composée que de l'année de survenance 2020. En effet, lors de création de ces bases, plusieurs complications sont apparues :

- Manque d'informations supplémentaires :
  Lorsque la base de données a été créée, seule la DSN récoltée durant les premiers mois de l'année 2021 été connue. La connaissance de cette année était donc trop peu maîtrisée. L'ensemble de la population sous risque mais également tous les sinistres survenus en 2021 n'étaient donc pas connus. Le fait de rajouter l'année 2021 à nos données aurait donc biaisé les résultats des modélisations.

– Survenance des arrêts 2020 sans égal :

  L'année 2020 est caractérisée par une sinistralité extrême. Ce phénomène à touché le monde entier y compris la France et le portefeuille clients d'AG2R. L'étude des indicateurs de sinistralité vient confirmer le caractère atypique de l'année 2020. En effet, les chocs survenus durant cette année sont sans précédent. Ils sont majoritairement causés par les différents confinements et la mise en place d'arrêts dérogatoires. Aujourd'hui, la sinistralité 2021 est toujours intense. Néanmoins, cette forte sinistralité subsiste dans des proportions bien moindres par rapport à 2020.

2.1.2 Motifs d'arrêts de travail

Une fois le cadre temporel défini, le périmètre des données va être restreint en fonction du risque considéré. En effet, dans cette étude, seuls les arrêts de travail entrainant une incapacité sont étudiés. De plus, certains motifs d'arrêts, qui possèdent des traitements et des mécanismes de remboursements particuliers sont exclus.

Dans un premier temps, tous les motifs d'entrée en arrêt de travail sont catégorisés dans le cahier des charges de la Déclaration Sociale Nominative. Ces différents motifs permettent une segmentation précise et sont classés de la manière suivante :

01 - maladie

02 - maternité

03 - paternité / accueil de l'enfant

04 - congé suite à un accident de trajet

05 - congé suite à une maladie professionnelle

06 - congé suite à un accident de travail ou de service

07 - femme enceinte dispensée de travail

08 - temps partiel thérapeutique

09 - adoption

10 - [FP] - congé suite à une maladie imputable au service

11 - [FP] - congé de maladie des victimes ou réformés de guerre (art. 41)

12 - [FP]- congé de longue durée

13 - [FP] - congé de longue maladie

99 - annulation

Tous les motifs relatifs à la fonction publique sont désignés par la nomenclature [FP]. Ils sont numérotés de 10 à 13 dans cette liste et font l'objet d'un traitement spécifique. Ils vont donc être exclus pour la suite des travaux. En effet, leur mécanisme de remboursement est différent du régime général.

De plus, les arrêts liés à la parentalité, qui regroupent par exemple la maternité, la paternité, l'accueil d'enfants ou encore les congés pour adoption, ne sont pas pris en compte dans la base de données finale. Ces congés sont respectivement numérotés, dans le cahier des charges de la DSN, 02, 03 et 09. En effet, ils sont exclus car eux aussi bénéficient d'un traitement particulier.

C'est la sécurité sociale qui prend en charge le versement des prestations en cas de sinistre, sous réserve de remplir certaines conditions. Dans le cas d'un congé maternité, la salariée doit être assurée sociale depuis au moins 10 mois. Elle doit aussi avoir travaillé au moins 150h au cours des trois derniers mois ou 600h au cours des 12 derniers mois précédent la grossesse. Enfin, elle doit également avoir perçu en cumulé 1 015 fois ou 2 030 fois le smic horaire, respectivement au cours des 6 ou 12 mois précédent la grossesse.

En outre, les temps partiels thérapeutiques, numérotés 08 dans la DSN, sont également retirés de cette étude. Les mi-temps thérapeutiques sont des arrêts assez spéciaux. Ils font généralement suite à un premier arrêt de travail. C'est lors de la reprise de poste que le temps partiel peut être proposé, en fonction de l'état et des besoins du salarié.

Enfin, toutes les périodes d'arrêt ayant comme statut le numéro 99, relatif au motif « annulé » sont retirées de la base de données.

Finalement, au cours de cette étude, les motifs d'arrêts conservés sont ceux pour lesquels un contrat de prévoyance complémentaire peut être souscrit. La liste finale comprend donc les congés pour maladie, les congés suite à un accident de trajet, une maladie professionnelle, un accident de travail ou de service et enfin les congés pour femme enceinte dispensée de travail.

Dans un second temps, s'il existe des motifs d'entrée en incapacité, il existe également différents motifs de sortie de cet état. En effet, l'arrêt de travail peut se clôturer des suites de la guérison du salarié, donc par la reprise totale du travail. L'individu peut également sortir de l'incapacité par son passage en invalidité : à la suite de la période réglementaire maximale de l'incapacité soit 36 mois ou à la suite de l'aggravation de son état, nécessitant le passage en invalidité. Somme toute, le décès est aussi une raison de sortir de l'incapacité.

2.1.3 Segmentation des contrats de prévoyance

Afin de représenter au mieux les caractéristiques de la population sous risque, il convient de segmenter les produits de prévoyance en fonction de leurs marchés. Trois segments distincts sont comptabilisés de la manière suivante :

- Le segment Standard :

  Chaque compagnie d'assurance possède ses propres gammes de contrats, prédéfinis à l'avance et comportant des garanties obligatoires. L'offre est donc construite en fonction des attentes supposées du secteur cible et la tarification s'appuie sur ses principales caractéristiques. C'est ce qu'on appelle le segment standard. En effet, ces contrats intègrent, de manière très macroéconomique, la démographie et les caractéristiques de l'entreprise assurée. Au sein du groupe AG2R La Mondiale, ces contrats sont à destination des entreprises comptant moins de 100 salariés mais aussi des Travailleurs Non-Salariés (TNS).

- Le segment Sur-Mesure :

  Pour les plus grosses entreprises (de 100 têtes et plus), les sociétés d'assurance peuvent proposer des contrats dits sur-mesure. En fonction des besoins spécifiques d'une entreprise, ces contrats permettent d'adapter les garanties couvertes ou non (possibilité d'ajout ou de suppression de clauses). En effet, la tarification dépend du barème de garanties désiré par l'entreprise et de ses propres caractéristiques (secteur d'activité, âge moyen, niveau de vie, caractéristiques démographiques ou toutes autres spécificités de l'entreprise).

En général, les produits sur-mesure procurent un double avantage. D'abord pour les salariés, qui bénéficient d'une couverture optimale avec des garanties adaptées à leurs besoins. Ensuite, pour les entreprises, qui s'assurent de faire des économies en ayant une tarification en adéquation avec les prestations utilisées.

– Le segment Convention Collective Nationale (CCN) :

  Enfin, la grande majorité des entreprises sont rattachées à une convention collective qui régit l'ensemble des droits des salariés d'un même secteur et qui complète le droit du travail. En général, ces conventions sont négociées entre les syndicats représentants les salariés et les organisations patronales. Elles imposent à l'employeur la mise en place d'une solution de prévoyance pour tout ou partie de ses salariés, mais aussi de respecter des règles particulières sur son choix de prévoyance collective.

  Les produits de ce segment sont donc un mélange des caractéristiques des produits standards et des produits sur-mesure. En général, le produit CCN fournit des garanties minimales. La CCN a ensuite la possibilité de rendre obligatoire la souscription de ce contrat négocié ou de simplement le conseiller à ses membres.

  Les entreprises d'un même secteur d'activité se regroupent en CCN afin de multiplier le nombre d'individus. Elles vont donc avoir une masse salariale suffisamment importante, et ainsi bénéficier d'un pouvoir de négociation plus pertinent.

2.2 Construction de la base de données

Afin de construire une base de données aussi propre que possible, l'intégralité des branches d'activités disponibles dans notre portefeuille est conservée, mais aussi, l'ensemble des trois segments de contrats présentés ci-dessus (Standard, Sur-Mesure, CCN).

Pour rappel, les données utilisées sont issues de la DSN. De nombreux avantages sur la connaissance de la sinistralité du portefeuille en découlent. Aujourd'hui, grâce à la DSN, tous les arrêts sont connus, y compris ceux qui ont une durée inférieure à celle de la franchise. Auparavant, avec les données disponibles dans nos bases internes, seuls les arrêts d'une durée supérieure à la franchise apparaissaient. En effet, ce sont les arrêts pour lesquels la complémentaire indemnise ses assurés. Cependant, la différence entre ces deux processus de récolte d'information n'est pas négligeable. En effet, les arrêts courts représentent une partie importante du nombre total d'arrêts, mais ils jouent aussi un rôle significatif sur l'incidence moyenne du portefeuille.

Ensuite, la DSN a l'avantage d'être régulière : réception d'une mise à jour des informations de manière mensuelle. Cela nous permet d'avoir des données plus fiables sur les salariés.

Néanmoins, c'est la première fois que des données provenant de la DSN vont servir à la construction de modèles au sein de notre équipe. C'est pourquoi il est donc nécessaire de modifier et retraiter les données brutes fournies par la DSN pour ensuite, les adapter à nos besoins spécifiques.

2.2.1 Cheminement et construction des différentes tables

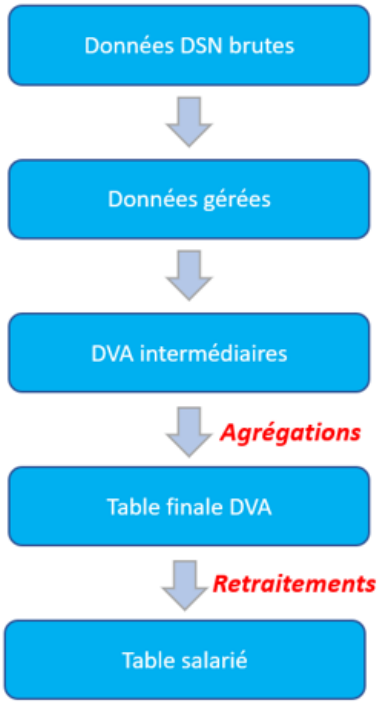


FIGURE 2.1 – Schéma des étapes de construction des bases de données

Tout d'abord, le point de départ se compose de la base de données brutes fournies par la DSN. Cette base contient des informations sur le salarié, sur ses fiches de paie, etc.

Dans cette base, un certain nombre d'informations ne va pas servir pour notre besoin. Effectivement, toutes les colonnes qui ne sont pas liées aux caractéristiques du salarié, à son contrat de travail ou à ses potentiels arrêts ne vont pas être nécessaires pour mener à bien notre étude.

Ensuite, ces données viennent alimenter de manière hebdomadaire les données gérées. Ce sont les données brutes de la DSN légèrement modifiées, puisque la nomenclature des champs est corrigée afin de rendre les tables exploitables.

Les données gérées représentent un ensemble de différentes tables. Ces tables sont regroupées par thème et contiennent chacune des informations précises telles que : les données sur le contrat de travail, sur les salariés, sur leurs arrêts, sur leurs entreprises affiliées, sur leurs adhésions à la prévoyance ou encore des données géographiques.

Par la suite, ce sont ces tables disjointes qui vont donner naissance aux tables appelées DVA. Ces tables contiennent respectivement des informations sur les salariés, sur les rémunérations ou sur les arrêts. Ce sont des tables intermédiaires à la construction de la base finale. À l'aide de règles de gestion, ces différents DVA intermédiaires sont regroupés pour ne former qu'une seule table finale. Elle se nomme la table finale DVA.

Dans l'optique de respecter les exigences liées au RGPD, aucune information permettant d'identifier un assuré n'est disponible dans ces tables. C'est pourquoi un salarié est caractérisé par un identifiant. L'identifiant est défini par une maille croisant le numéro de sécurité sociale, l'année de déclaration de la DSN, le numéro de SIREN, le numéro de contrat AG2R La Mondiale et les éventuels

arrêts. Cela veut donc dire qu'un individu qui possède deux contrats chez AG2R La Mondiale sera caractérisé comme deux salariés différents et aura au moins deux lignes dans cette base. De manière similaire, un individu qui est à temps partiel dans deux entreprises distinctes sera comptabilisé par deux identifiants différents.

Ensuite, la table DVA finale se caractérise de la manière suivante. Chaque salarié possède une première ligne de déclaration classique. Cette ligne se compose de toutes les informations non événementielles qui caractérisent le salarié et son contrat, toujours dans le respect du RGPD[1]. Ces informations proviennent de la dernière déclaration sociale nominative reçue pour cet individu.
Les lignes suivantes, associées au même identifiant, représentent les déclarations dites événementielles. En fonction du nombre d'événements déclarés, donc en fonction du nombre d'arrêts que cet identifiant a pu avoir durant la période considérée, zéro, une ou plusieurs lignes supplémentaires sont présentes dans cette base.

Enfin, de nombreux retraitements, qui seront énumérés dans la suite, viennent donner naissance à la table finale, nommée « table salarié ». Cette table est désormais exploitable et va servir pour toutes les modélisations futures. La base finale recense donc tous les individus présents dans les entreprises assurées par AG2R La Mondiale au cours des années d'observation 2018, 2019 et 2020, ayant eu ou non des arrêts de travail.

Le format de cette table se lit en croisant les informations relatvies aux assurés avec celles des années d'observation. Si bien que, pour chaque assuré de notre population, il y a une première ligne de déclaration, définissant les caractéristiques de cet individu. Ensuite, cette ligne est suivie d'une ligne par année d'observation, qui comptabilise et regroupe le nombre d'arrêts total, subit durant chacune des trois années. De plus, une variable précisant la durée totale de tous ces arrêts, en nombre de jours, vient compléter cette ligne.

Cette base va nous servir lors de la construction de nos modèles d'incidence. La variable cible, ou à expliquer, est le nombre d'arrêts que l'assuré a subi. En effet, le but est de modéliser l'incidence donc la probabilité de tomber en arrêt au cours d'une année. C'est pourquoi pour chaque profil de salarié et année de survenance, il est nécessaire d'estimer son nombre d'arrêts, pondéré par son exposition.

Il est impératif de pondérer les arrêts observés par l'exposition que cet individu a eu au cours de l'année d'observation. En effet, un individu qui est présent durant seulement 6 mois de l'année dans l'entreprise, aura une exposition de 50%. Si, au cours de cette période, il a eu deux arrêts de travail, cela revient à dire que ce même individu aurait en moyenne quatre arrêts, s'il était resté une année complète dans cette entreprise.

2.2.2 Zoom sur les variables utilisées

Désormais, nous allons énumérer toutes les variables présentes dans la « base salarié ». Pour ce mémoire, 14 variables ont été retenues.

– Année d'observation : cette variable indique, pour chaque identifiant, l'année d'observation ou l'année d'exposition considérée. En effet, il se peut qu'un individu soit présent toute l'année 2018 mais seulement la moitié de l'année 2019. Cette variable sert alors pour différencier les

1. Le Règlement Général sur la Protection des Données (RGPD) encadre le traitement des données personnelles sur le territoire de l'Union européenne.

arrêts et l'exposition d'un salarié durant chaque année d'observation. De plus, dans la base finale, il y a une ligne par année d'observation et par assuré, qui définit le nombre et la durée des arrêts. Elle peut donc prendre trois valeurs numériques distinctes : 2018, 2019 ou 2020.

- Nombre d'arrêts : c'est la variable cible. Toutes les modélisations auront pour but de prédire, le plus justement possible, ce nombre d'arrêts. Cette variable comptabilise pour chaque identifiant le nombre d'arrêts qu'il a eu durant chaque année d'observation entre 2018 et 2020. C'est une variable numérique, continue, qui prend des valeurs comprises entre 0 et 22.

- Age du salarié : cette variable indique l'âge du salarié au moment de la déclaration sociale nominative. C'est également une variable numérique et continue. Par la suite, un retraitement est fait sur cette variable afin de ne garder que les individus dont l'âge est compris entre 16 et 67 ans inclus. Ils représentent la population active et sont les seuls à pouvoir bénéficier d'un contrat de prévoyance.

- Catégories Socio Professionnelles (CSP) : c'est une variable catégorielle qui classe les individus selon leur fonction. Un regroupement autour de quatres postes ou catégories clefs a été effectué de la manière suivante :

  - Les cadres
  - Les employés
  - Les ouvriers
  - Les professions intermédiaires

- Identifiant du salarié : il s'agit d'un numéro définissant le salarié afin de respecter son anonymat. Il est attribué à chaque assuré en fonction de la maille choisie (n° SS, année DSN, n° SIREN, n° contrat, arrêt).

- Genre : cette variable est catégorielle et donne l'information sur le sexe de l'individu considéré. Elle prend deux modalités qui sont Femme et Homme.

- Libellé du contrat : c'est également une variable catégorielle à deux modalités : CDD ou CDI en fonction de la nature du contrat de travail.

- Ratio d'exposition : comme indiqué précédemment, cette variable est importante afin de comptabiliser le poids de chaque individu. En effet, cette variable numérique définie un ratio, donc un chiffre compris entre 0 exclu et 1 inclus, en fonction de la présence de l'individu au sein de l'entreprise, au cours de chaque année d'observation.

  Par exemple, un individu qui est présent toute l'année bénéficie d'un ratio d'exposition de 1. En revanche, un individu pour lequel son contrat de travail commence le 1er avril a, quant à lui, un ratio d'exposition de 0,75 car il n'est présent que durant 9 mois de l'année.

- SIREN (« Système d'Identification de Répertoire des ENtreprises ») de l'entreprise : cette variable sert à identifier l'entreprise dans laquelle travaille le salarié, selon un identifiant déterminé de manière unique et nationale. Ce numéro est composé d'une suite de 9 chiffres, il permet d'avoir des informations juridiques et financières sur l'entreprise. De plus, il est attribué à vie à une entreprise.

– Statut du salarié : cette variable est, elle aussi, catégorielle avec deux modalités, qui sont Cadre et Non Cadre, suivant le statut de l'assuré, défini dans son contrat de travail.

– Tranche d'ancienneté : cette variable résulte de la transformation d'une variable numérique et continue indiquant l'ancienneté du salarié au sein de son entreprise, calculée en années. Six tranches distinctes d'ancienneté sont définies comme suit :
  - Moins d'un an
  - Entre un et trois ans
  - Entre trois et cinq ans
  - Entre cinq et dix ans
  - Entre dix et vingt ans
  - Plus de vingt ans

– Tranche établissement : cette variable provient également d'une transformation de variable numérique et continue, indiquant le nombre de salariés par établissement. De la manière suivante, cinq modalités en découlent :
  - Entre 1 et 10 salariés
  - Entre 11 et 50 salariés
  - Entre 51 et 150 salariés
  - Entre 151 et 250 salariés
  - Plus de 250 salariés

– Tranche salaire annualisé : en se basant sur la valeur du SMIC annuel de janvier 2022, soit 15 228€ net, et de la pyramide des salaires en France, la variable continue définissant le salaire annuel net est remaniée pour former six tranches distinctes de salaire annualisé, comme suit :
  - Inférieur à 15 000€
  - Entre 15 000€ et 24 000€
  - Entre 24 000€ et 36 000€
  - Entre 36 000€ et 48 000€
  - Entre 48 000€ et 72 000€
  - Supérieur à 72 000€

– Type de gestion : c'est une variable catégorielle à trois modalités. Les modalités sont choisies afin de définir qui s'occupe de la gestion des contrats :
  - Gestion directe
  - Gestion déléguée
  - Uniquement gestionnaire du contrat

2.3 Retraitements et règles de gestion

Les données issues de la DSN sont brutes, c'est-à-dire sans aucun retraitement préalable. Pour l'étude du taux d'incidence, la base de données doit respecter certaines contraintes métiers et valider des règles de gestion. Dans la suite, une liste non exhaustive de retraitements est détaillée.

2.3.1 Règles de gestion concernant la sélection du périmètre

Dans cette section, une succession de retraitements est réalisée afin de se focaliser sur le périmètre désiré.

2.3.1.1 Définition du périmètre par application de filtres sur les données

- Initialement, les données sont séparées en deux bases selon les années d'observation. La première base concerne les données provenant de la crise sanitaire, observées sur 2020. La deuxième, se compose des données antérieures à la période de la Covid-19.
- Comme indiqué précédemment, seuls les individus qui ont une activité professionnelle nous intéressent dans ce mémoire. C'est pourquoi un filtre sur les salariés dont l'âge est compris entre 16 et 67 ans inclus est mis en place.
- De même, un filtre est établi sur les arrêts, pour ne conserver que ceux qui proviennent d'un congé maladie, d'un congé suite à un accident de trajet, une maladie professionnelle, un accident de travail ou de service ou pour une femme enceinte dispensée de travail. Tous les arrêts relatifs à la maternité, la paternité, à l'adoption ou encore au mi-temps thérapeutique sont supprimés de la base.
- Ensuite, des retraitements sur les variables de dates sont opérés. En effet, dans le but de mieux appréhender les rechutes d'arrêts, un traitement sur la date d'entrée en arrêt et sur la date de sortie d'arrêt doit être mis en place.
- Ensuite, à cause du manque de recul nécessaire pour comptabiliser la durée exacte des arrêts survenus en 2020 (car l'incapacité dure au maximum 36 mois), les arrêts sont bornés à 365 jours, soit un an. En effet, tous les arrêts survenus en 2018 ou en 2019 qui dépassent 365 jours sont tronqués et limités à un an. Ce retraitement a de fortes conséquences dans la modélisation de la durée d'arrêt. En revanche, pour calculer la probabilité de tomber en arrêt, cette modification ne va pas perturber les résultats finaux.

2.3.1.2 Modification des variables catégorielles

Dans un second temps, certaines variables ne sont pas directement exploitables de la DSN fournie. Elles font donc l'objet de différentes transformations.

Pour commencer, plusieurs modalités de variables catégorielles sont, soit inadaptées, soit trop floues, soit trop imprécises pour les besoins de notre étude et nécessitent donc un recodage :

- Libellé du contrat : initialement, cette variable indique le type de contrats de travail du salarié. En effet, selon la durée du contrat, l'activité de l'employeur ou encore la nature du travail confié au salarié, la nature du contrat diffère.

  À l'origine, cette variable contient 14 modalités distinctes. Finalement, dans l'optique de simplifier et de rendre exploitable cette variable, seulement deux modalités sont mises en place : CDD et CDI.

- Catégorie Socio-professionnelle : au départ, dans la base de données, cette variable correspond au statut conventionnel du salarié. Elle compte dix modalités, dont certaines qui ne sont que très peu représentées. Un regroupement en quatre modalités distinctes est alors opéré : Cadres, Employés, Ouvriers et Professions intermédiaires.

- Statut salarié : avant toute modification, cette variable définit le statut du salarié et contient cinq modalités. Pour les besoins de cette étude, seules deux modalités suffisent. C'est pourquoi elle ne contient désormais que les modalités Cadre ou Non cadre.

- Type de gestion : initialement, la variable a 29 modalités et chacune d'entre elles correspond à un code définissant le gestionnaire du contrat de prévoyance. Afin que cette variable dévienne intéressante à intégrer dans nos modèles, les 29 anciennes modalités sont transformées en trois nouvelles catégories distinctes.

2.3.1.3 Transformation du format des variables

Enfin, le format de certaines variables continues a du être modifié de façon à créer des catégories ou des tranches.

- Salaire annuel : premièrement, le salaire annuel de l'individu, présent dans les bases initiales, est transformé en salaire annualisé.

  Cela veut dire que, en fonction de la durée de présence en entreprise du salarié, le montant de son salaire est retravaillé afin de le ramener à une année entière. En effet, si un individu n'est présent qu'une partie de l'année, la variable relative à son salaire annuel comprend un biais.

  Le salaire annualisé est construit en additionnant tous les salaires mensuels perçus au cours de l'année, divisé par le ratio d'exposition de cet indentifiant. Par conséquent, pour un individu qui n'est présent dans l'entreprise que durant six mois, son salaire annualisé sera la somme de ces six mois de salaire divisé par 0,5.

  Deuxièmement, cette variable est découpée en différentes tranches distinctes. Effectivement, dans ce portefeuille, le panel de salaires annualisés est compris entre presque 10 000€ et plus de 300 000€.

  Le fait de conserver cette variable de manière continue n'est pas intéressant. En effet, entre un salarié qui gagne 40 000€ annuel et un autre qui gagne 40 010€, la différence ne peut être significative. C'est pourquoi une transformation en tranches semble le plus juste et le plus naturel pour observer l'effet du salaire sur le taux d'incidence.

- Ancienneté : le découpage en tranches d'ancienneté annuelle a été effectué pour essayer de capter les différents profils d'individus. En effet, en fonction du nombre d'année passé au sein d'une entreprise, le taux de turnover et le nombre d'arrêts observé varient fortement.

- Etablissement : le choix du découpage a été effectué afin d'avoir un nombre significatif de salariés dans chaque catégorie mais surtout en fonction des différents seuils de taille d'entreprise définis pour les déclarations sociales.

  En effet, les microentreprises ou très petites entreprises ont un effectif compris entre 1 et 10 salariés. Les petites entreprises se composent d'un effectif compris entre 11 et 50. Ensuite, viennent les moyennes entreprises avec un effectif qui se situe entre 51 et 250 personnes. Cette dernière catégorie est scindée en deux pour avoir des modalités représentatives de la population et équitables en termes de nombre d'individu.

2.3.2 Règles de gestion concernant les arrêts de travail

Dans cette partie, des retraitements sur des caractéristiques particulières des arrêts de travail sont mis en place.

2.3.2.1 Analyse des différentes franchises

Au cours de ce mémoire, le terme « *norme franchise 0 jour* » désigne l'incidence dès le premier jour de l'arrêt de travail. De manière similaire, la « *norme franchise 30 jours* » indique l'incidence à partir du 31ème jours, c'est à dire le nombre d'arrêts qui atteignent une durée au moins égale à 31 jours. Dans la suite, seuls des modèles avec une norme franchise de 0 jour seront présentés.

Ce terme de franchise est étudiée indépendemment de la réelle franchise de garantie définie dans le contrat. Pour rappel, la franchise du contrat est, dans la majorité des cas, comptée en nombre de jours. Elle désigne le laps de temps entre le moment où l'individu tombe en arrêt et le moment où son contrat de prévoyance prend effet. La prise d'effet se caractérise par le premier jour à partir duquel il va être indemnisé. C'est donc une période de temps incompressible pendant laquelle l'assuré ne touche aucune indemnisation de la part de l'organisme complémentaire, en cas d'arrêt de travail.

Après une étude de différents contrats de prévoyance, huit franchises prédominent et semblent majoritairement souscrites. En complément de ce mémoire, des modèles d'incidence avec norme franchise de 3j, 7j, 10j, 30j, 60j, 90j ou encore 365 jours seront mis en places. Chacune de ces normes d'incidence servira ensuite pour tarifer les contrats avec ces franchises associées.

La franchise constitue un élément primordial du contrat de prévoyance. Elle joue un rôle important dans la comparaison des organismes assureurs. C'est pourquoi il est nécessaire de l'inclure dans les modèles. Selon la franchise souscrite, la norme tarifaire proposée à l'individu sera différente.

Afin de mieux comprendre le phénomène de franchise, voici un exemple avec une franchise de 10 jours.

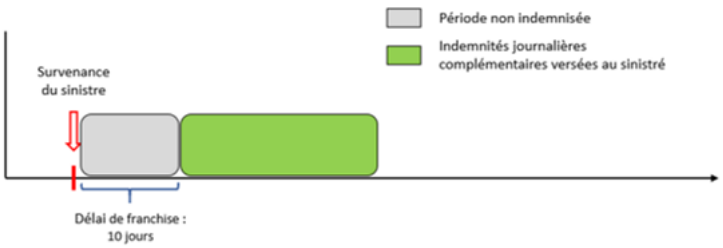


FIGURE 2.2 – Schéma explicatif du principe de franchise

2.3.2.2 Prise en compte du phénomène de rechute

Supposons qu'un individu soit victime d'un arrêt de travail pour une quelconque raison. À l'issue de cet arrêt, et si le salarié est guéri, il va reprendre le travail. Cependant, après la reprise de son emploi, il peut arriver que la pathologie initiale s'aggrave ou qu'une nouvelle lésion liée au premier arrêt apparaisse. Dans ce cas, le médecin du salarié peut constater ce qu'on appelle une rechute. Pour que le terme de rechute soit employé, il faut donc que la raison du nouvel arrêt soit en lien avec la raison de l'arrêt précédent.

Afin de comptabiliser au mieux le nombre de rechutes dans nos bases de données, il est nécessaire de préciser quels sont les critères qui vont les caractériser.

- Les deux arrêts doivent concerner le même individu ;
- Il faut que ce soit deux arrêts consécutifs ;
- Le motif des deux arrêts de travail doit être similaire (ex : accident de travail) ;
- Le délai entre deux arrêts est inférieur à 4 jours.

Le nombre de jours entre deux arrêts est relativement faible, cela permet de capter uniquement les arrêts issus d'une rechute et non un deuxième arrêt qui pourrait avoir le même motif mais n'avoir aucun lien avec le premier.

Lorsqu'une rechute est comptabilisée, la ligne correspondant à l'arrêt initial fait l'objet d'un traitement particulier. Dans un premier temps, il est donc nécessaire de regrouper le nombre de rechutes par individu et par année de déclaration DSN. En effet, une fois cette opération faite, toutes ces rechutes sont rattachées et regroupées dans un seul et unique arrêt. Cet arrêt comprend donc l'arrêt initial et toutes les rechutes qui en découlent.

Ensuite, il va être important de calculer au plus juste la durée d'arrêt, en nombre de jours, rechutes comprises. Pour ce faire, deux méthodes sont opposées :

- La durée totale de l'arrêt, en nombre de jours, pour chaque individu est la somme du nombre de jours de chaque arrêt (l'arrêt initial ainsi que toutes les rechutes suivantes).
- La durée totale de l'arrêt, en nombre de jours, pour chaque individu se calcule en faisant la différence entre la date de retour à l'emploi, après la dernière rechute, et la date de dernier jour travaillé, avant le debut de l'arrêt initial.

De prime abord, la première proposition semble être la meilleure, car plus précise. Cependant, après avoir comparé les deux méthodes, davantage d'erreurs étaient constatées avec cette première manière de calculer. En effet, en faisant la somme globale des durées de chaque arrêt, de nombreux chevauchements d'arrêts sont pris en compte.

Ces chevauchements entrainent une forte surestimation du nombre de jours d'arrêts et la prise en compte de plusieurs doublons sur les durées. Ce problème peut être due à des erreurs de gestion ou à des erreurs lors de la déclaration sociale nominative. En effet, la date de début d'arrêt (ou de début de rechute) est quasi systématiquement bien remplie cependant, c'est la date de fin d'arrêt qui est souvent erronée.

En revanche, l'utilisation de la deuxième méthode réduit les erreurs commises. Elle semble plus précise. Au plus, une surestimation de la durée de 4 jours est réalisée. Cela correspond au délai, entre deux arrêts consécutifs, pour définir la rechute. Finalement, en accord avec la direction, c'est la deuxième méthode de calcul qui a été privilégiée dans la suite de cette étude.

En outre, il est important de noter que lorsqu'on comptabilise une rechute, le délai de franchise ne s'applique pas comme lors de l'arrêt initial. La franchise entre en jeu seulement pour l'arrêt initial.

Voici un exemple concret pour un arrêt qui est suivi de 3 rechutes.

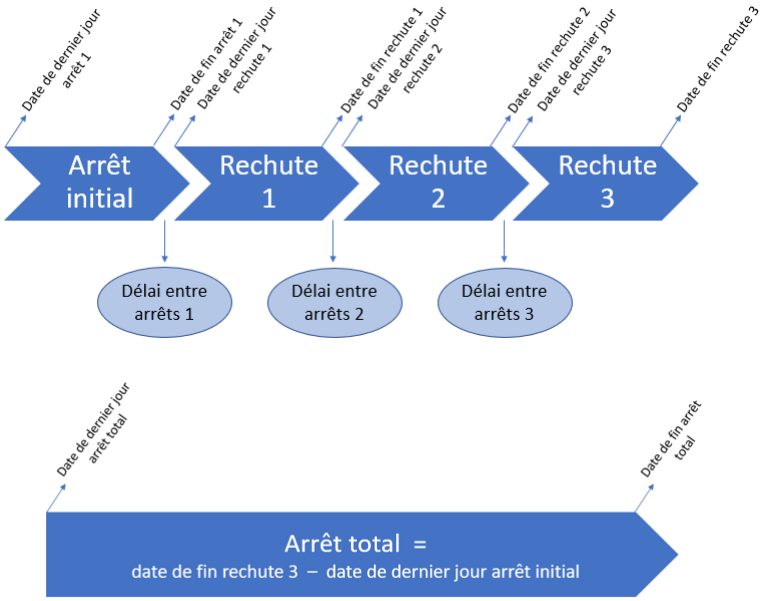


FIGURE 2.3 – Schéma explicatif du calcul des rechutes

Dans cet exemple, l'arrêt total se calcule en faisant la somme de : la durée de l'arrêt initial, la durée de la rechute 1, la durée de la rechute 2 et de la rechute 3, mais aussi du délai entre arrêts n° 1, du délai entre arrêts n° 2 et enfin du délai entre arrêts n° 3.

2.3.3 Règles de gestion concernant les valeurs aberrantes

Au cours de cette ultime partie concernant les retraitements de la base, des contrôles sont effectués pour éviter au maximum les anomalies dans les données.

2.3.3.1 Valeurs manquantes

Après vérification, la base finale ne contient aucune valeur manquante.

2.3.3.2 Prise en compte des doublons et des chevauchements d'arrêts

En réalisant de nombreux contrôles, des doublons ou des chevauchements d'arrêts ont été detecté. Un doublon est définit lorsque deux ou plusieurs lignes concernant le même identifiant et le même arrêt sont présentes dans la base. En général, c'est majoritairement la valeur du salaire qui évoluait

entre les différentes lignes de doublons. Ce problème concerne environ 0,1% des lignes. En analysant ces lignes de doublons, la solution envisagée est de ne conserver que la ligne qui semble être la plus cohérente et la plus juste, en supprimant de la base toutes les autres lignes.

De plus, quelques lignes ont été identifiées comme ayant des dates d'arrêts (date de dernier jour travaillé ou date de reprise du travail) qui se chevauchent. Heureusement, cette erreur n'est que très rare et concerne uniquement des lignes modifiées.

En effet, certaines lignes subissent des modifications après l'envoi de la DSN. Par exemple, s'il y a une erreur de frappe ou q'un arrêt se prolonge à la suite d'une aggravation de l'inaptitude au travail, la déclaration est modifiée. Pour un identifiant et un arrêt considéré, l'hypothèse retenue est de conserver uniquement la ligne définissant le minimum des dates de derniers jours travaillés et le maximum des dates de reprises.

2.3.3.3 Contrôle des valeurs aberrantes

Un ensemble de tests est effectué pour supprimer au maximum les valeurs aberrantes de la base de données. Certaines de ces valeurs extrêmes sont facilement identifiables alors que d'autres le sont beaucoup moins.

Dans la suite, une liste non exhaustive des différents retraitements est présentée. Dans la majorité des cas et lorsque l'incohérence ne pouvait être résolue, ces lignes aberrantes ont été supprimées de la base. En effet, l'ensemble de toutes ces valeurs incorrectes ne représente qu'une infime partie de cette base et leur suppression est donc sans conséquence sur la qualité finale.

Dans un premier temps, l'âge du salarié est analysé, quelques valeurs incohérentes sont identifées. Elles sont désormais exclues du portefeuille.

Ensuite, la durée des arrêts a été recalculée. Plusieurs résultats étaient bien trop importants pour être réalistes. En effet, un arrêt d'une durée égale à plus de 10 000 jours avait été détecté. Pour rappel, l'incapacité dure au maximum 36 mois. Au dela, si l'arrêt est toujours en cours, l'individu bascule dans l'état d'invalidité. C'est tout à fait possible qu'un individu reste en invalidité durant 27 ans, néanmoins comme seul le risque incapacité est étudé dans ce mémoire, cette ligne doit être supprimée.

Enfin, quelques soucis avec les variables relatives aux montants de salaire ont été rencontrés. Des montants négatifs ont été répertoriés ainsi que des montants correspondants aux indemnités perçues par le salarié à la suite d'un arrêt. Pour certaines de ces lignes, une correction a pu être apportée en se référant aux autres variables de salaire présentes dans la base. Tandis que pour les autres, aucune correction n'a pu être toruvée, elles ont donc été supprimées.

2.3.3.4 Corrélation entre les variables

Il est évident qu'au vu du nombre de variables utilisées dans la « base salarié », certaines d'entre elles sont corrélées les unes aux autres.

– Il y a une corrélation non négligeable entre l'âge du salarié et les tranches d'ancienneté. En effet, il n'est pas possible pour un individu d'une vingtaine d'années d'avoir plus de dix ans d'ancienneté dans la même entreprise. Par conséquent, les tranches d'ancienneté les plus

fortes recensent une majorité de personnes ayant un âge avancé. Cependant, après des petites études sur ces deux variables, il est important de les conserver dans la base finale.

– Ensuite, il y a une très forte corrélation entre la modalité cadre de la variable relative à la catégorie socio-professionnelle et la modalité cadre de la variable statut salarié. Néanmoins, dans l'ensemble des modèles construits, la variable statut salarié n'est jamais significative, donc seule la modalité cadre de la variable CSP sera prise en compte dans les modélisations.

– De plus, il semble y avoir une corrélation entre le salaire annualisé du salarié et son âge. En effet, de manière logique, plus l'individu est âgé, plus il acquiert de l'expérience et du savoir faire et donc meilleur son salaire est. Pour tenter de minimiser cet effet, des tranches de salaire sont préférées. De plus, ces deux variables sont significatives dans les modèles et semblent expliquer des aspects différents sur la fréquence d'arrêt. C'est pourquoi elles sont toutes les deux conservées dans les modèles.

– Enfin, il peut y avoir une corrélation entre la variable statut salarié et la variable sur les salaires annuels. Effectivement, un individu cadre aura tendance à avoir un meilleur salaire qu'un individu non cadre.

2.4 Statistiques descriptives

2.4.1 Statistiques globales sur le portefeuille

Les statistiques présentées au cours de cette partie sont effectuées sur la population d'assurés d'AG2R La Mondiale. Pour chacune des années observées, la population sous risque est comptabilisée ci-dessous.

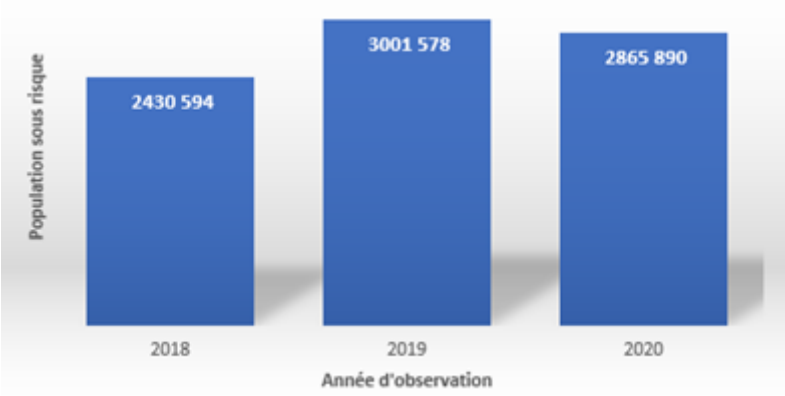


FIGURE 2.4 – Évolution de la population assurée

La population sous risque correspond au nombre d'assurés qui possèdent un contrat de prévoyance au sein du groupe AG2R La Mondiale. Elle atteint un pic pour l'année 2019, avec plus de 3 millions de salariés comptabilisés. En se basant sur des tests réalisés depuis la mise en place de la DSN mensuelle, la qualité et la précision des informations transmises s'améliorent d'année en année.

La maitrise de la DSN fournie par les entreprises et la signature d'affaires nouvelles sont les deux principales causes de la forte hausse constatée en 2019.

En revanche, l'année 2020 est marquée par la diminution de la population assurée au sein de notre portefeuille français. Cette réduction avoisine les 4% et s'explique notamment par l'augmentation du taux de chômage durant la crise sanitaire. En effet, le graphique suivant provient de l'INSEE et présente les évolutions du taux de chômage en France métropolitaine entre 2003 et 2022.

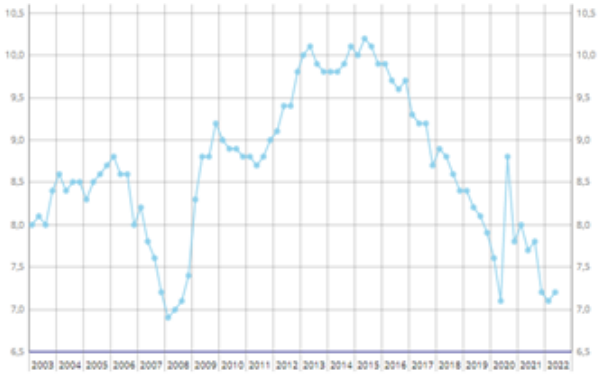


FIGURE 2.5 – Évolution du taux de chômage en France métropolitaine
[2]

Depuis la moitié de l'année 2015, le taux de chômage a une tendance décroissante et monotone. Cependant, le taux de chômage observé début 2020, qui était au plus bas depuis plus d'une dizaine d'années, a subi un pic d'augmentation. En effet, cette hausse est de l'ordre de 25% puisque le taux de chômage passe de 7,1% au début de l'année 2020 à 8,8% en fin d'année. Cette hausse brutale du chômage additionnée au gel des embauches durant toute l'année explique la diminution de notre population d'assurés.

Dans le but d'approfondir les événements qui touchent la population sous risque, l'évolution du nombre d'entreprises, recensées dans notre base de données, est désormais étudiée.

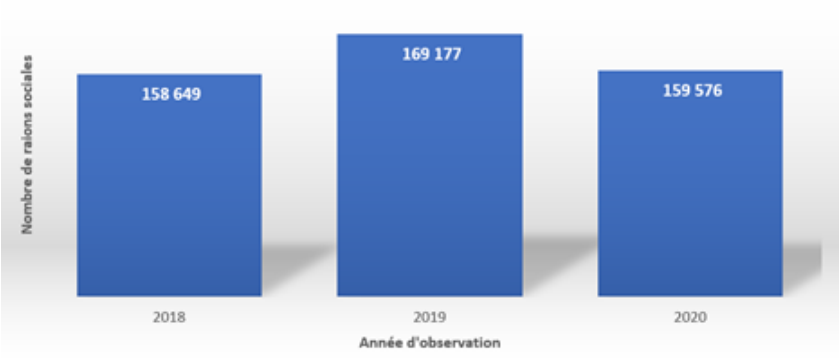


FIGURE 2.6 – Évolution du nombre de raisons sociales

Comme précédemment, le nombre de raisons sociales atteint son maximum en 2019. Néanmoins, ce nombre est quasiment identique pour les années d'observation 2018 et 2020, mais presque 7% plus faible qu'en 2019. Finalement, le nombre de raisons sociales recensées dans ce portefeuille suit les variations de la population sous risque.

2. Source : INSEE

Ensuite, grâce à l'apport de la DSN, une analyse sur la population sinistrée par rapport à celle non sinistrée est réalisée.

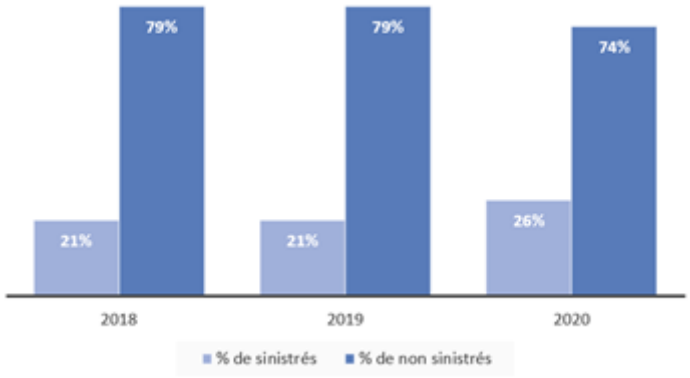


FIGURE 2.7 – Répartition des sinistrés et des non sinistrés par année

Les années 2018 et 2019 se comportent de manière identique sur cette répartition. Durant ces années, un peu plus d'un cinquième (21%) des assurés sont victimes d'un ou plusieurs arrêts de travail. En contrepartie, c'est presque 8 salariés sur 10 qui n'ont aucun arrêt au cours de cette période. C'est une des raisons pour laquelle ces deux années sont regroupées dans une seule et même base pour la construction des modèles futurs.

En revanche, cette répartition évolue sur l'année 2020. Une augmentation de la population sinistrée est constatée, pour atteindre un taux égal à 26%. Par opposition, le pourcentage de non sinistrés diminue et se stabilise à 74%. Ce graphe confirme nos suppositions : la Covid-19 a impacté la probabilité de tomber en arrêt au sein de ce portefeuille. Le rapport entre les sinistrés et les non sinistrés passe de 0,26 à 0,35.

Désormais, le nombre de sinistres comptabilisé pour chaque année d'observation est présenté. La déclaration des arrêts de travail est différente de la DSN classique mensuelle. En effet, elle fait partie des déclarations dites événementielles et doit être transmise dès la survenance de l'arrêt ou du phénomène exceptionnel (rupture de contrat, reprise du travail, ...). Par conséquent, le nombre d'arrêt observés en 2018 est surement un peu sous-estimé.

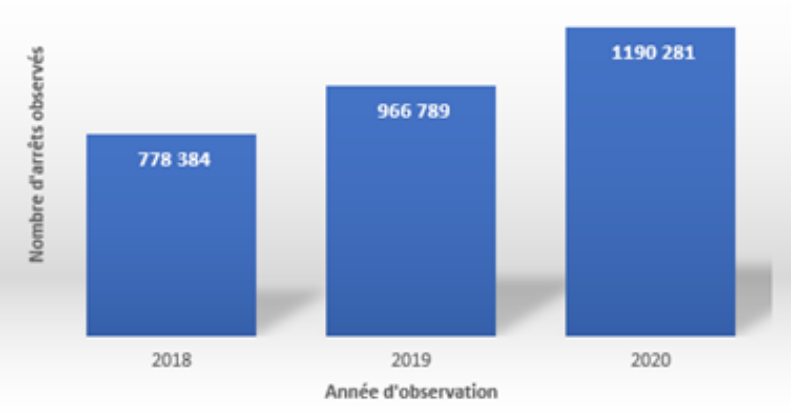


FIGURE 2.8 – Évolution du nombre d'arrêts observé

Le nombre d'arrêts observés dans notre portefeuille augmente de manière linéaire au cours du temps. Entre 2018 et 2019, la hausse du nombre de sinistre avoisine les 24%. Les erreurs lors de la déclaration événementielle et l'augmentation de la population sous risque expliquent cette hausse.

Entre 2019 et 2020, il y a encore une augmentation de presque 23%. Malgré une diminution du nombre d'assurés, le nombre d'arrêts ne cesse d'augmenter au cours de cette période de crise. Cette augmentation soudaine de sinistres va se traduire par une explosion des prestations versées aux assurés.

Dans la suite, l'exposition moyenne des assurés au sein de leur entreprise est analysée. L'exposition correspond au temps de présence du salarié dans l'entreprise au cours de l'année considérée, donné en pourcentage. Par exemple, un salarié embauché le 1er avril 2019 qui est toujours en contrat aujourd'hui aura donc une exposition de 0% pour l'année 2018, de 9/12 soit 75% en 2019 et de 100% en 2020.

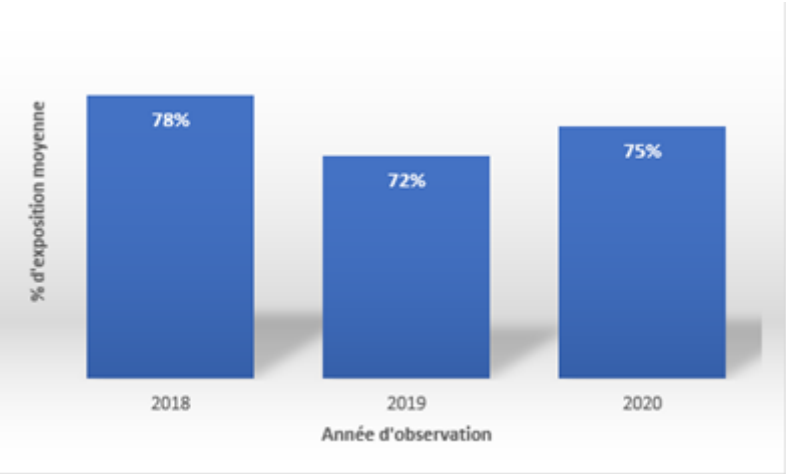


FIGURE 2.9 – Évolution de l'exposition moyenne

L'année 2018 est donc caractérisée par l'exposition la plus élevée. Une exposition égale à 78% veut donc dire que les assurés sont présents dans l'entreprise en moyenne 10 mois de l'année sur 12. L'exposition est donc un bon indicateur pour quantifier à la fois le turnover, les nouvelles embauches mais également les fins de contrats (licenciement ou retraite) par exemple.

La baisse de 6% d'exposition moyenne concernant l'année 2019 s'explique par plusieurs raisons. Premièrement, le marché de l'emploi va très bien et le nombre de déclarations d'embauche ne cesse d'augmenter. Deuxièmement, selon les chiffres de l'INSEE, le taux de turnover atteint un pic et se situe entre 5 et 15% au cours de cette année. Néanmoins, la bonne santé du marché de l'emploi n'a pas résisté à la crise sanitaire qui a frappé le pays en 2020. Les nouvelles embauches ont subi un sérieux revers au cours de cette année. C'est principalement ce qui explique la hausse d'exposition sur l'année 2020.

Enfin, l'incidence moyenne est étudiée. Pour rappel, l'incidence caractérise la probabilité que l'assuré tombe en arrêt au cours de l'année considérée. Cette variable est au cœur de toutes nos modélisations.

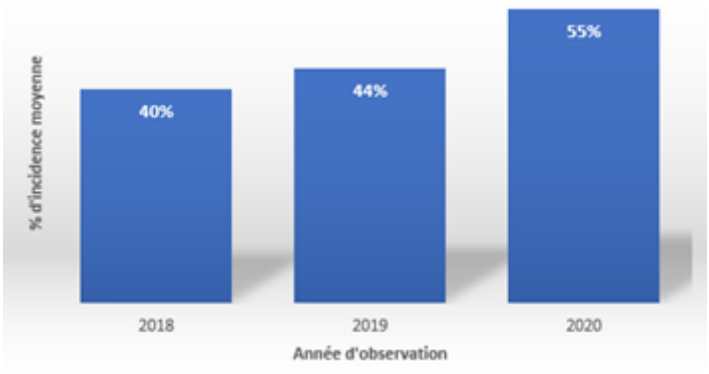


FIGURE 2.10 – Évolution de l'incidence moyenne

L'incidence de l'année 2018 est de 40%. Celle de 2019 augmente légèrement (+4%) mais reste dans le même ordre de grandeur. En effet, le nombre d'arrêts a augmenté plus rapidement que la population sous risque et ce nombre d'arrêts semble réparti de manière plus homogène entre les individus.

Un focus sur l'année impactée par la pandémie mondiale est fait. L'incidence moyenne explose comparée aux années précédentes, +15% par rapport à 2018 et +11% par rapport à 2019. Désormais, en moyenne, un assuré présent dans le portefeuille a une incidence de 55%. Une fois de plus ce graphe confirme l'hypothèse que la crise de la Covid-19 a principalement heurté la fréquence d'arrêt.

De plus, il est important de faire la distinction entre les deux points suivants. D'un côté, le nombre de salariés qui ont eu un arrêt en 2020 représente 26% des assurés de ce portefeuille, soit un salarié sur quatre. D'un autre côté, l'incidence moyenne du portefeuille est de 55%. Elle désigne le nombre d'arrêts observés en 2020 divisé par l'exposition annuelle. Ce qui signifie que chaque personne faisant un arrêt en 2020, en fait 2,1 en moyenne.

2.4.2 Étude sur les variables qui influencent le nombre d'arrêts

L'analyse des phénomènes les plus marquants est réalisée au cours de cette partie. Par conséquent, ces statistiques se concentrent sur les variables qui entraînent des enseignements majeurs. L'études des autres variables est présente en annexe 4.3.5 .

- Clé de lecture des graphiques ci-dessous

Le graphique combine un histogramme (en dessous) et des courbes présentant les taux d'incidence (au-dessus). L'histogramme présente le nombre de salariés pour chacune des catégories de la variable considérée (genre, CSP, tranches d'établissement...).

L'histogramme se lit à l'aide de l'axe de gauche, qui correspond à un nombre d'individus, comptabilisé en milliers de personnes. Au-dessus, il y a une courbe d'incidence pour chaque année d'observation. Ces courbes sont construites comme la somme des nombres d'arrêts divisée par la somme des ratios d'exposition pour chacune des catégories de la variable considérée, exprimée en pourcentage. Chaque point de la courbe représente l'incidence d'une modalité de la variable étudiée. Ils sont ensuite reliés pour former une courbe. L'axe de droite, exprimé en pourcentage, sert à la lecture de l'incidence.

Trois couleurs sont utilisées. Chaque couleur correspond à une unique année d'observation et est défini comme suit :

Année 2018 : Année 2019 : Année 2020 :

2.4.2.1 Répartition des salariés par tranches d'âge

Pour faciliter l'analyse, l'âge des salariés du portefeuille n'est pas étudié comme tel. Le caractère continu de la variable est conservé. Néanmoins, l'analyse porte sur des tranches d'âge, d'intervalles égaux à cinq ans (excepté pour la dernière tranche, 7 ans). L'âge moyen des assurés présents dans le portefeuille est relativement stable puisqu'il oscille entre 38,4 ans en 2018 et 37,7 ans en 2020. Malgré des variations minimes, le fait que la population sous risque de 2020 soit légèrement plus jeune va entraîner de fortes conséquences sur l'étude de l'incidence. Mécaniquement, si l'âge moyen du portefeuille est inférieur cela implique des assurés en meilleure santé et avec un salaire inférieur. Par conséquent, les indemnités versées en cas d'arrêt de travail ne sont pas les mêmes.

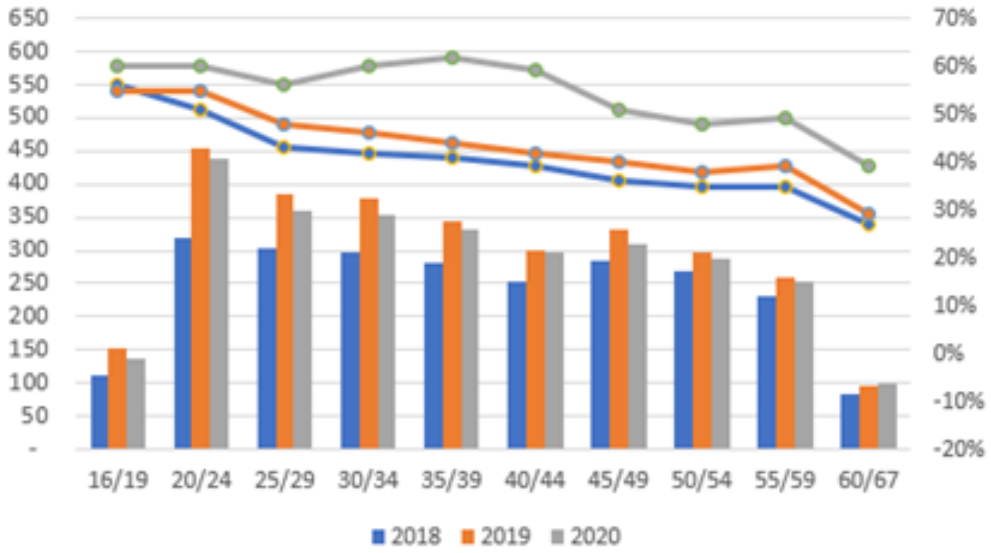


FIGURE 2.11 – Répartition des assurés en fonction de leur âge

Étudier la répartition par tranches d'âge permet de se rendre compte de quelques points importants. Premièrement, les assurés les plus jeunes (moins de 19 ans) et les plus âgés (plus de 60 ans) sont présents dans des proportions bien plus faibles que les autres tranches d'âge. Deuxièmement, pour les trois années d'observation, la tranche la plus représentée comprend les salariés dont l'âge est compris entre 20 et 24 ans. Enfin, l'évolution de toutes les tranches d'âge semble suivre celle de la population sous risque.

S'agissant des taux d'incidence, les années 2018 et 2019 se comportent de manière similaire. Les individus les plus jeunes (entre 16 et 25 ans) font davantage d'arrêts. Ils comptabilisent une moyenne de 0,55 arrêt par an et par salarié. En revanche, les salariés les plus âgés (entre 60 et 67 ans) ont un taux d'incidence relativement inférieur au reste du portefeuille. Ce taux est égal à 28%.

Par ailleurs, pour les années 2018 et 2019, l'incidence évolue de manière décroissante en fonction de l'âge et de façon presque linéaire. Des similitudes existent pour l'année frappée par la pandémie. En effet, la courbe des taux est également décroissante en fonction de l'âge. Les moins de 25 ans

ont toujours l'incidence la plus forte (60%) et les plus de 60 ans, l'incidence la plus faible (40%). En revanche, la tendance n'est clairement pas linéaire. Un nouveau phénomène semble se produire pour les individus dont l'âge se trouve entre 30 et 45 ans. Une bosse apparaît. Principalement durant les périodes de confinements, de nombreux salariés de cet intervalle d'âge ont dû poser des arrêts dérogatoires, pour motif de garde d'enfants par exemple, ou ont été contraints au chômage partiel. Ces motifs d'arrêts n'existaient pas auparavant. Par conséquent, cela entraîne une sur-déclaration d'arrêts qui explique une partie de cette hausse d'incidence.

Par ailleurs, la crise sanitaire a également touché les salariés de tous les âges. La courbe d'incidence relative à l'année 2020, en gris, est formellement supérieure aux autres. La hausse du taux d'incidence est moins marquée chez les jeunes, avec seulement +4%. Pour les autres tranches d'âge, le gap d'incidence lié à la crise sanitaire est compris entre +7% et +20%.

	[16 ;19]	[20 ;24]	[25 ;29]	[30 ;34]	[35 ;39]	[40 ;44]	[45 ;49]	[50 ;54]	[55 ;59]	[60 ;67]	**Moy**
2018 + 2019	47%	56%	70%	76%	79%	81%	83%	86%	88%	81%	**75%**
2020	50%	58%	70%	76%	80%	81%	83%	86%	88%	81%	**75%**

Sur les deux bases de données, l'exposition évolue de manière croissante en fonction de l'âge. Les plus jeunes ont une exposition inférieure à la moyenne. Deux raisons principales rentrent en jeu. Premièrement, depuis quelques années, le taux d'emploi des jeunes ne cesse de croître. Deuxièmement et de manière générale en France, le turnover concerne majoritairement de jeunes salariés. Ils changent d'emploi plus régulièrement dans l'optique d'évoluer plus rapidement dans leur carrière professionnelle et parce qu'ils ont en général moins de contraintes personnelles.

À contrario, les salariés d'un âge plus avancé ont souvent une situation personnelle plus stable et ont d'autres attentes au niveau professionnel. Ils préfèrent rester dans une entreprise où ils se sentent bien, maintenir leur équilibre vie privée vie professionnelle. Ils sont donc moins souvent sujets au turnover. Par conséquent, le taux d'exposition des individus âgés de 40 ans et plus est plus élevé. Il dépasse les 80%.

2.4.2.2 Répartition des salariés par Catégorie Socio-Professionnelle

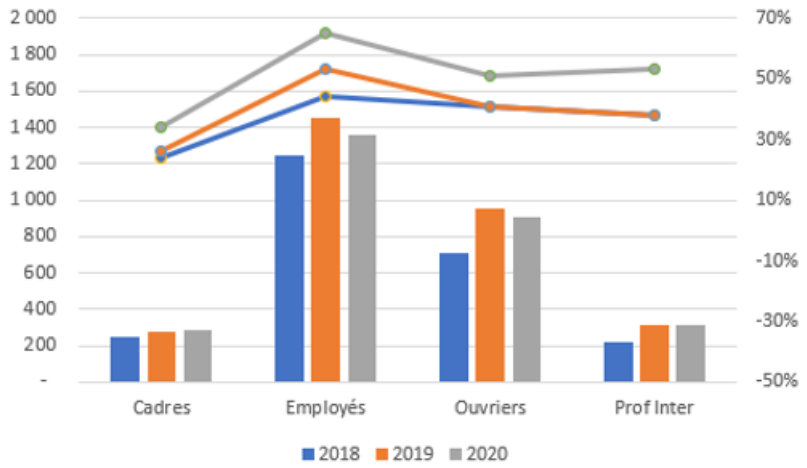


FIGURE 2.12 – Répartition des assurés en fonction de leur CSP

L'analyse se porte désormais sur la catégorie socio-professionnelle des salariés. Ce portefeuille se compose d'une majorité d' « Employés » et ceux pour les trois années d'observation. De plus, les « Cadres » ainsi que ceux qui exercent des professions intermédiaires (techniciens, agents de maîtrise ou contremaîtres par exemple) sont les moins représentés dans notre portefeuille. En moyenne, il y a environ 30 000 salariés par modalité.

De plus, la hausse de la population sous risque entre 2018 et 2019 semble plus importante pour les CSP « Employés » et « Ouvriers ». Pour rappel, la hausse de la population est en grande partie absorbée par une population masculine. En France, les salariés ouvriers sont principalement représentés par des hommes. Ensuite, la baisse générale du nombre de salariés dans le portefeuille entre 2019 et 2020 ne se répercute pas sur les catégories « Cadres » et « Professions Intermédiaires ».

Pour les CSP cadres, ouvriers et professions intermédiaires, le taux d'incidence au cours de l'année est identique en 2018 et en 2019. En revanche, une hausse de l'incidence est observée chez les employés en 2019. Elle passe de 44% en 2018 à 53% en 2019.

Ensuite, la crise sanitaire impacte de manière significative chacune des catégories socio-professionnelles puisque l'augmentation moyenne du taux d'incidence est de plus de 10% sur l'ensemble des quatre modalités.

	Cadres	Employés	Ouvriers	Prof. Inter.	**Moy**
2018 + 2019	92%	69%	76%	86%	**75%**
2020	92%	69%	76%	86%	**75%**

Concernant les deux bases étudiées, l'évolution de l'exposition a exactement la même allure. Cependant, les cadres et les professions intermédiaires tirent clairement l'exposition moyenne vers le haut avec respectivement un taux de 92% et 86%. En nombre de mois, cela équivaut à une présence entre dix et onze mois par an.

Ensuite, les ouvriers ont une exposition qui se stabilise au cours du temps aux alentours de l'exposition moyenne. En revanche, les salariés employés ont l'exposition la plus faible. En effet, en 2018/2019 comme en 2020, leur taux de présence n'est que de 69%, soit environ 8 mois par an.

2.4.2.3 Répartition des salariés par taille d'entreprise

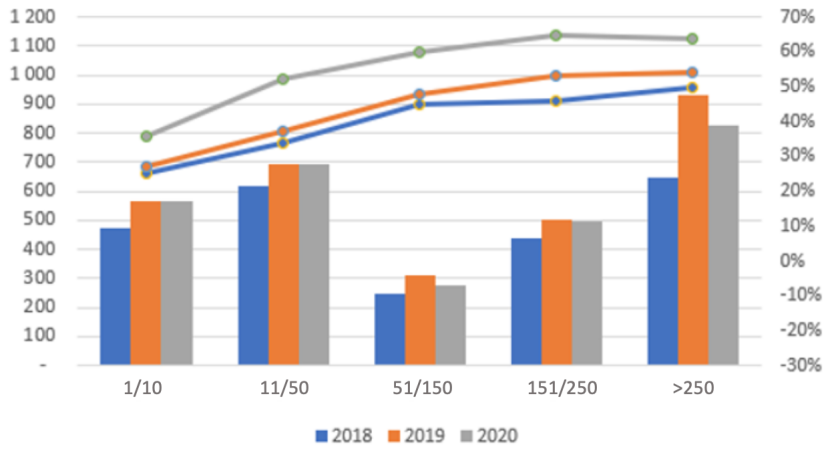


FIGURE 2.13 – Répartition des assurés en fonction de la taille de leur entreprise

L'étude des statistiques descriptives se porte désormais sur la taille des entreprises. Environ un tiers des salariés de ce portefeuille occupent un poste dans une entreprise de moins de 50 personnes. À contrario, quasiment 30% des assurés font partie de grands groupes de plus de 250 salariés. Entre ces deux extrêmes, les assurés sont répartis dans des entreprises de tailles diverses.

Concernant l'année 2019, la hausse de la population est principalement absorbée par les salariés des grandes entreprises. En effet, pour cette catégorie l'augmentation du nombre d'individus est de 30% alors que pour les autres modalités elle n'est que de 10%. En 2020, le même phénomène se produit. La baisse de la population touche quasi uniquement les entreprises de plus de 250 salariés. Effectivement, cette diminution est de l'ordre de 11%. Pour les entreprises de taille plus modeste, la répartition du nombre de salariés est similaire à l'année 2019.

Quelles que soient les années, les courbes d'incidence semblent suivre la même tendance de croissance. Plus la taille de l'entreprise augmente et plus le taux d'incidence des employés semble fort. En général, les employés de petites structures sont moins bien couverts par leur complémentaire lors d'arrêts de travail. Les couvertures négociées par les grosses entreprises comprennent généralement davantage de garanties et à des tarifs souvent préférentiels grâce à la mutualisation du risque.

	1 à 10 sal.	11 à 50 sal.	51 à 150 sal.	151 à 250 sal.	+250 sal.	**Moy**
2018 + 2019	74%	74%	73%	75%	77%	**75%**
2020	75%	75%	73%	76%	76%	**75%**

L'exposition est constante quelle que soit la taille de l'entreprise et quelle que soit l'année observée. Effectivement, que l'entreprise emploie uniquement 1 ou plus de 250 salariés, l'exposition est proche de la moyenne, entre 73% et 77%, soit environ 9 mois de présence annuelle.

2.4.2.4 Répartition des salariés par tranches de salaire annuel

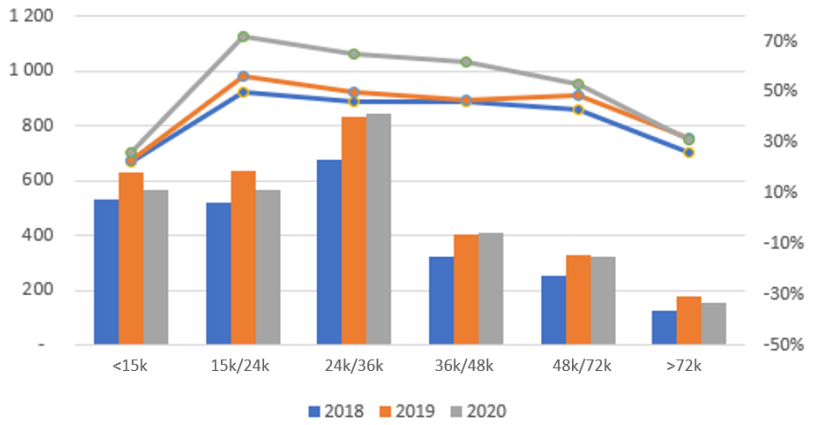


FIGURE 2.14 – Répartition des assurés en fonction de leur tranche de salaire

La répartition des salaires annuels semble suivre l'allure de la pyramide des salaires française. En effet, une grande majorité de nos assurés gagnent entre quinze et trente-six mille euros par an. Cela correspond à un salaire mensuel compris entre le SMIC et 3 000€. De plus, tous les salariés compris dans la tranche inférieure gagnent moins que le salaire minimum interprofessionnel de croissance (SMIC). Ce sont majoritairement des individus qui travaillent à temps partiel ou sur des missions de courte durée, tel que l'intérim. Ensuite, plus le salaire annuel augmente et plus le nombre d'individus diminue.

Entre les années 2018 et 2019, l'évolution démographique du portefeuille suit la même tendance pour chaque tranche de salaire. Quel que soit le niveau de salaire, une hausse de 15% est observée. Néanmoins, ce sont principalement les salariés qui ont de modestes revenus qui sont touchés par la baisse de population relative à l'année 2020. Pour les salaires inférieurs à 24 000€, une diminution de 10% du nombre d'individus présents dans le portefeuille est constatée, contre une stagnation pour les autres tranches.

Pour chacune des années observées, la plus forte incidence est représentée par les individus qui ont un revenu mensuel compris entre le SMIC et 2 000€. En dessous de ce palier, les salariés ont une incidence plus faible. Pour rappel, dans cette tranche se trouve principalement des salariés à temps partiel. Ils sont donc dans des situations moins stables et plus précaires que les autres individus de ce portefeuille. Par conséquent, une baisse de revenu causée par un arrêt de travail serait trop difficile à supporter pour eux. Au-dessus de ce palier, l'incidence décroît avec l'augmentation du salaire. Les individus avec les salaires les plus élevés occupent des postes de bureaux, donc paradoxalement moins contraignants physiquement. De plus, ils ont généralement plus de responsabilités. Ils ont donc plus de mal à s'absenter de leur entreprise, ce qui explique un taux d'incidence plus faible.

	< 15k€	[15k ;24k[	[24k ;36k[	[36k ;48k[	[48k ;72k[	> 72k€	**Moy**
2018 + 2019	65%	74%	76%	73%	74%	74%	**75%**
2020	74%	64%	71%	77%	76%	81%	**75%**

Alors que l'exposition est plutôt stable sur la base antérieure à la crise sanitaire, elle l'est beaucoup moins sur les données en 2020. En effet, avant la pandémie, quel que soit le salaire perçu, les salariés sont présents en entreprise durant les trois quarts de l'année, soit environ 9 mois par an. En revanche, durant l'année 2020, l'exposition par tranche de salaire est un peu plus volatile. Une tendance générale se dégage. Les individus qui perçoivent de modestes revenus ont une exposition plus faible. Une différence d'exposition moyenne de 17% est observée entre ces deux extrémités. De plus, un phénomène semble se produire pour la tranche de salaire comprise entre 24 K€ et 36 K€. En effet, un abaissement de l'exposition moyenne est observé, pour atteindre 71%.

2.4.2.5 Répartition des salariés par tranches d'ancienneté

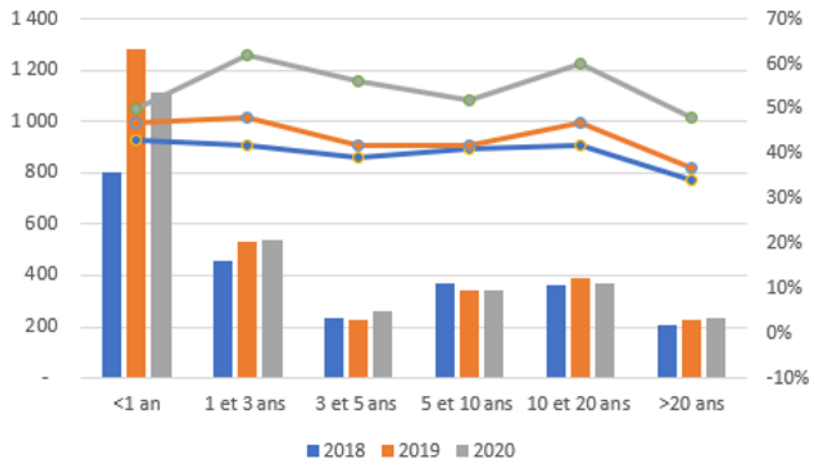


FIGURE 2.15 – Répartition des assurés en fonction de leur tranche d'ancienneté

Force est de constater que les salariés qui ont une ancienneté inférieure à un an sont en majorité dans le portefeuille. Suivant l'année d'observation, ils représentent entre 32% et 43% de la population

globale. De plus, pour cette tranche d'ancienneté, la répartition du nombre d'individus fluctue au fil des années. En effet, entre 2018 et 2019, une hausse de 40% est recensée. En revanche, entre 2019 et 2020 c'est un abaissement de presque 16% qui se dégage. Une majorité de ce portefeuille semble très enclin au turnover et ne travaillera donc pas pour la même entreprise tout au long de sa carrière professionnelle.

Les autres modalités sont plus stables au cours du temps. Le nombre d'individus présents dans chaque modalité ne varie que légèrement entre 2018 et 2020.

Concernant les trois années d'observation, deux pics d'incidence sont identifiables. Les anciennetés comprises entre un et trois ans et entre dix et vingt ans sont donc plus touchées par l'absentéisme. En effet, plusieurs raisons peuvent être responsables de cette hausse. Après un an d'ancienneté et une fois passé le temps d'adaptation à une nouvelle équipe, les éventuelles périodes d'essais, ou la baisse d'implication et de motivation liée à une situation confortable entraînent les salariés à faire plus d'arrêts.

En outre, au fur et à mesure que l'ancienneté augmente, les salariés acquièrent de nouveaux avantages lors de la prise en charge des arrêts de travail, notamment par la sécurité sociale. Prenons l'exemple de la loi de mensualisation où pour chaque tranche d'ancienneté de cinq ans, le salarié bénéfice d'une durée supplémentaire de 10 jours pour le maintien de son salaire. En revanche, pour les autres modalités, l'incidence semble plus stable et se situe aux alentours de 40% pour les années 2018 et 2019 et proche de 55% pour l'année 2020.

Par définition, l'exposition est corrélée avec l'ancienneté du salarié en entreprise. C'est pourquoi le tableau récapitulant les différentes expositions par modalité n'est pas présenté.

3 Construction des modèles théoriques

Au cours de la partie précédente, centrée sur le portefeuille d'assurés, le périmètre des données a été défini. Ensuite, les différentes étapes de construction de la base finale sont énumérées. Enfin, une analyse est réalisée sur les variables influentes par le biais de statistiques descriptives.

Dans cette troisième partie, les différents modèles mis en place pour ces travaux sont détaillés pour en comprendre leur fonctionnement. Premièrement, différents items nécessairent pour l'implémentation des modèles sont listés, notamment les méthodes de validation croisée. Deuxièmement, le présentation générale des différents modèles de prédictions est détaillée. D'abord, cette étude s'articule autour de la mise en place des modèles linéaires généralisés, puis des modèles qui reposent sur l'algorithme de boosting et enfin des modèles utilisés par un outil de tarification appelé Akur8.

3.1 Méthodes préalables pour l'implémentation des modèles

Les assurés qui composent ce portefeuille sont répartis dans la France entière. Par conséquent, la volumétrie des données utilisées est assez importante. C'est pourquoi toutes les modélisations et les calculs présents dans ce mémoire ont été effectués à l'aide du langage Python.

3.1.1 Transformation des variables catégorielles

Pour que les variables catégorielles soient prises en compte dans les modèles de Machine Learning, elles doivent, au préalable, être transformées. Cette modification se caractérise par la division des variables catégorielles en variables dichotomiques distinctes.

À l'aide de la commande « *get_dummmies()* » du package panda sous Python, ce processus est appliqué à chacune des variables catégorielles de la base de données. Cette modification permettra de déterminer l'individu de référence et de mieux segmenter le portefeuille.

3.1.2 Optimisation par validation croisée

La méthode dite de *« Cross validation »* permet d'améliorer les performances d'un modèle prédictif de Machine Learning. Cette technique consiste à mettre de côté un échantillon de l'ensemble de données. Sur ce fragment, le modèle n'est pas entraîné. Plus tard, il servira pour évaluer les performances du modèle.

Il existe plusieurs techniques de validation croisée. Deux d'entre elles sont utilisées au cours de ce mémoire : la séparation de la base finale en deux sous-ensembles et la technique des K-folds.

a) Train-Test Split : Séparation de la base

En Machine Learning, avant d'exécuter un algorithme sur l'ensemble des données, il est possible de les diviser en deux sous-ensembles. En effet, cela va permettre d'augmenter les performances des modèles. Ce processus s'appelle la méthode *Train-Test Split*. Le premier sous-ensemble correspond à l'ensemble de *formation* ou *d'entrainement*. Il est souvent compris entre 80% et 90% du portefeuille global. Le second, moins conséquent, correspond à l'ensemble de *test*.

Lorsque le modèle de Machine Learning est mis en place, le but n'est pas qu'il soit performant sur les données d'entrainement mais plutôt sur les données de tests. En effet, la validation de modèle consiste à mesurer la performance de celui-ci sur de nouvelles données, inconnues auparavant.

De plus, faire cette scission permet de réduire les problèmes de sous ou sur-apprentissage :

– Sous-apprentissage ou « under-fitting » en anglais :

  En Machine Learning, un modèle d'apprentissage est construit pour atteindre un objectif fixé. Or, il se peut que l'apprentissage automatique n'assimile pas correctement les facteurs de segmentations, indispensables pour aboutir à de bonnes prédictions. Il n'intègre donc pas les tendances générales et ne peut se généraliser correctement aux données d'entrainement. Ce modèle est donc sous-adapté, il possède un biais élevé. Ce biais entraine l'augmentation des erreurs de manière importante.

  Les modèles complexes sont moins soumis au problème de sous-ajustement car les paramètres s'ajustent et apprennent des informations compliquées.

– Sur-apprentissage ou « over-fitting » en anglais :

  De manière contraire, le modèle risque de se spécialiser sur les données d'entrainement et ne pas parvenir à se généraliser lorsque les données d'entrée changent. Cela veut dire que la précision sur les données d'entrainement est très voire trop poussée. Au contraire, la précision de l'ensemble de test est détériorée. C'est donc ce qu'on appelle le sur-ajustement.

  Par conséquent, si le modèle présente de trop bons résultats et des performances quasi parfaites sur les données d'entrainement, il faut se méfier. Les modèles très complexes, avec une variance élevée, souffrent de ce problème. Il y a donc un compromis à faire entre variance élevée et biais élevé.

C'est pourquoi, pour pallier cet éventuel problème, l'ensemble des données est scindé en deux. Cette division s'opère aussi bien sur la variable cible, le nombre d'arrêts de travail, que sur la base contenant toutes les autres variables. Dans cette étude, 80% des données vont servir pour la base d'entrainement et les 20% restant pour la base de test. Ce processus de division est effectué à l'aide de la fonction « train_test_split() » de la librairie Scikit-Learn sous Python.

b) La technique des K-folds :

Comme pour la méthode présentée précédemment mais de manière plus robuste, la validation croisée par K-folds peut également résoudre les problèmes de sur et sous-apprentissage. Cette méthode est plus avantageuse car elle permet en plus un rééchantillonnage (« re-sampling » en anglais) des données. Il va permettre d'estimer les valeurs extrêmes ou rares avec une probabilité faible et les valeurs intermédiaires avec une probabilité plus importante. Ainsi, le sous-échantillon construit est plus représentatif et plus harmonieux. De plus, par rapport aux autres approches de « cross validation », elle aboutit souvent sur le modèle le moins biaisé.

L'algorithme de K-folds s'articule de la manière suivante. Tout d'abord, l'ensemble des données est séparé de manière aléatoire, en K groupes de taille similaire. Le paramètre « K » fait référence à un nombre de groupes selon lequel l'échantillon est divisé. La valeur K doit être ni trop basse, ni trop élevée.

Une valeur de K trop haute entraine un modèle moins biaisé. Cependant, la variance risque d'augmenter et ainsi conduire à un problème de surajustement. De manière générale, les valeurs de K sont comprises entre 5 et 10 en fonction de la taille des données initiales. Cette technique avec K égal à 2 revient à séparer la base selon la méthode *Train-Test Split*.

À chaque étape, un groupe est mis de côté. Ensuite, le modèle est entraîné en utilisant les K-1 autres folds. Ce modèle va ensuite être validé par le dernier fold non utilisé. Pour chaque itération, le score et les erreurs sont notés. Ce processus se répète jusqu'à ce que chacun des K folds apparaîssent à la fois au sein des données d'entrainements et de tests. Finalement, l'erreur globale du modèle est calculée en faisant la moyenne des erreurs de chaque itération.

Voici un exemple de cross validation avec la méthode de K-folds, lorsque K égal à cinq.

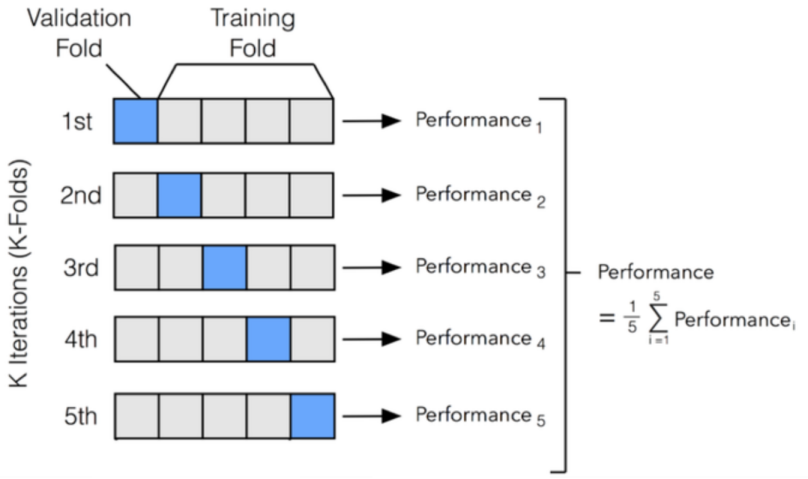


FIGURE 3.1 – Segmentation de la base de données pour la Croos-Validation
[1]

Ce processus est mis en place à l'aide des fonctions « *cross_val_scor* » et « *cross_val_predict* » de la bibliothèque « Scikit-Learn » sous Python.

3.2 Le Modèle Linéaire Généralisé, GLM

En tarification, de nombreux modèles prédictifs utilisent les modèles linéaires généralisés (GLM). En effet, les GLM ont gagné en popularité dans les applications d'assurance, grâce à leurs améliorations significatives par rapport aux modèles univariés.

3.2.1 Principe et fonctionnement des GLM

Les GLM sont une généralisation de la régression linéaire simple, dans laquelle une seule variable prédictive est considérée. Ce modèle a fait son apparition avec John Nelder et Robert Wedderburn en 1972, afin de répondre à la complexification des problèmes statistiques.

Le modèle GLM consiste donc à étudier l'éventuelle relation entre une variable à expliquer, souvent désignée par la lettre Y et p variables explicatives, généralement notées $X_1, \ldots, X_n$.

Les modèles GLM sont principalement utilisés dans deux situations :

1) Lorsque l'on cherche à prédire des données de type comptage (nombre d'essais marqués durant un match de rugby, nb d'arrêts observés, ...)

2) Lorsque l'on cherche à prédire des données de type binaire (Malade ou non, vivant ou mort, ...)

1. Source : https ://www.i2tutorials.com/machine-learning-tutorial/machine-learning-k-fold-cross-validation/

Dans notre étude, la réponse voulue est de type comptage. Les valeurs sont donc bornées inférieurement par 0 et ne comptent que des valeurs entières positives. Théoriquement, elles suivent donc une distribution de Poisson, avec un paramètre lambda inconnu. Par définition, l'espérance et la variance de la variable réponse doivent être égales au paramètre lambda de la distribution de Poisson.

Les modèles GLM se composent de trois éléments principaux :

- La composante aléatoire ou principale : Elle représente la variable ou le vecteur $(Y_1, \ldots, Y_n)$ de réponse, donc à expliquer. Dans notre étude, il s'agit d'une unique variable, notée Y. La densité de la variable aléatoire doit appartenir à la famille de loi exponentielle.

- La composante déterministe ou systématique :
  Elle désigne le prédicteur linéaire, noté η. Dans le cas de variables explicatives $X = (X_1, \ldots, X_n)$ et d'un vecteur à estimer $\beta = (\beta_1, \ldots, \beta_n)$, le prédicteur est un vecteur de taille *n*, de la forme suivante :
  $$\eta(X) = \sum_{i=1}^{n} X_i \times \beta_i.$$

- Une fonction de lien : Elle est souvent notée g. Elle doit être réelle, dérivable et strictement monotone. Elle relie la composante principale Y avec la composante systématique η, de la manière suivante :
  $$g_n(E[Y]) = \beta_0 + \beta_1 \times X_1 + \ldots . \beta_n \times X_n,$$
  ou alors :
  $$\eta(X) = g(\mu) \; avec \; \mu = E[Y].$$

De plus, la relation entre ces variables n'est pas exacte. Il existe un terme d'erreur à rajouter aux modèles, de sorte qu'une réalisation de la variable prédite soit la somme d'une partie observée (expliquée par les variables explicatives) et une partie inobservée (un bruit aléatoire), comme suit :

$$y_i = \beta_1 \times x_{1i} + \beta_2 \times x_{2i} + ... + \beta_n \times x_{ni} + \epsilon_i.$$

Le terme « ϵ » capte les insuffisances du modèle. Ces résidus se composent principalement de :

- L'écart entre la réalité observée et les prédictions du modèle GLM.

- L'ensemble des variables qui ne sont pas prises en compte dans le modèle.

- Les fluctuations liées à l'échantillonnage.

L'objectif premier des GLM est d'estimer les coefficients β_0 et $\beta_1, \ldots, \beta_n$ associés aux variables explicatives $X_1, \ldots, X_n$. En effet, le coefficient β_0 désigne la modalité de référence du modèle, tandis qu'un coefficient $\beta_(i)$, pour i allant de 1 à n, s'interprète de la manière suivante : il correspond à la variation estimée de *Y* lorsque la variable X_i augmente de 1 unité, toutes choses étant égales par ailleurs.

La démarche de la modélisation consiste à estimer les paramètres $\beta_1, \ldots, \beta_n$ par la méthode du maximum de vraisemblance, en considérant $y = (y_1, ..., y_n)$ comme étant une réalisation de l'échantillon de n variables aléatoires indépendantes, $(Y_1, ..., Y_n)$, dont les fonctions de densité f_{Y_i}, pour chaque i, sont issues de la famille exponentielle. La vraisemblance en y s'écrit donc :

$$\mathcal{L}(y; \beta, \phi) = \prod_{i=1}^{n} f(y_i; \omega_i, \phi) = \prod_{i=1}^{n} f(y_i; x_i\beta, \phi).$$

Avec $\omega_i = g(E(Y_i)) = x_i\beta$
Dans ces conditions, la log-vraisemblance s'écrit :

$$\ell(y; \beta, \phi) = \sum_{i=1}^{n} [\frac{y_i x_i \beta - b(x_i\beta)}{\gamma_i(\phi)} + c(y_i, \phi)].$$

La valeur maximale de $\ell(y; \beta, \phi)$ est obtenue en résolvant l'équation aux dérivées partielles suivante :

$\begin{cases} \frac{\partial \ell(y;\beta,\phi)}{\partial \beta_j} = 0, \\ \frac{\partial \ell(y;\beta,\phi)}{\partial \phi} = 0, \end{cases}$ avec j allant de 1 à n.

La résolution d'un tel système ne possède pas toujours de solution explicite. Néanmoins, en pratique, il existe différents algorithmes d'optimisation itératifs pour résoudre ce système et obtenir les estimations du maximum de vraisemblance. Les deux méthodes les plus couramment utilisées sont l'algorithme de *Newton-Raphson* et l'algorithme du score de *Fisher*.

Une fois que les coefficients sont estimés, la qualité d'ajustement du modèle peut être testée. Dans le cadre des GLM, elle est principalement mesurée à l'aide de la déviance du modèle ou de la statistique du χ^2 de Pearson.

La déviance mesure l'écart entre le modèle GLM ajusté et un modèle dit saturé. Le modèle saturé est défini comme le modèle parfaitement ajusté qui possède autant de paramètres que d'observations. En notant, *M* le modèle ajusté et *S* le modèle saturé, la déviance du modèle *M* s'écrit donc comme suit :

$$\mathcal{D}(M) = -2\ \Phi(log(\frac{\mathcal{L}_M}{\mathcal{L}_S})) = -2\ \Phi(\mathcal{L}(y, \hat{\beta}_M) - \mathcal{L}(y, \hat{\beta}_S)),$$

avec $\hat{\beta}_M$ l'estimateur du maximum de vraisemblance de β dans le modèle ajusté, noté *M* et $\hat{\beta}_S$ l'estimateur du maximum de vraisemblance de β dans le modèle saturé, noté *S*.

Asymptotiquement, $\mathcal{D}$ suit une loi de χ^2 à $n - p$ degrés de liberté, avec p le nombre de variables explicatives du modèle.

De plus, la statistique du χ^2 de Pearson est aussi employée pour comparer les écarts entre le modèle ajusté et les données observées. Elle se définit par :

$$\chi^2 = \sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{Var(\hat{\mu}_i)},$$

où $\hat{\mu}_i = g^{-1}(x_i\ \beta)$. L'écart observé n'est pas significatif au niveau α si la valeur observée de la statistique χ^2 est supérieure au quantile $\chi^2_{n-p,1-\alpha}$.

Enfin, pour évaluer le pouvoir explicatif global du modèle et la pertinence des variables explicatives

dans la prédiction, des tests de significativité sont mis en place.

3.2.2 Hypothèses à vérifier

Les modèles sont relativement simples à implémenter. Néanmoins, quelques hypothèses doivent être vérifiées au préalable.

Premièrement, certaines hypothèses concernent le modèle dans sa globalité.

a) Les variables explicatives, notées X, doivent être observées sans erreur et ne contenir aucune valeur aléatoire.

C'est le cas dans notre jeu de données. L'intégralité de l'information provient de la DSN. Les données sont donc précises et majoritairement correctes. De plus, grâce aux retraitements réalisés au préalable, les variables utilisées ne comportent pas d'erreur.

b) Absence de colinéarité parfaite :

Il ne doit pas avoir de relation linéaire parfaite entre les variables explicatives du modèle. Dans ce portefeuille, certaines catégories de différentes variables sont légèrement interdépendantes. Cependant, aucune colinéarité parfaite n'est constatée dans cette base.

c) Nombre d'observations supérieur au nombre de paramètres :

Dans les deux bases de données, plusieurs millions d'observations sont contenus. Ensuite, suivant le nombre de variables inclue dans le modèle, le nombre de paramètres varie. En revanche, dans tous les modèles réalisés, le nombre de paramètres oscille entre 7 et 31. L'hypothèse est donc largement vérifiée.

Deuxièmement, des hypothèses impliquant les résidus sont à vérifier :

a) Normalité des résidus :

Pour vérifier que les résidus suivent une loi normale, le test de Shapiro-Wilk est nécessaire. Il est défini selon l'hypothèse suivante : H_0 : les résidus sont distribués selon une loi normale

On rejette l'hypothèse H_0 si la p-value est inférieure à 5%.

A l'aide de la fonction « shapiro() », de la bibliothèque « scipy.stats » en Python on trouve :

Statistique de Test : 0.9866

p-value : 0.8821

La p-value est donc supérieure à 0.05, l'hypothèse nulle de ce test ne peut pas être rejetée. Les résidus des modèles GLM suivent donc une loi normale.

b) En moyenne, le modèle est bien spécifié :

Cela veut dire que l'espérance des résidus du modèle est en moyenne nulle, soit que $E(\epsilon_i) = 0$.

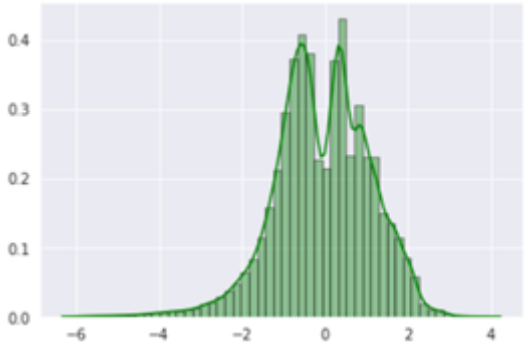


FIGURE 3.2 – Histogramme de la répartition des résidus

Malgré une proportion plus faible en 0, l'histogramme ci-dessus prend ses valeurs dans un périmètre proche de 0. C'est grâce au calcul de la moyenne des résidus que ce test peut être validé. En effet, les résidus sont globalement centrés puisque : $E(\epsilon_i) = -0.056$.

c) Homoscédasticité des erreurs :

L'homoscédasticité correspond au fait que la variance des erreurs du modèle est constante : $V(\epsilon_i) = \sigma^2$, pour tout i, avec σ une constante réelle.

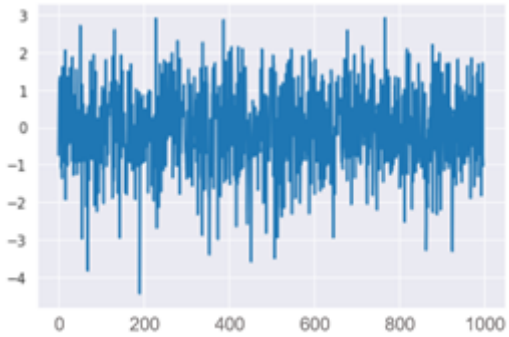


FIGURE 3.3 – Répartition des résidus

d) Absence d'autocorrélation des erreurs :

Le modèle comporte des erreurs corrélées lorsque la covariance entre deux termes d'erreur est non nulle. En effet, l'absence de corrélation des erreurs est définie dans le cas suivant : $E(u_i; u_j) = 0$, pour tout $i \neq j$.

Pour tester cette hypothèse, le test de Durbin & Watson est adopté. Ce test permet de détecter une autocorrélation des résidus d'ordre 1, selon la forme suivante :

$$\epsilon_t = \rho \times \epsilon_{t-1} + \gamma_t,$$

où $\gamma_t = BB(0, \sigma^2)$

L'hypothèse nulle du test est : $H_0 : \rho = 0$

La statistique DW est la suivante :

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2},$$

où les e_i sont les résidus du modèle.

Les valeurs du test de Durbin & Watson varient entre 0 et 4 :

- Valeur proche de 2 : indique de la non-autocorrélation
- Valeur proche de 0 : indique une corrélation positive
- Valeur proche de 4 : indique une corrélation négative

A l'aide de la fonction « durbin_watson() » de la librairie statsmodels de Python, la valeur de la statistique de DW est égale à 1.9888. L'hypothèse est donc vérifiée.

e) Exogénéité du modèle :

Le modèle GLM est exogène si les erreurs sont indépendantes des variables explicatives : $Cov(u_i; X_i) = E(u_i x X_i) = 0$, pour tout i.

Une fois que toutes ces hypothèses sont vérifiées, il faut analyser la qualité globale du modèle. Dans un premier temps, la statistique de Pearson est employée pour mesurer la qualité d'ajustement du modèle, définie comme suit :

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i \times (1 - \hat{\mu}_i)} \sim X^2_{n-p-1},$$

où y_i une réalisation de la variable cible et $\hat{\mu}_i$ la moyenne de la variable cible.

X_1^2 : Loi de Khi-Deux de paramètre $n - p - 1$.

Dans un second temps, pour savoir si les variables ont une importance dans notre modèle, il faut tester la significativité des coefficients. Pour cela le test de Wald est adopté. Cela revient à tester l'hypothèse $H_0 : \beta_i = 0$ et ce pour tout i = $1...q$.

Le test s'appuie sur la statistique de test suivante :

$$W = \frac{\hat{\beta}_i^2}{Var(\hat{\beta}_i)} \sim X_1^2.$$

Si la p-value est inférieure à 5%, l'hypothèse nulle est rejetée.

3.2.3 Rappel sur les lois possibles et les fonctions de liens

La variable réponse, notée Y, doit appartenir à la famille exponentielle. La fonction de densité de cette distribution doit dépendre des paramètres θ et ϕ comme indiqué ci-dessous :

$$f(y|\theta,\phi) = exp(\frac{y*\theta - b(\theta)}{a(\phi)} + c(y,\phi)).$$

Densité qui se compose des éléments suivants :

- a(.) : une fonction de $\mathbb{R}\to\mathbb{R}^*$
- b(.) : une fonction de $\mathbb{R}\to\mathbb{R}$, de classe C^2 inversible
- c(.) : une fonction de $\mathbb{R}^2\to\mathbb{R}$
- θ : un paramètre réel naturel
- ϕ : un paramètre réel de dispersion
- y : la variable représentant la cible

Il s'ensuit quelques propriétés qui découlent de l'appartenance à cette famille. Si Y appartient à cette famille exponentielle, alors :

$$\mu = E(Y) = b^{'}(\theta),$$

$$Var(Y) = a(\phi) * b^{''}(\theta) = a(\phi) * V(\mu),$$

où $V(\mu)$ est appelée fonction variance.

Selon la nature de la variable réponse, différentes fonctions de lien peuvent être introduites dans les modèles. Quelques exemples de lois appartenant à la famille exponentielle sont donnés dans le tableau ci-dessous.

Loi de Y	Notation	θ	b(θ)	a(Φ)	E[Y]	$V(\mu) = \frac{Var(Y)}{a(\Phi)}$
Binomiale	$\mathcal{B}(n,p)$	ln($\frac{p}{1-p}$)	$n$ln$(1+e^{\theta})$	1	np	$np(1-p)$
Poisson	$\mathcal{P}(\lambda)$	ln(λ)	e^{θ}	1	λ	λ
Normale	$\mathcal{N}(m,\sigma^2)$	m	$\frac{\theta^2}{2}$	σ^2	m	1
Exponnentielle	$\mathcal{E}(\lambda)$	$\frac{1}{\lambda}$	ln(θ)	1	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$\mathcal{G}(\alpha,\beta)$	$\frac{1}{\alpha\beta}$	ln(θ)	- $\frac{1}{\alpha}$	$\frac{1}{\alpha\beta}$	$\frac{1}{\alpha\beta^2}$

- La loi Binomiale est utilisée pour expliquer des variables binaires.
- La loi Normale est majoritairement employée pour prédire une réponse quantitative continue, qui prend ses valeurs dans l'ensemble des réels.

- La loi de Poisson convient parfaitement pour modéliser un processus de comptage ou une fréquence.
- La loi Gamma est très performante pour modéliser des valeurs réelles strictement positives.
- La loi Inverse Gaussienne convient pour les mêmes modèles que la loi Gamma.
- La loi Binomiale Négative, comme la loi de Poisson, est adaptée pour les modèles de comptage. La particularité de cette loi est qu'elle permet généralement de résoudre un problème de sur-dispersion.

3.2.4 Avantages et Inconvénients

Les points forts de ce modèle sont les suivants :

- Les GLM permettent de modéliser des réponses diverses : $\in \mathbb{R}, \in \mathbb{R}^+, \in \mathbb{N}, \in [0;1]$, etc.
- Les hypothèses à vérifier et les paramètres sont simples et peu nombreux.
- Le modèle est facilement interprétable. Il capte d'éventuelles relations linéaires entre une variable cible et des variables explicatives. Il en résulte également un unique coefficient par variable, qui permet de quantifier l'impact de chaque variable sur les prédictions.
- De nombreux tests sur l'ajustement du modèle (avec par exemple la qualité globale du modèle ou la significativité de coefficients) et sur les résidus peuvent être réalisés.
- Les GLM permettent d'ajouter et d'analyser les effets de variables explicatives mélangées entre elles. En effet, pour apporter une analyse plus fine ou pour capter des effets non linéaires, il est possible de croiser deux ou plusieurs variables. Par exemple, en croisant les variables âge et genre, il est possible d'étudier l'effet de l'âge sur la variable cible séparément en fonction du sexe de l'individu. Le premier coefficient rattaché à la variable âge indique l'influence de la modalité de référence, prenons par exemple le genre féminin. La somme du premier et du second coefficient indique désormais l'influence des hommes sur la variable à expliquer.

En revanche, les GLM présentent quelques limites :

- La normalité n'est souvent pas vérifiée dans la réalité.
- Il peut y avoir des problèmes de convergence et donc un risque de modèle non négligeable.
- Le modèle devient compliqué à utiliser si le nombre de variables et de modalités devient trop important. En effet, il est préférable de dissocier les variables catégorielles en plusieurs variables binaires. Chaque variable dichotomique est associée à un coefficient. Le nombre de coefficients à estimer peut donc augmenter de manière exponentielle.
- Le modèle ne permet pas de tester les effets de seuils et la dépendance entre les variables explicatives.

3.3 Le modèle Extreme Gradient Boosting, XGBoost

L'algorithme « Extreme Gradient Boosting » se présente aujourd'hui comme l'un des modèles de Machine Learning les plus performants. Toujours dans l'objectif de challenger l'outil Akur8, le choix de cet algorithme semble naturel.

Grâce à une étude menée sur différents modèles de Machine Learning tel que l'algorithme CART ou Random Forest, le modèle XGBoost est le plus concluant. De plus, ce modèle est principalement connu pour ses fortes performances lors des concours Kaggle.

L'algorithme XGBoost se base sur l'algorithme de Gradient Boosting, qui lui-même repose sur le principe du boosting.

3.3.1 Principe du Boosting

Le boosting fait partie des modèles d'apprentissage d'ensemble, qui consiste à former plusieurs modèles utilisant le même algorithme d'apprentissage. C'est une technique d'intelligence artificielle qui permet d'assembler un grand nombre d'apprenants faibles pour former un apprenant fort, beaucoup plus efficace.

- Les apprenants ou classifieurs faibles, *« weak learner »* en anglais, sont des algorithmes avec de faibles performances individuelles. Ils ont donc un pouvoir de prédiction contestable. Ces modèles sont souvent victimes de sur-apprentissage. Ils ne peuvent pas gérer ni classer des données si celles-ci varient trop par rapport aux données d'entrainement.

- Les apprenants ou classifieurs forts, *« strong learner »* an anglais, ont donc des performances meilleures et une précision plus élevée pour les prédictions.

L'idée derrière le principe du boosting est donc que plusieurs petits algorithmes peuvent être plus performants qu'un seul gros. C'est pourquoi plusieurs « weak learner » sont construits les uns après les autres, pour former un « strong learner ». Dans la suite de ce mémoire, le principe du Boosting est employé avec des algorithmes d'arbre de décision.

En Machine Learning, les arbres de décisions sont des structures de données qui reposent sur la segmentation de l'ensemble de données initial, la « *racine* » de l'arbre. À chaque « *nœuds* » de l'arbre une question est posée. En fonction de la réponse, la population est scindée en deux ou plusieurs sous-groupes. De manière itérative, la population est divisée jusqu'à ce qu'il ne reste plus que des classes homogènes, appelées « *feuilles* » .

Voici un exemple d'arbre de décision pour prédire si le salarié sera victime ou non d'un arrêt de travail, sur un sous-échantillon de notre base de données.

FIGURE 3.4 – Exemple d'arbre de décision

Le premier nœud du haut segmente la base de données selon la question : « le salarié est-il ouvrier ? ». S'il n'est pas ouvrier, la segmentation s'arrête, une première feuille de l'arbre est consitutée, avec la prédiction « no arrêt ». Ensuite, si le salarié est ouvrier, le nœud suivant divise la base en fonction de l'âge (plus ou moins âgé de 26 ans) et ainsi de suite jusqu'aux feuilles finales.

3.3.2 Fonctionnement du Boosting

Le Boosting repose donc sur l'idée de construire plusieurs apprenants faibles pour former un apprenant fort. Dans un premier temps, des poids égaux sont attribués à toutes les observations. Le processus de boosting crée alors un modèle de base, le premier « weak learner ». Ce modèle est ensuite entrainé sur les données. Il ne crée que des arbres simples, c'est-à-dire avec un unique nœud et donc une seule division des données. À partir des résultats obtenus, il effectue ensuite des prédictions pour chaque échantillon de données.

Ensuite, l'algorithme évalue les bonnes et mauvaises prédictions du modèle. Les futurs apprenants faibles construits vont venir corriger les erreurs des précédents. C'est pourquoi, pour les observations mal classées ou les erreurs d'estimation, le poids qui leur est attribué augmente fortement. Le poids affecté aux bonnes prédictions peut éventuellement diminuer. La pondération des données prédites va alors servir pour la construction de l'arbre suivant.

Un second modèle est alors construit pour tenter de corriger les erreurs présentes dans le modèle initial. Il est donc entrainé avec les données pondérées obtenues lors de la première étape. L'objectif de cette étape est donc de maximiser le nombre de bonnes prédictions pour les observations avec un poids élevé. Ce processus est répété au cours de chaque étape de création d'un nouvel arbre.

Le schéma suivant synthétise l'algorithme sur la création de trois arbres.

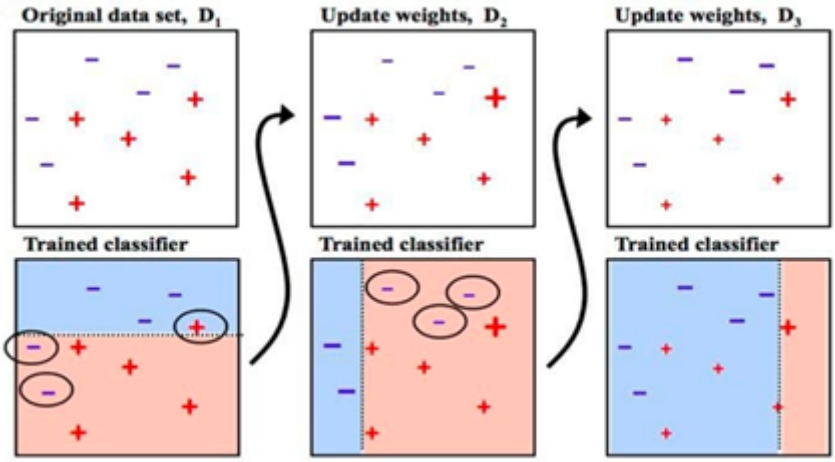


FIGURE 3.5 – Schéma du processus de Boosting
2

Les prédictions du dernier modèle seront donc les prédictions globales pondérées, fournies par les anciens arbres de décision. Le processus s'arrête lorsque les erreurs d'entrainements deviennent inférieures à un seuil choisi au préalable.

Le boosting attribue également des poids plus importants pour les modèles qui ont d'excellentes prédictions. Ils vont donc avoir une forte influence dans la décision finale.

3.3.3 Le Gradient Boosting

Le principe du boosting peut se modéliser avec différents algorithmes. Il en existe deux principaux qui sont le Boosting adaptatif (AdaBoost) et le Boosting de gradient. Dans la suite de ce mémoire, l'algorithme de descente du gradient (« Gradient Boosting machine ») est retenu.

L'algorithme Gradient Boosting a beaucoup de points communs avec le modèle classique Adaboost, inventé par Freund et Schapir en 1996. Effectivement, les deux processus se basent sur la construction d'un ensemble de « weak learner », créés les uns après les autres formant un « strong learner » final. Néanmoins, contrairement à Adaboost, les classifieurs faibles du Gradient Boosting ont tous des poids égaux, quelle que soit leur performance. La particularité de Gradient Boosting est qu'il essaye de prédire à chaque étape non pas les données elles-mêmes mais les résidus.

Par conséquent, le Gradient Boosting se base sur le principe de minimisation de la fonction de perte de l'algorithme de descente du gradient. Les erreurs ou résidus du modèle correspondent à l'écart entre la variable observée, notée y, et la valeur prédite, notée $\hat{y}$, en fonction des différentes variables explicatives $X = (X_1, \ldots, X_n)$.

Dans cette étude, les résidus sont représentés par la fonction de perte la plus couramment utilisée, notée L. Elle correspond à l'erreur quadratique moyenne (MSE) et se définit comme suit :

$$L(y, \hat{y}) = MSE = \frac{1}{n} * \sum (y - \hat{y})^2.$$

2. Source : https://iq.opengenus.org/gradient-boosting/

À chaque étape, l'objectif est donc de minimiser cette fonction L, c'est-à-dire de rechercher un nouveau classifieur $\tilde{f}$ tel que :

$$\tilde{f} = arg\min_{y} E[L(y, \hat{y})].$$

Le premier « weak learner » $\tilde{f_0}$ du processus est simplement la moyenne des observations. Ce premier arbre est donc très peu efficace. La fonction de perte à minimiser s'écrit donc :

$$\tilde{f_1} = arg\min_{y} \sum_{i=1}^{n} L(y_i, \tilde{f_0}).$$

Le nouvel apprenant faible $\tilde{f_1}$ comprend donc $\tilde{f_0}$ plus un terme provenant des résidus, représenté par la fonction g. Il est donc plus performant :

$$\tilde{f_1}(x) = \tilde{f_0}(x) + g(x), \forall x.$$

Ensuite, au cours des étapes suivantes, les apprenants faibles sont entrainés pour prédire au mieux les écarts précédents. L'objectif est de corriger uniquement les prédictions qui ont été mal prédites. L'arbre de décision suivant va donc améliorer les prédictions pour lesquelles l'erreur est élevée sans détériorer celles pour qui l'erreur est faible.

En effet, au lieu de prédire y, le second arbre va prédire $y - \hat{y}$. En faisant la somme de la prédiction du second arbre et du premier on devrait obtenir y. On répète ainsi cette opération de manière itérative, jusqu'à ce que l'erreur soit suffisamment faible.

Pour chaque étape $k \in 1, \ldots, K$

$$\tilde{f_k} = \tilde{f_{k-1}} + arg\min_{y} \sum_{i=1}^{n} L(y_i, \tilde{f_{i-1}}).$$

3.3.4 Le modèle XGBoost

Finalement, c'est le modèle Extreme Gradient Boosting qui est retenu lors de nos modélisations. Cet algorithme est conçu comme une implémentation optimisée de Gradient Boosting. Lui aussi se base sur un assemblage de classifieurs faibles qui prédisent les résidus. Néanmoins, sa particularité provient des « weak learner ». Ce sont des arbres décisionnels élagués. En effet, les arbres qui ne sont pas assez bons sont élagués. Certaines branches sont coupées jusqu'à ce qu'ils soient assez performants. S'ils ne prédisent pas assez juste après élagage, ces arbres sont supprimés du modèle. Cette méthode s'appelle le « pruning ».

Ainsi, XGBoost s'assure de ne conserver que des bons apprenants faibles. Il est optimisé dans son implémentation sous-jacente. Afin d'être le plus performant, il se focalise sur le temps d'exécution, la flexibilité et les performances du modèle. En effet, il va rajouter des tirages aléatoires sur les observations et les covariables, dans le but de décorréler les arbres entre eux.

Il y a également d'autres différences avec le modèle de Gradient Boosting :

– Le modèle XGBoost fonctionne selon le principe de parallélisation des tâches. Ceci est possible grâce à la nature permutable des boucles utilisées pour la construction des apprenants faibles. Il peut donc exploiter la puissance des ordinateurs multicœurs. Les performances augmentent

puisque toute surcharge dans le calcul est compensée par la parallélisation des tâches. Il est alors possible d'entrainer les modèles sur de très grands ensembles de données.

– Il dispose d'un large panel d'hyperparamètres. En tant qu'utilisateur, cela nous permet donc d'avoir un meilleur contrôle sur l'implémentation. En effet, il est possible de réduire au maximum le taux d'erreur en faisant varier les paramètres. Néanmoins, le tunning de tous ces paramètres peut être long et fastidieux.

– Le modèle du Gradient Boosting utilise une fonction de perte pour évaluer la qualité des prédictions. En revanche, le modèle XGBoost introduit en plus une fonction de régularisation. Elle permet de contrôler la robustesse et de pénaliser la complexité du modèle. L'objectif de XGBoost est donc de minimiser une fonction qui est la somme de la fonction de perte et de la fonction de régularisation.

– Le modèle XGB prend en compte les valeurs manquantes dans les bases de données.

Comme pour les autres modèles de Machine Learning basés sur des arbres de décisions, XGBoost est sensible au problème de sur-apprentissage. Cependant, il peut être corrigé en limitant la profondeur des arbres (diminuer le paramètre *max_depth*) ou en réalisant les modèles sur des échantillons restreints de l'ensemble des données, à l'aide de méthodes de cross-validation par exemple.

3.3.5 Hyperparamétrisation et optimisation : Grid Search

Un hyperparamètre est un type de paramètre, en général externe au modèle, défini avant le début du processus d'apprentissage. Pour optimiser les performances du modèle, il est nécessaire de régler ces hyperparamètres, grâce à un GridSearch par exemple.

XGBoost se caractérise par ses nombreux hyperparamètres qui augmentent la complexité du modèle. Tous ne sont pas indispensables et ne méritent pas un traitement particulier. Néanmoins, une liste non exhaustive des paramètres les plus courants, qui sont en général modifiés, est présentée ci-dessous :

– *max_depth* : il définit la profondeur maximale de chaque arbre de décision. La valeur par défaut est 6. Il est important de faire varier cette variable pour augmenter les performances du modèle ou résoudre un problème de sur-apprentissage.

– *learning_rate* : c'est le taux d'apprentissage du modèle. C'est un facteur de pondération qui réduit le poids des caractéristiques à chaque étape de boosting. Il doit être compris entre 0 et 1 et sa valeur par défaut est 0,3. En général, une valeur proche de 0,01 est utilisée pour débuter, puis ajuster au besoin. Augmenter le taux d'apprentissage permet d'accélérer la vitesse de calcul. Si le modèle est plus rapide sans perdre en performance, il est possible d'augmenter le nombre d'estimateurs pour améliorer sa qualité.

– *n_estimators* : il correspond au nombre d'arbres dans notre ensemble. Cela équivaut au nombre d'étapes du boosting. La valeur doit être entière et positive. La valeur par défaut est 100. En général, plus l'ensemble de données est volumineux, plus le nombre d'arbres a besoin d'être grand.

– *Colsample_bytree* : il représente le pourcentage de colonnes à échantillonner aléatoirement pour chacun des arbres. Ce paramètre permet de réduire les problèmes de sur-ajustement.

- *subsample* : ce paramètre représente la fraction d'observations à segmenter pour chacun des arbres. La valeur doit être comprise entre 0 et 1. La valeur par défaut est 1. Des valeurs trop faibles empêchent le sur-apprentissage mais peuvent entrainer un biais trop fort. Un sous-échantillon à 0,5 signifie que 50% des données d'apprentissage sont utilisées pour construire les arbres de décision.
- *Alpha* : c'est le paramètre de régularisation L1 sur les poids (Lasso Regression). Sa valeur par défaut est 0 mais elle peut prendre n'importe quel entier. Augmenter sa valeur rend le modèle plus conservateur. Cela permet principalement d'augmenter les performances de vitesse du modèle.
- *Lambda* : c'est le paramètre de régularisation L2 sur les poids (Ridge Regression). Sa valeur par défaut est 1 mais elle peut être égale à n'importe quel entier. Augmenter sa valeur permet principalement de réduire le sur-ajustement.
- *Gamma* : c'est un autre paramètre de régularisation pour l'élagage des arbres. Il indique la réduction de perte minimale nécessaire pour construire un arbre supplémentaire. Par défaut, Gamma est égal à 0. Il peut prendre sa valeur dans tous les entiers.
- *min_sample_leaf* : le nombre minimum de caractères qu'une feuille terminale doit avoir.
- *min_sample_split* : le nombre minimum de caractères qu'un nœud doit avoir pour être divisé.
- *max_features* : le nombre de variables à prendre en compte pour construire chaque arbre. En général, il est égal à la racine carrée du nombre de colonnes de notre base.
- *eval_metric* : ce paramètre spécifie quelle fonction de perte est retenue : MSE, MAE, RMSE ou log loss pour les modèles de classification.

Par la suite, pour simplifier et réduire le temps de tunning des paramètres, une recherche de grille, grid search en anglais, est mise en place. Pour chaque combinaison des paramètres spécifiés en entrée, cette méthode détermine les paramètres à retenir pour obtenir le modèle le plus performant.

En outre, comme notre base de données est volumineuse et que l'optimisation des paramètres prend beaucoup de temps, il est nécessaire de segmenter la recherche des paramètres optimaux. D'abord, l'optimisation des paramètres principaux, liés aux arbres de décision : *max_depth* et *n_estrimators*. Par exemple, si dans le grid search, cinq valeurs différentes sont définies pour chacun de ces deux paramètres, au total ce n'est pas moins de vingt-cinq modèles distincts qui sont testés, en combinant les différents paramètres.

Puis, les autres paramètres généraux sont étudiés (*subsample*, *colsample_bytree*, *learning_rate*, *min_sample_leaf*, *min_sample_split*, *max_feature*).

Enfin, c'est au tour des paramètres de régularisation d'être réglés. En général, il faut commencer par l'optimisation d'*Alpha*, puis *Lambda* et pour finir *Gamma*.

3.3.6 Avantages et Inconvénients

Les forces de ce modèle sont les suivantes :

- Flexibilité : XGBoost peut effectuer différentes tâches d'apprentissages telles que la régression, la classification ou encore le classement. Il peut également s'appuyer sur différentes fonctions de perte (RMSE, MSE, MAE ou encore log loss).
- Facilité d'implémentation : XGBoost est un algorithme connu et développé dans plusieurs langages de programmation.
- Réduction du biais : le biais correspond à l'inexactitude des résultats. Les algorithmes de boosting sont construits de sorte qu'ils apprennent de leurs erreurs et s'améliorent au fur et à mesure des itérations. Cela permet donc de réduire considérablement les erreurs de prédictions. C'est donc un modèle performant et robuste. De plus, il intègre directement un processus de validation croisée.
- Il est capable de capter des effets non linéaires entre les covariables.

Toutefois, XGBoost présente quelques limites :

- Nombre conséquent de paramètres : pouvoir choisir différents paramètres peut-être un avantage, puisque l'utilisateur semble avoir le contrôle des modèles qu'il construit. Cependant, cela entraine une forte augmentation de la complexité. De plus, analyser et optimiser chacun des paramètres rallonge les temps de calculs de manière exponentielle.
- Interprétabilité du modèle plus compliquée : l'algorithme du boosting est considéré comme une « boîte noire ». Même si le fonctionnement général est relativement clair à comprendre, cela se complexifie lorsqu'il faut chercher d'où provient une erreur ou simplement pour interpréter les résultats en sortie de modèle.
- Les algorithmes nécessitent des ordinateurs puissants pour les calculs.
- Sensible au sur-apprentissage : malgré son processus de validation croisée intégré, XGBoost présente un problème récurrent. Il reste très sensible au sur-apprentissage.

3.4 Le modèle outil, Akur8

Enfin nous allons présenter un logiciel de tarification actuarielle, nommé Akur8. Cet outil propose aux utilisateurs une combinaison entre la transparence et l'interprétabilité des résultats des modèles classiques de tarifications (GLM, ...) et l'automatisation des algorithmes de Machine Learning (CART, XGBoost, ...).

Akur8 permet de rationnaliser le cycle entier de tarification, en partant de la modélisation actuarielle jusqu'à la production. En effet, la refonte des normes tarifaires a un coût élevé. Akur8 propose donc une solution flexible et facile à prendre en main. De plus, avoir un outil spécifique pour la création de norme permet de réduire le risque opérationnel.

C'est donc un énorme gain de temps d'utiliser une solution aussi visuelle et ergonomique qu'Akur8. Le temps économisé peut donc être mis à profit pour analyser les résultats ou travailler sur d'autres sujets.

3.4.1 Présentation des Modèles Additifs Généralisés (GAM)

Plus précisément, la solution Akur8 se base sur des modèles additifs généralisés, en anglais GAM : Generalized Additive Model. À l'origine, les GAM ont été développés en 1990 par T. Hastie et R. Tibshirani dans l'objectif de mélanger des modèles linéaires et des modèles additifs.

À la différence des GLM, les modèles GAM intègrent de l'additivité dans la composante déterministe. Ils permettent donc de décrire une relation non-linéaire entre des prédicteurs et la variable réponse. Cela veut dire que les variables explicatives qui définissent la composante déterministe sont désormais représentées dans le modèle par des fonctions de lissage inconnues. En effet, le modèle s'écrit désormais :

$$g(E[Y]) = \beta_0 + \sum_{j=1}^{p} f_j(X_j),$$

où :

- g : la fonction de lien utilisée ;
- E[Y] = $\hat{y}$: le vecteur des prédictions ;
- X_j : le vecteur prédictif de la variable j, avec j allant de 1 à p. p désigne le nombre de variables dans notre modèle ;
- f_j (avec j allant de 1 à p) : la fonction de lissage de la variable explicative X_j. Les $\beta_j * x_j$ d'une régression linéaire sont remplacés par les fonctions $f_j(X_j)$. Ce sont des fonctions pas à pas constantes par morceaux. Elles permettent de prendre en compte tout type de forme non paramétrique. Ces fonctions lisses sont donc inconnues.

Étant donné que la fonction de lissage $f(X_i)$ est locale et pas spécifiquement linéaire, l'ampleur de variable explicative X_i peut varier en fonction de sa relation avec la variable réponse. Autrement dit, contrairement à un coefficient fixe β_i associé à la variable X_i, la fonction f peut changer tout au long du gradient x_i.

Désormais, la variable cible Y s'écrit selon le modèle suivant :

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon,$$

où $(\epsilon_t)_t$ est un processus indépendant de loi normale centrée représentant l'erreur du modèle.

L'objectif des GAM est d'estimer les fonctions f. Selon la famille d'appartenance (fonctions polynomiales, lisseurs de régression linéaire locale, splines cubiques de lissage, fonctions trigonométriques, etc) de ces fonctions, plusieurs méthodes d'estimation sont possibles. La plus courante reste l'algorithme d'ajustement arrière, backfitting algorithm en anglais. Il se défini de la manière suivante :

De l'écriture précédente de Y, on en déduit :

$$Y - \beta_0 - \sum_{k=1}^{p} f_{k\neq j}(X_k) = f_j(X_j) + \epsilon.$$

En passant à l'espérance conditionnelle sachant X_j on trouve :

$$E[Y - \beta_0 - \sum_{k=1}^{p} f_{k\neq j}(X_k)|X_j] = f_j(X_j).$$

Cela signifie donc que la valeur de la fonction de lissage associée à la covariable X_j peut s'exprimer en fonction des autres variables explicatives. L'algorithme tire son nom du fait qu'à chaque étape, la valeur prise par la fonction associée à une variable est calculée à partir des valeurs prises par les fonctions des autres variables à l'étape précédente.

L'algorithme de backfitting résout donc ces p équations de manière itérative :

1) Initialisation des valeurs des fonctions de lissage, tel que :

$$\beta_0 = \frac{1}{n} \times \sum_{j=1}^{n} y_i.$$

C'est-à-dire que la valeur initiale β_0 est égale à la moyenne des observations de la variable réponse Y. Par conséquent, pour chacun des lisseurs, une valeur de départ $f_1^0, f_2^0, \ldots, f_p^0$ est déterminée.

2) Définition de f_j^1, pour j allant de 1 à p tel que :

$$f_j^1 = E_j[Y - \beta_0 - \sum_{k=1}^{p} f_{k\neq j}^0(X_k)|X_j].$$

Durant cette étape, le calcul de f_1 avec un seuil égal 1 est effectué à partir des f_j au seuil 0.

3) Ce processus est répété pour obtenir les fonctions $f_2^1, f_3^1, \ldots, f_p^1$.

4) Tant que les fonctions f_j^k ne convergent pas, l'itération se poursuit pour k = 1,. . .,p.

Il se peut que les fonctions ne convergent jamais. Au préalable, il faut donc définir un seuil de lissage, au-delà duquel l'algorithme s'arrête.

L'algorithme s'arrête donc lorsque les fonctions f_j convergent ou lorsque le seuil de lissage est atteint. Selon la méthode d'ajustement choisie (en général le maximum de vraisemblance), le degré de lissage est estimé à l'aide d'une validation croisée ou d'une régression pénalisée.

3.4.2 Caractéristiques des modèles sous Akur8

Les modèles GAM présentés sous Akur8 sont plus difficiles à ajuster. En effet, pour chacune des variables, un nombre conséquent de paramètres peut être estimé. À contrario, les GLM estiment un unique coefficient par variable. C'est ce qui explique que seul des effets linéaires sont capturés.

Néanmoins, les deux modèles GLM et GAM sont construits selon le même principe de maximisation de la log vraisemblance :

$$\hat{\beta} = arg\max_{\beta} p(y|\hat{y}(X)) = arg\max_{\beta} LogLikelihood(x, y, \beta),$$

avec $p(y|\hat{y}(X))$ qui suit la distribution choisie pour le modèle considéré. L'objectif premier et donc d'estimer les coefficients β_i. Par conséquent, le résultat final est de la forme :

$$\hat{Y}(X) = Offset \times \beta_1 \times \beta_2 \times ... \times \beta_n,$$

où n est le nombre de coefficients.

Or, la particularité d'Akur8 est qu'il permet un contrôle du sur-apprentissage dans le processus d'ajustement. En effet, les coefficients $\hat{\beta}$ ne sont pas directement estimés. Des hypothèses préalables sont insérées dans le but de limiter le surajustement des coefficients $\beta_{i,j}$ adaptés à un GLM dit « naïf ». Ces hypothèses peuvent concerner les transformations (polynomiales ou groupements de modalités par exemple) réalisées. Néanmoins, les hypothèses les plus correctrices interviennent sur l'ajustement du modèle en s'appuyant sur la crédibilité du cadre bayésien. Cela permet d'assurer la cohérence entre les différents coefficients créés.

En intégrant directement ces hypothèses, l'approche par maximum de vraisemblance devient :

$$\hat{\beta} = \underset{\beta}{arg\max}\ LogLikelihood(x, y, \beta) + log(p_{hypothse}(\beta)),$$

où $log(p_{hypothse}(\beta))$ intègre les hypothèses préalables.

Parmi les hypothèses intégrées dans le modèle, il y a notamment :

— Les coefficients suivent une loi gaussienne, centrée en 0 : Ridge Régression ;

— Les coefficients suivent une distribution de Laplace, centrée en 0 : Lasso régression ;

Les régressions de Lasso sont populaires car ce sont de bons outils pour la sélection de variables. Cependant, bien qu'il soit très puissant, le modèle Lasso n'est pas directement applicable. Des hypothèses supplémentaires sur les coefficients doivent être introduites.

C'est pourquoi deux coefficients consécutifs :

– Sont plus susceptibles d'être proches que très éloignés s'ils sont significativement différents.

– Ont les mêmes coefficients s'ils ne sont pas significativement différents.

Pour vérifier cette hypothèse, les fonctions $f_j(.)$ sont supposées constantes par paliers. L'hypothèse de nullité *H0* se définit ainsi :

$$\beta_{i,j} = \beta_{i+1,j}$$

Cette hypothèse est comparée aux données disponibles via la statistique de test Khi-Deux. Si l'hypothèse est rejetée alors $\beta_{i,j} \neq \beta_{i+1,j}$ et les paliers correspondant ne seront pas égaux. La fonction f_j ne sera donc pas constante en tous points. En revanche, si l'hypothèse est vérifiée alors $\beta_{i,j} = \beta_{i+1,j}$.

Cette méthode d'estimation n'est pas la plus simple. Cependant, elle génère des groupements de valeurs telles ques les f_j acquièrent une structure de fonctions constantes par morceaux.

Ainsi, un seuil de rejet faible, relatif à de fortes hypothèses et à une faible variance, conduira à des modèles bruyants. En effet, cela entraîne un grand nombre de coefficients distincts et donc de nombreux paliers pour les fonctions f_j. Néanmoins, ce modèle sera sujet au sur-apprentissage puisque les coefficients estimés sont trop proches des observations fournies en entrée.

À l'inverse, un seuil de rejet élevé, correspond à des hypothèses faibles et à une forte variance. Il aboutit à de larges classes de coefficients égaux et donc à un nombre réduit de paliers. Les modèles ainsi créés seront plus robustes mais sans doute moins performants. [3]

Par conséquent, Akur8 va donc proposer différents modèles. L'utilisateur pourra ensuite choisir celui qu'il préfère en fonction des contraintes dont il dispose.

3.4.3 Conception de modèles à l'aide de l'outil Akur8

3.4.3.1 Génération de modèles

Avant d'utiliser la solution Akur8, il faut au préalable construire et retraiter une base de données. Cette base va ensuite servir pour l'exécution des modèles.

Ensuite, les objectifs de modélisation doivent être clairement définis. Pour lancer un processus de construction de modèles, il faut d'abord spécifier la variable cible, une variable temporelle [4], une variable d'exposition et enfin une variable de stratification. Cette dernière sert à partitionner les données lors du processus de cross-validation.

De plus, en fonction des objectifs de modélisation, le type de loi doit être défini. Pour modéliser la fréquence d'arrêts de travail pour le risque incapacité deux lois s'offrent à nous :

— La loi de Poisson.

— La loi Binomiale Négative.

Ensuite, il faut sélectionner et intégrer toutes les variables explicatives. Les variables disponibles à postériori de l'étude, les variables aberrantes ou comportant des données manquantes, ainsi que les variables d'origine inconnue sont à proscrire de cette sélection. Les variables retenues sont alors susceptibles ou non d'être discriminantes dans l'estimation de la variable cible.

Enfin, il est indispensable de définir les paramètres suivants :

— Le pas de lissage :
Il indique le nombre de niveaux de lissage. Par exemple avec un lissage de 8, l'outil générera, pour chaque niveau de parcimonie, 8 modèles différents. Il y aura alors tout un spectre de modèles, partant de modèles très sensibles [5] et allant jusqu'à des modèles très robustes [6]. Par conséquent, un modèle peu lissé sera fidèle aux données tandis qu'un modèle très lisse sera plus robuste mais plus éloigné des données.

— Le pas de parcimonie :
La valeur de parcimonie correspond à un groupe de modèles ayant le même nombre de variables. Par défaut, l'outil choisit un pas égal à 5. Akur8 va alors générer cinq groupes de modèles. Chaque groupe est rattaché à un nombre spécifique de variables dans l'intervalle choisi.

— L'intervalle du nombre de variables à retenir :
Les modèles créés par l'algorithme d'apprentissage automatique comprendront un certain nombre de variables, appartenant à l'intervalle choisi. Il est possible de sélectionner entre 5

3. Voir Annexe 4.3.5 : Comparaison de modèles avec différents seuils de rejet.
4. Cette variable est très utile pour assurer la stabilité des modèles dans le temps.
5. Un modèle sensible avec un faible niveau de lissage est sujet au sur-ajustement.
6. Un modèle robuste avec un fort niveau de lissage est sujet au sous-ajustement.

et 30 variables. Les modèles générés sont alors de complexité simple (5 variables) à moyenne (30 variables).

Le choix de ces paramètres définit le nombre de modèles créés. Des valeurs élevées entraînent un gain de précision dans la création de modèles, mais aussi un temps de calcul plus long. En effet, pour chaque combinaison de niveaux de lissage et de parcimonie, cinq modèles sont construits :

— Quatre modèles de validation croisée sur le sous-ensemble de modélisation de la base de données.

— Un modèle sur le sous-ensemble de modélisation.

L'intégration de la validation croisée permet d'avoir une estimation fiable des performances des modèles.

3.4.3.2 Visualisation des résultats

Une fois le processus terminé, différents modèles sont alors proposés dans un GridSearch. Pour chacun des modèles créés, il affiche la performance en fonction du nombre de variables inclues. Par conséquent, ce graphique offre une visualisation claire du compromis possible entre la complexité du modèle et ses performances.

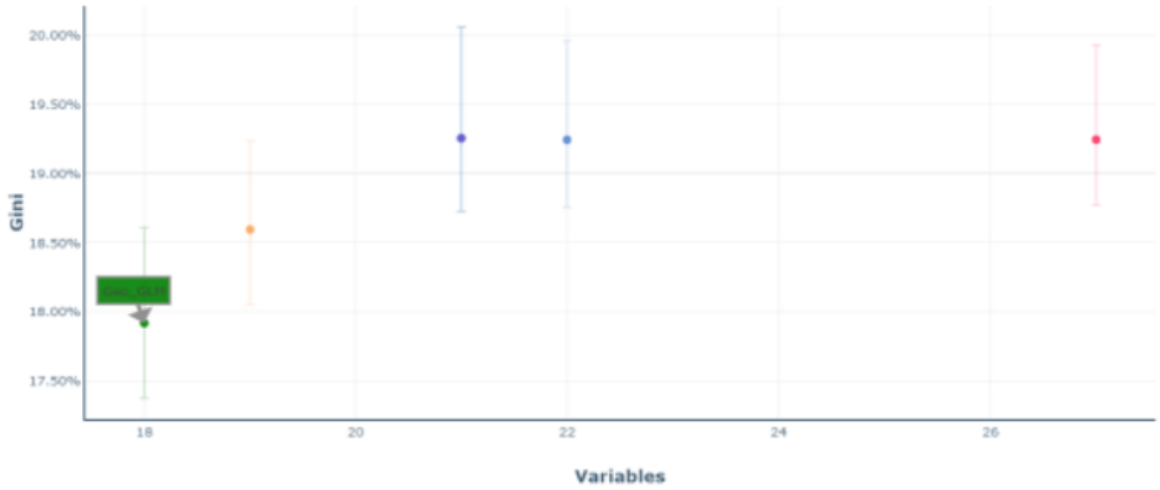


FIGURE 3.6 – Exemple de GridSearch

C'est à l'aide de ce graphique que l'utilisateur va choisir le modèle qu'il souhaite retenir. Un compromis entre la complexité du modèle et ses performances est à faire en fonction de ses besoins. En effet, un modèle avec de nombreuses variables sera plus difficile à interpréter. De plus, le coût des calculs sera plus important.

3.4.3.3 Inspection du modèle retenu

Après avoir sélectionné un modèle, un aperçu de diverses informations le concernant est consultable :

– Le spread de variables :
Il permet d'étudier les variables retenues dans le modèle ainsi que leur importance. Deux spreads sont proposés dans Akur8 :

- Le spread 100/0 : Il permet de visualiser l'influence de la variable observée sur la variable à prédire. Il est calculé de la manière suivante :

$$Spread_{100/0} = \frac{Max(Coefficients)}{Min(Coefficients)} - 1.$$

- Le spread 95/5 : C'est une mesure identique au spread précédent, après suppression des coefficients correspondant aux 5% les plus élevés et aux 5% les plus faibles du jeu de données. Il permet de ne pas tenir compte des valeurs extrêmes. Il représente donc une vision plus robuste de l'impact de la variable.

  Voici ci-dessous une illustration de ces deux spreads.

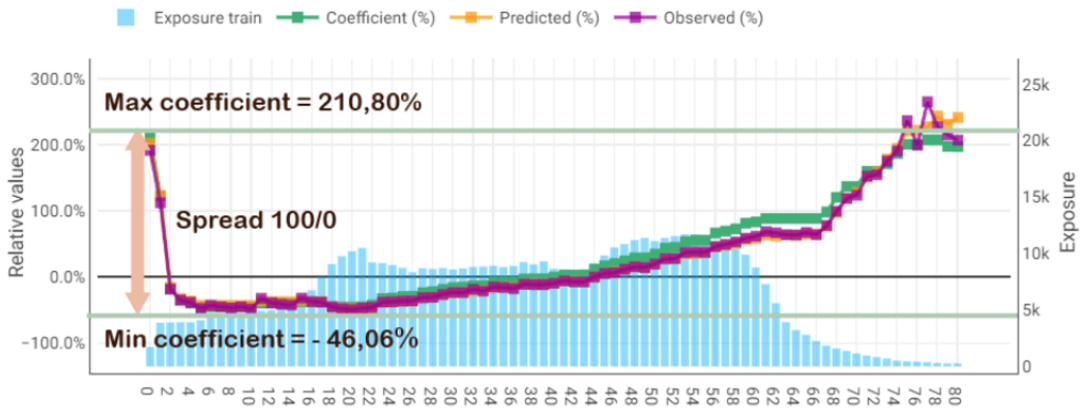


FIGURE 3.7 – Exemple de Spread

Le coefficient le plus élevé est de +210,8% et le plus faible de -46,06%. Le spread 100/0 associé à cette variable est donc calculé de la manière suivante :

$$Spread_{100/0} = \frac{100\% + 210,8\%}{100\% - 46,06\%} - 1 = \frac{310,8\%}{53,94\%} - 1 = 476\%.$$

Après suppression des coefficients correspondant à 5% des profils présentant le risque le plus élevé, le coefficient le plus élevé est de +80,2%. Après suppression des coefficients correspondant à 5% des profils présentant le risque le plus faible, le coefficient le plus bas est de -31,7%. Le spread 95/5 est donc :

$$Spread_{95/5} = \frac{100\% + 80,2\%}{100\% - 31,7\%} - 1 = \frac{180,2\%}{68,3\%} - 1 = 164\%.$$

– Les résidus quantiles randomisés :
  Ils servent à valider ou non l'adéquation de la loi utilisée par le modèle avec la distribution des données. L'adéquation est confirmée lorsque les résidus sont centrés par rapport à l'axe des ordonnées. Une tâche uniforme doit apparaître.

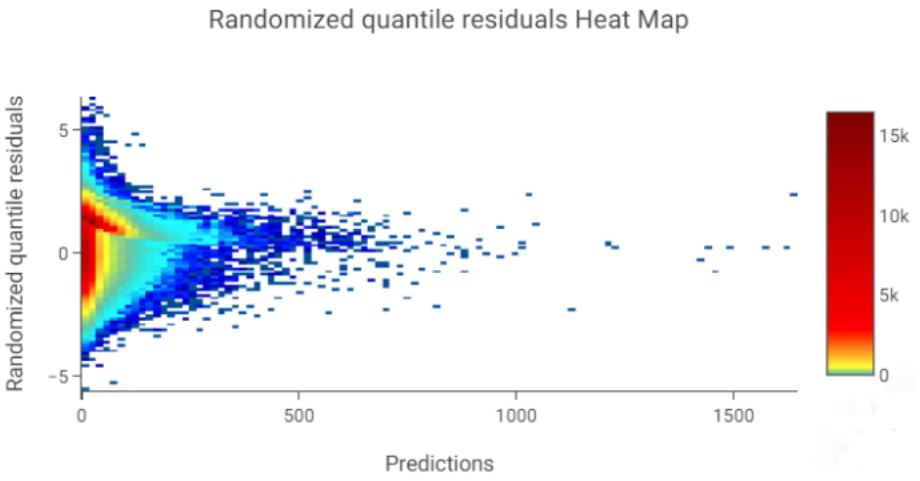


FIGURE 3.8 – Exemple de résidus

– Les courbes de Lorenz et de Lift :
  Ces courbes sont explicitées dans les parties suivantes.

– Les métriques de performance :
  Tous les indicateurs de performance utilisés sont détaillés dans les parties suivantes.

3.4.3.4 Personnalisation du modèle

Une fois que le modèle est choisi, l'utilisateur peut encore le personnaliser pour l'améliorer. Dans un premier temps, pour résoudre des problèmes d'ajustement, il est possible de modifier le niveau de lissage du modèle. Ce lissage est global, c'est à dire qu'il s'applique à l'ensemble des variables du modèle.

Un fort niveau de lissage entraîne un modèle plus robuste. Cependant, des signaux pertinents peuvent ne pas être captés. En revanche, un niveau de lissage trop faible est synonyme de sur-ajustement. Le modèle va capter du bruit et les estimations ne seront pas cohérentes. Le niveau de lissage optimal varie donc en fonction du modèle créé. Pour être le meilleur possible, il faut faire un compromis entre la robustesse de segmentation et le signal inclus dans le modèle.

Voici une image de deux niveaux de lissage pour la variable âge.

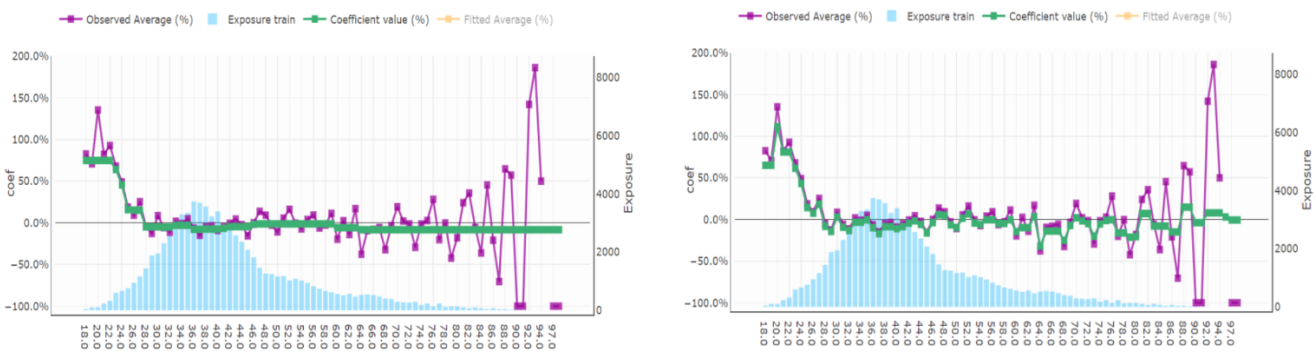


FIGURE 3.9 – Exemple de lissage : à gauche un modèle à faible lissage et à droite un modèle avec lissage optimal

Dans un second temps, des interactions entre les variables peuvent être ajoutées. Celles-ci sont ajustées de manière à ce que les effets sur la variable unique ne soient pas affectés. En effet, les interactions sont construites sur les effets résiduels.

Prenons l'exemple de l'ajout d'un terme d'interaction entre l'âge et le genre. Si le coefficient associé à cette nouvelle variable s'avère significatif, l'effet de l'âge des femmes sur le nombre d'arrêts peut être étudié indépendamment de celui des hommes.

Enfin, l'utilisateur récupère les coefficients retenus dans le modèle. Ils vont servir de base pour aboutir aux lois d'incidence. De plus, il peut vérifier la stabilité du modèle en le relançant sur de nouvelles données.

3.4.4 Avantages et Inconvénients

Les points forts de la solution proposée par l'outil Akur8 sont les suivants :

- Solution pour la tarification assurantielle :

  Cet outil propose donc une automatisation du processus de tarification en assurance. Grâce à l'intelligence artificielle et à la théorie sous-jacente, Akur8 propose un processus de pricing rapide, auditable et avec un bon pouvoir prédictif.

- Facilité de prise en main :

  L'outil mis à disposition est très ergonomique et facile d'utilisation. En effet, de nombreux graphiques rendent l'interface très visuelle. De plus, il ne demande pas de connaissances pointues sur la théorie des modèles sous-jacents. L'interface « clique-bouton » donne la possibilité à n'importe quel utilisateur d'utiliser cette solution.

- Flexibilité et stabilité :

  Les modèles théoriques sur lesquels se basent Akur8 présentent de meilleures performances. De plus, les prédictions obtenues sont plus robustes que lors des modèles classiques de Machine Learning.

- Gain de temps :

  Cet outil permet d'éviter de coder manuellement chacune des étapes de définition et de construction de modèles. De plus, lorsqu'un projet est lancé, différents modèles sont créés pour que l'utilisateur puisse choisir celui qui convient le mieux à ses attentes.

  Le nombre de modèles créés dépend du nombre de variables intégrées et des niveaux de lissage choisis. Le nombre récupéré est ensuite multiplié par le nombre de folds nécessaires pour le processus de cross-validation. En effet, la validation croisée est directement intégrée dans la construction des modèles.

- Contrôle important sur les modèles :

  Bien que ce soit un outil « clef en main », la place de l'utilisateur reste importante pour la bonne réalisation des modèles. En effet, suivant les contraintes et les besoins de l'utilisateur, il est possible de créer des modèles très spécifiques. L'utilisateur contrôle du début à la fin la construction des modèles, presque comme s'il faisait manuellement ces modèles à l'aide d'un logiciel de programmation par exemple.

Néanmoins, cet outil présente quelques limites.

– Utilisation de l'intelligence artificielle :

  Le monde de l'assurance est très réglementé. Le processus de tarification en assurance comprend donc beaucoup de contraintes. Cependant les modèles de Machine Learning utilisant l'IA sont souvent de nature « boîte noire ». Même si les développeurs ont fait énormément d'efforts sur ce point, il reste des zones d'ombre dans la construction des modèles.

– Tendance à moyenner les prédictions :

  En prenant en compte des hypothèses de la théorie de la crédibilité couplées avec les différents niveaux de lissage, les modèles ont tendance à prédire des résultats qui gravitent autour de la moyenne du portefeuille. Cette méthode sert à donner moins d'importance aux comportements extrêmes et aux groupes de population qui sont moins représentés. Cependant, certaines classes de salariés ont une fréquence d'arrêt largement sous-estimée par le modèle. Dans la suite, si les prédictions sont érronées alors la tarification qui en découle est également biaisée. Néanmoins, cet effet se compense. L'outil Akur8 permet de réduire l'effet d'une sur-incidence et au contraire d'augmenter l'effet d'une sous-incidence.

4 Choix du meilleur modèle et analyse des résultats

Après avoir présenté les éléments sous-jacents et détaillé la construction des trois modèles GLM, XGBoost et Akur8, nous allons étudier et comparer les résultats obtenus sur nos données.

Au cours de cette partie, nous présenterons d'abord les différents indicateurs de performance, nécessaires dans la sélection du meilleur modèle.

Par la suite, les trois algorithmes de modélisations seront confrontés les uns aux autres. A l'aide de plusieurs critières, le plus performant sera retenu.

Ensuite, ce modèle sera appliqué sur chacune de nos deux bases, pour analyser les évolutions de l'incidence et les conséquences dues à la crise sanitaire, sur notre portefeuille.

En dernier lieu et dans l'objectif de répondre à la problématique de ce mémoire, les résultats produits par ces modèles seront résumés et les conséquences sur le monde de la prévoyance détaillées. De plus, une étude focalisée sur les changements caractéristiques des arrêts de travail, en particulier concernant l'évolution des profils types, sera mise en place.

4.1 Présentation des indicateurs de performance

Pour rappel, les trois modèles employés pour la modélisation de l'incidence sont les GLM, XGBoost et l'outil Akur8. Ces trois modèles ne fonctionnent pas de la même façon et utilisent des algorithmes qui sont de complexité variable. Toutefois, dans le but de ne retenir que le modèle le plus performant, différents indicateurs de comparaison sont mis en place.

Pour obtenir le meilleur modèle de prédiction, deux critères principaux vont guider la comparaison :

- La performance du modèle :
  À travers ce critère, l'objectif est de quantifier à quel point le modèle arrive à segmenter la population par rapport au risque considéré. Un bon modèle sera un modèle capable de prédire correctement chaque catégorie homogène de population, sur des données d'entraînement mais principalement sur de nouvelles données. Plus un modèle est robuste, meilleure est sa capacité d'adaptation.

  Pour mesurer la performance des modèles, des indicateurs tels que l'indice de Gini, la courbe de Lift ou encore la courbe de Lorenz sont mis en place.

- La qualité du modèle :
  Le critère de la qualité du modèle se base sur la théorie de l'information. En effet, lors de la construction des modèles une partie de l'information est perdue. Ce critère permet donc d'estimer cette perte d'information lors de la modélisation du nombre d'arrêts par salarié et par an. Par conséquent, les caractéristiques qui influent le plus sur notre variable cible sont mises en valeur.

  Pour mesurer la qualité du modèle, s'intéresser aux erreurs RMSE ou MSE, au coefficient R^2 ou à l'analyse de la significativité des variables est nécessaire.

Somme toute, le modèle qui prédit le mieux l'incidence et qui présente les plus faibles résidus est défini comme le plus performant. Ceci reste vrai pour n'importe quelle base fournie en entrée.

Pour comparer des modèles de Machine Learning, il existe donc un grand nombre de métriques. Néanmoins, il est indispensable de mettre en place des indicateurs communs aux trois modèles pour mener à bien une analyse précise et pertinente.

Les principales métriques employées sont les suivantes :

- La courbe de Lorenz et l'indice de Gini :
  La courbe de Lorenz permet d'illustrer la répartition d'une variable cible dans une population donnée. Elle est également connue sous le nom de courbe de Gini ou « the Cumulative Accuracy Profile » (CAP) en anglais. C'est un outil très important pour quantifier la performance d'un modèle.

  Cette courbe a été développée en 1905 par M. O. Lorenz. Initialement utilisée pour le partage de la richesse dans une population donnée, elle permettait d'évaluer aisément les inégalités de répartition des revenus. En abscisse, se trouvait le pourcentage de la population, trié de manière croissante par rapport à leur richesse détenue.

Il est alors possible de quantifier la part cumulée des richesses détenue par x% des individus les moins riches de la population.

Au cours du temps, cette courbe a été adaptée dans différents domaines. Dans ce mémoire, elle représente la part cumulée du risque « *nombre d'arrêts de travail* », expliquée par la population assurée.

Elle est donc construite en ordonnant les prédictions et les observations. L'axe des abscisses indique les salariés triés par ordre décroissant en fonction de leur nombre d'arrêts prédits. L'axe des ordonnées classe par ordre croissant la part cumulée du risque observé. Cette courbe permet donc de quantifier la qualité des prédictions du modèle.

Voici un exemple de courbe de Lorenz pour mieux comprendre son fonctionnement.

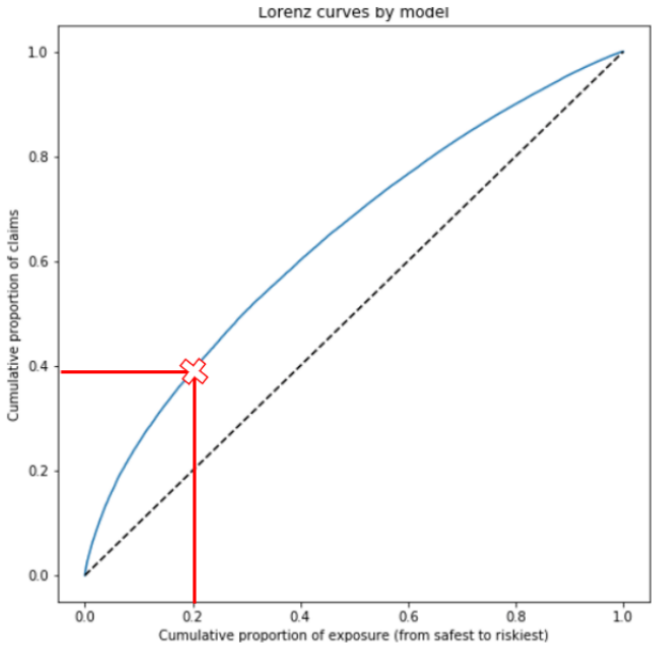


FIGURE 4.1 – Exemple de courbe de Lorenz

Le point identifié par la croix rouge indique que 20% des assurés du portefeuille qui ont le plus d'arrêts prédits, permettent d'expliquer quasiment 40% du nombre d'arrêts total, observé dans la base de données.

La diagonale en pointillés représente le cas d'égalité. Tous les individus de ce portefeuille sont donc égaux vis à vis du risque considéré. Par conséquent, tout au long de la diagonale, x% de la population explique x% du nombre d'arrêts prédit.

Une interprétation de la courbe de Lorenz peut être faite à l'aide du coefficient de Gini. En effet, cet indicateur a été introduit par C. Gini en 1912. Il se calcule en prenant deux fois la valeur de l'aire comprise entre la courbe de Lorenz et la diagonale en pointillés.

Souvent utilisé en statistiques, le Gini mesure la capacité de segmentation d'un modèle. Dans notre étude, le Gini indique la capacité du modèle à identifier le risque considéré.

La valeur de cet indice est comprise entre 0 et 1, exprimée en pourcentage. L'objectif est donc de maximiser ce coefficient. Plus le Gini est élevé et plus le pouvoir de segmentation est fort.

– La courbe de Lift :

C'est aussi une courbe qui permet de quantifier les performances d'un modèle. Comme pour la courbe de Lorenz, les prédictions sont ordonnées. Toutefois, dans le cas de la courbe de Lift, le tri des prédictions s'effectue de manière croissante, en les regroupant dans vingt groupes distincts. Chacun des groupes représente donc 5% de la population totale assurée.

Pour chaque groupe, la prédiction moyenne du modèle (points jaunes), ainsi que la moyenne du nombre d'arrêts observés (points violets) sont calculées. Ensuite, ces moyennes sont respectivement reliées les unes aux autres pour former les courbes ci-dessous. En jaune, la courbe correspond au nombre d'arrêts prédits et en violet c'est la courbe des arrêts observés.

Voici un exemple de courbe de Lift.

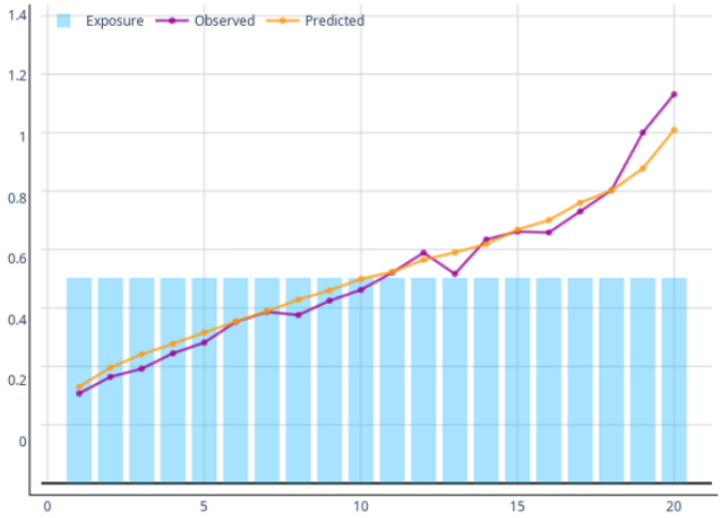


FIGURE 4.2 – Exemple de courbe de Lift

En bleu, légendé par « *Exposure* », se trouvent les 20 groupes d'individus. Les salariés qui figurent dans le groupe le plus à gauche du graphe (points de la courbe de Lift les plus bas) correspondent aux individus qui ont l'incidence la plus faible. À contrario, les points les plus à droite représentent les salariés qui ont de fortes incidences.

– Le coefficient R-squarred :

Cet indicateur est plus connu sous le nom de coefficient de détermination ou R^2. C'est une mesure statistique qui détermine la proportion de variance de la variable cible qui provient des variables explicatives. Le coefficient de détermination mesure donc la qualité d'ajustement du modèle considéré.

Il ne peut prendre que des valeurs comprises entre 0 et 1. Il permet d'expliquer $R^2\%$ des variations de Y, la variable cible. Plus ce coefficient est proche de 1, plus le modèle est en adéquation avec les données collectées et plus il est de bonne qualité. De sorte que, les variables utilisées expliquent très bien la variable cible. Inversement, un coefficient proche de 0 indique un modèle de mauvaise qualité.

Le coefficient de détermination se calcule comme suit :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSE}{SST},$$

avec :

— n le nombre d'observations ;

— y_i la valeur de la i^{ème} variable cible ;

— $\hat{y}_i$ sa valeur prédite ;

— $\bar{y}$ la moyenne du nombre d'arrêts observés ;

Dans le cas des GLM, la variance totale des données observées, en anglais « Sum of Squared Total », est égale à la somme de la variance expliquée par le modèle (SSE) et de la variance expliquée par les résidus (SSR).

Dans la réalité, en présence de bases de données bruyantes, la valeur du R^2 n'est jamais très élevée, en particulier pour des risques complexes comme la prévoyance. En général, il ne dépasse pas quelques points de pourcentage.

– Les erreurs RMSE et MSE :

RMSE signifie Root Mean Square Error et MSE Mean Square Error. L'erreur MSE est donc le carré de RMSE. Ces deux indicateurs mesurent l'écart entre les valeurs prédites par le modèle et les valeurs observées dans la réalité.

Le terme « erreur » représente donc la différence entre le prédit et l'observé. Les modèles les plus performants sont donc ceux qui minimisent ces deux indicateurs.

Voici la formule du RMSE :

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

On peut éventuellement multiplier cette formule par des pondérations attribuées aux observations :

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

$$avec \sum_{i=1}^{n}\omega_i = 1.$$

Il est important de noter que ces deux métriques sont sensibles aux données extrêmes, très éloignées de la moyenne.

– Les erreurs MAE :

Cet indicateur signifie Mean Absolute Error. Comme pour les deux métriques ci-dessus, l'erreur MAE mesure la différence entre le prédit et l'observé. Le MAE est construit de telle sorte qu'il limite l'impact des valeurs extrêmes sur l'erreur totale :

$$MAE = \sum_{i=1}^{n} |y_i - \hat{y}_i|.$$

De même, pour chercher le meilleur modèle il faut minimiser cette métrique.

4.2 Comparaison des modèles

La comparaison de modèles se déroule en deux étapes majeures. Dans un premier temps, une analyse de la significativité des variables est faite. Le but est de définir quelles vont être les variables retenues. Dans un second temps, les performances des modèles sont confrontées, d'abord grâce aux indicateurs précédemment définis et ensuite, à l'aide d'une analyse sur l'incidence prédite, variable par variable.

Le processus d'analyse et de détermination du meilleur modèle s'avère être identique pour nos deux bases d'étude. Afin d'éviter les répétitions, la démarche détaillée par la suite ne concerne que le périmètre observé en période de crise sanitaire.

4.2.1 Rappel du périmètre

Dans cette partie, les trois modèles GLM, XGBoost et Akur8 sont comparés. Pour que la comparaison ait du sens, les hypothèses prises et les périmètres étudiés doivent être identiques pour ces trois modèles. C'est pourquoi il a fallu commencer par construire plusieurs modèles en faisant varier le nombre de variables explicatives et/ou les lois sous-jacentes introduites.

L'ajout d'une variable supplémentaire dans un modèle a un coût. L'objectif est donc de maximiser la performance en minimisant le nombre de variables. Pour ce faire, un premier test de performance est effectué avec des modèles composés uniquement de trois variables. En effet, les normes tarifaires déjà en place ne se basent que sur trois variables principales, qui sont l'âge, le genre et la CSP.

Ensuite, un second modèle est testé. Il se compose de toutes les variables misent à disposition dans cette étude.

Enfin, de manière successive, les variables sont ajoutées ou retirées des modèles testés, les unes après les autres. Cela permet d'identifier lesquelles il faudrait retenir pour aboutir au modèle le plus performant. À chaque étape, les différentes métriques sont analysées.

Finalement, le modèle retenu qui présente les meilleures performances se compose de cinq variables explicatives, qui sont : le salaire annuel du salarié, la taille de son entreprise, sa CSP, son âge et enfin son genre.

En parallèle, des tests sont effectués sur les fonctions de lien utilisées dans les modèles GLM et Akur8. Ces tests se focalisent principalement sur la loi de poisson et la loi binomiale négative, qui sont les plus courantes et les plus adéquates dans la modélisation d'une fréquence. Avec l'outil Akur8, les résultats observés avec ces deux lois sont très similaires pour ne pas dire identiques. En revanche, avec les modèles implémentés sous Python, la distribution de poisson donne de meilleures métriques (Gini et R^2 majoritairement). Le choix final se porte donc en faveur de la distribution de poisson.

En outre, pour maximiser les performances du modèle XGBoost, il est nécessaire d'optimiser les paramètres du modèle. Ceci a été réalisé à l'aide d'un « Grid Search », comme présenté dans la partie *3.3.5* de ce mémoire.

4.2.2 Analyse des indicateurs de performance

Après avoir défini le nombre et le nom des variables explicatives à introduire dans les modèles, l'analyse se poursuit autour des cinq métriques numériques et des deux courbes présentées précédemment.

Le tableau ci-dessous récapitule l'ensemble des indicateurs numériques pour le modèle 2020 composé de cinq variables, construit d'abord avec les GLM, puis XGBoost et enfin Akur8.

	GLM	XGBost	Akur8
Gini	27,44%	28,99%	26,67%
R^2	4,5%	5,61%	7,99%
RMSE	1,437	1,422	1,136
MSE	2,211	2,023	1,29
MAE	0,876	0,833	0,738

Dans un premier temps, il est important de remarquer que pour chacun de ces modèles, les indicateurs présentés sont du même ordre de grandeur. Ils sont tous relativement proches, ce qui permet de confirmer la bonne réalisation de toutes les modélisations et l'acceptabilité des résultats obtenus.

Pour rappel, plus le coefficient de Gini et le R^2 sont élevés, plus le modèle est performant. Au contraire, plus les erreurs RMSE, MSE et MAE sont faibles et meilleur le modèle est.

Les GLM sont les modèles les moins complexes. D'après ce tableau, ils présentent des performances inférieures aux autres modèles. Néanmoins, l'indice de Gini du modèle GLM est légèrement meilleur que celui d'Akur8. Le pouvoir de segmentation de ces deux modèles est donc proche, mais meilleur pour le GLM.

Ensuite, le modèle XGBoost a un coefficient de Gini plus important que les deux autres algorithmes. En revanche, pour les autres métriques, Akur8 semble être la solution la plus performante. En effet, le coefficient R^2 est supérieur de 3,5% par rapport au GLM et de plus de 2% comparé à XGBoost. De plus, les indicateurs relatifs aux erreurs (RMSE, MSE et MAE) sont plus faibles et donc meilleurs avec l'outil Akur8.

Grâce à cette analyse, une première idée du modèle à retenir semble apparaître. Néanmoins, ce n'est pas suffisant. Désormais, la comparaison se poursuit à travers l'étude des différentes courbes de Lift et de Lorenz. Le but est donc de confirmer ou de réfuter l'idée initiale.

FIGURE 4.3 – Comparaison des courbes de Lift

Les deux courbes de Lift les plus à gauche, respectivement relatives aux modèles GLM et XGBoost, proviennent d'un code Python alors que la dernière est réalisée à l'aide de l'outil Akur8.

Pour les GLM, le modèle prédit très bien les incidences faibles. En effet, pour les groupes de points les plus à gauche du graphe, l'observé en violet et le prédit en jaune se superposent. Cependant, pour les salariés qui ont une forte incidence (à droite du graphe), le modèle a plus de difficulté à s'adapter. Par conséquent, les individus qui ont une forte fréquence d'arrêts et qui possèdent donc des caractéristiques particulières, ne sont pas idéalement spécifiés par ce modèle. De fait, le modèle prédit une sur-incidence de presque 10% par rapport à la réalité.

En outre, la courbe du nombre d'arrêts prédits est toujours au-dessus de celle de l'observé. Par conséquent, les GLM semblent souffrir d'une légère tendance à la surestimation. Pour une compagnie d'assurance, ce modèle paraît plus prudent. Avec quelques ajustements, il peut tout à fait être retenu pour aboutir à une nouvelle norme tarifaire.

Ensuite, les moyennes des prédictions pour chaque groupe se situent dans un intervalle compris entre 0 et 1,12. Toutefois, les moyennes des arrêts observés sont comprises entre 0 et 1,02, et sont toujours inférieures aux prédictions. Ainsi, les salariés qui se trouvent dans le groupe le plus à gauche, ont en moyenne aucun arrêt durant l'année 2020. À contrario, ceux qui se trouvent dans le dernier groupe, le plus à droite, ont en moyenne 1,02 arrêts par an.

En ce qui concerne le modèle XGBoost, les prédictions semblent très proches de l'observé, excepté pour le dernier groupe de population qui possède l'incidence la plus forte. Pour ce dernier groupe l'erreur de prédiction est importante. En effet, le modèle prédit un nombre d'arrêts plus faible que la réalité. L'écart de prédiction est de l'ordre de 8%. Par conséquent, ce modèle aussi ne capte pas correctement les caractéristiques liées à une forte incidence.

Dans ce cas précis, il faut faire attention. Le fait de sous-estimer le nombre d'arrêts fait augmenter drastiquement le risque de ce modèle d'un point de vue actuariel. En effet, une sous-estimation du nombre d'arrêts entraîne une sous-tarification de cette classe d'assurés. Par conséquent, en fin de cycle de production, l'assureur se retrouve avec un mauvais ratio *S/P*. Au final, un produit tarifé avec ce modèle risque d'attirer en priorité les mauvais risques et ensuite de faire fuir les bons. En effet, si la tarification initiale est trop basse, il va falloir réhausser le prix du contrat en question. Les salariés qui ont la possibilité de trouver un couverture moins cher chez un autre assureur représentent les bons risques. Ils vont potentiellement quitter le portefeuille et partir à la concurrence. Ainsi, cela entraîne une hausse de la proportion de mauvais risques et donc une mutualisation plus difficile dans ce genre de situation.

Le modèle XGB a donc plus de difficultés à segmenter la population. Néanmoins, l'analyse des intervalles des différentes moyennes d'incidence semble conclure le contraire. Effectivement, les moyennes des valeurs observées sont désormais comprises entre 0,104 et 1,496 alors que l'intervalle était [0 ; 1,02] pour les GLM. En effet, il semble donc que XGB segmente mieux la population en classes homogènes.

Enfin, les prédictions du modèle Akur8 semblent se regrouper autour du nombre d'arrêts moyen. Akur8 surestime les faibles incidences et sous-estime les incidences plus fortes. Néanmoins, comme ce modèle a de meilleurs indicateurs globaux sur les erreurs, les prédictions qui en découlent sont donc plus proches de la réalité.

De plus, les moyennes du nombre d'arrêts observé sont comprises entre 0,13 et 1,13. La segmentation du modèle Akur8 semble être à mi-chemin entre celle des GLM et XGBoost. L'analyse de l'écart du vingtième groupe montre une sous-estimation de quasiment 17% par rapport à la réalité, ce qui est bien plus que le modèle XGB. Pour les incidences extrêmes, le modèle Akur8 semble donc beaucoup plus risqué.

Désormais, les courbes de Lorenz sont examinées.

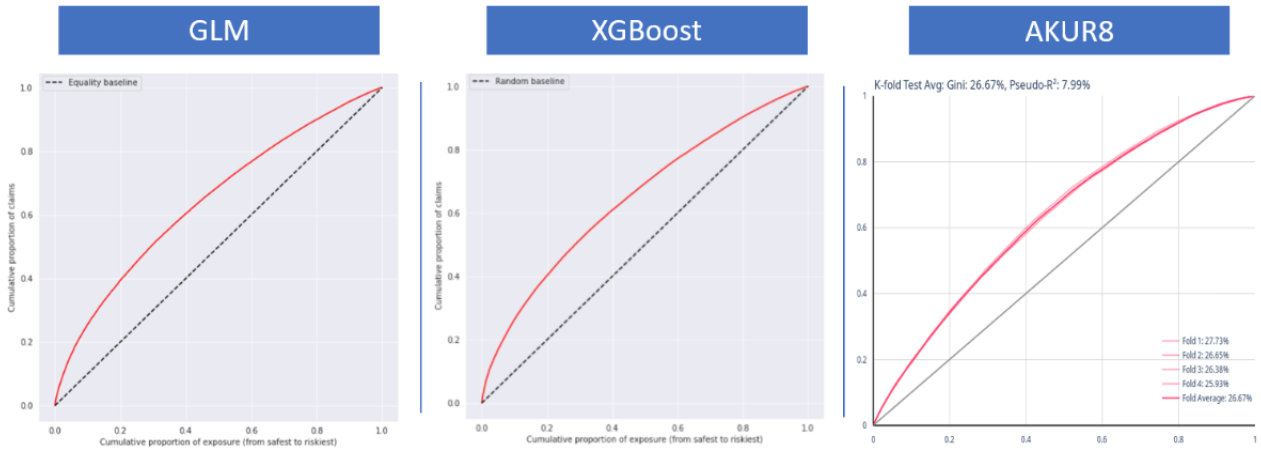


FIGURE 4.4 – Comparaison des courbes de Lorenz

Pour rappel, la courbe de Lorenz est la représentation graphique du coefficient de Gini. Le modèle XGBoost a un Gini de presque 29% alors que les GLM et Akur8 ont un coefficient un peu inférieur. Visuellement, la différence entre ces courbes n'est clairement pas flagrante. Par exemple, avec les modèles XGB et GLM, 20% des salariés qui ont le plus d'arrêts expliquent environ 40% du risque. Toutefois, pour le modèle Akur8, ces mêmes 20% d'individus ne caractérisent plus que 35% du risque considéré.

Cet écart vient confirmer l'analyse faite sur les courbes de Lift. En effet, l'estimation des fortes incidences est moins bien effectuée dans le modèle Akur8, comparé aux GLM et XGBoost.

Finalement, l'étude de ces deux courbes ne permet pas d'isoler avec certitude le modèle le plus performant. C'est pourquoi il semble nécessaire de la compléter avec un comparatif de chaque variable du modèle. Il repose sur la confrontation entre l'incidence prédite par les différents modèles et l'incidence réellement observée, modalité par modalité.

4.2.3 Analyse de l'incidence variable par variable

Cette opposition est réalisée sur chacune des cinq variables présentes dans le modèle. Le genre ne comporte que deux modalités donc la mise en place d'un histogramme est préférée pour analyser les écarts d'incidence. En revanche, comme les autres variables ont davantage de modalités, une dérivée de loi d'incidence est construite pour chacune d'elles. En effet, grâce aux résultats de nos différents modèles sur l'incidence, le taux de nouveaux cas en arrêt est affiché en fonction de chaque modalité.

Dans les cinq graphiques qui vont suivre, le même code couleur est employé. En effet, l'incidence observée durant l'année de la crise sanitaire est représentée par la couleur verte. Ensuite, une couleur est affectée à chacun des modèles précédemment construits : le bleu pour les GLM, le gris pour XGBoost et le orange pour Akur8.

Tout d'abord, l'analyse se porte sur le genre.

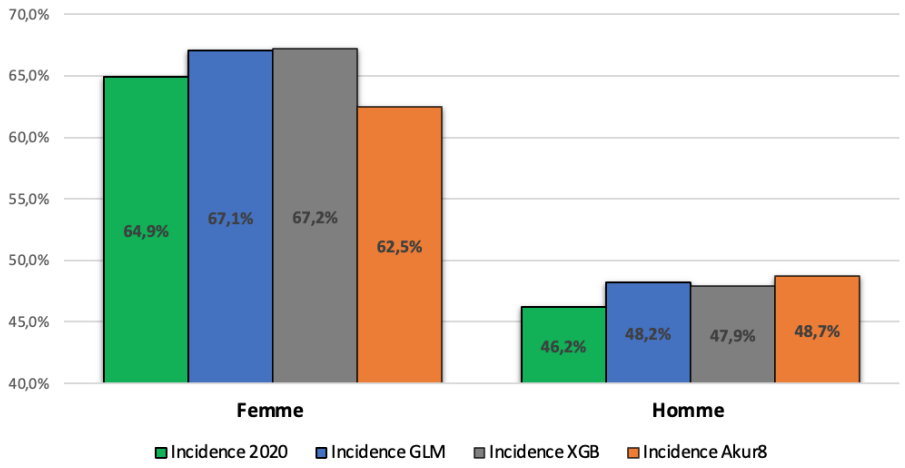


FIGURE 4.5 – Histogramme de l'incidence en fonction du genre

L'histogramme ci-dessus présente les différents taux d'incidence pour les femmes, du côté gauche et pour les hommes, du côté droit.

Dans la réalité, pour l'année 2020, les femmes sont victimes d'un plus grand nombre d'arrêts que les hommes. En effet, leur incidence est de 64,9%, alors qu'elle n'est que de 46,2% pour les hommes. Les trois modèles intègrent bien la sur-incidence des femmes, dans des proportions plus ou moins justes.

Le modèle GLM a une tendance globale à surestimer le nombre d'arrêts prédits. Effectivement, pour les deux modalités homme et femme, les prédictions sont surestimées de respectivement 2,1% et 2%. De plus, pour cette variable, les modèles GLM et XGB prédisent quasiment les mêmes taux d'incidence. Pour les femmes, l'incidence est aux alentours de 67% tandis que pour les hommes elle n'est que de 48%.

Pour rappel, la solution Akur8 a tendance à ajuster ses prédictions autour de la moyenne. Pour une variable comme le genre, le modèle sous-estime donc la modalité avec une forte incidence, en la rapprochant de la moyenne du portefeuille. À contrario, il surestime l'incidence plus faible. En effet, l'incidence est sous-estimée de 2,4% pour les femmes, alors qu'elle est surestimée de 2,5% pour les hommes.

Néanmoins, comme on a pu le voir avec les courbes de Lift, les caractéristiques d'une forte incidence sont difficilement prises en compte. La modalité femme fait partie de ces caractéristiques. De manière générale, XGB sous-estime ce type de modalité. Cependant, la catégorie femme est ici surestimée. Une forme de corrélation entre les variables semble se dégager de ces résultats. En effet, le modèle va donc sous-estimer dans des proportions plus importantes les autres modalités représentatives d'une forte incidence.

Ensuite, l'étude se poursuit avec la catégorie socio-professionnelle du salarié.

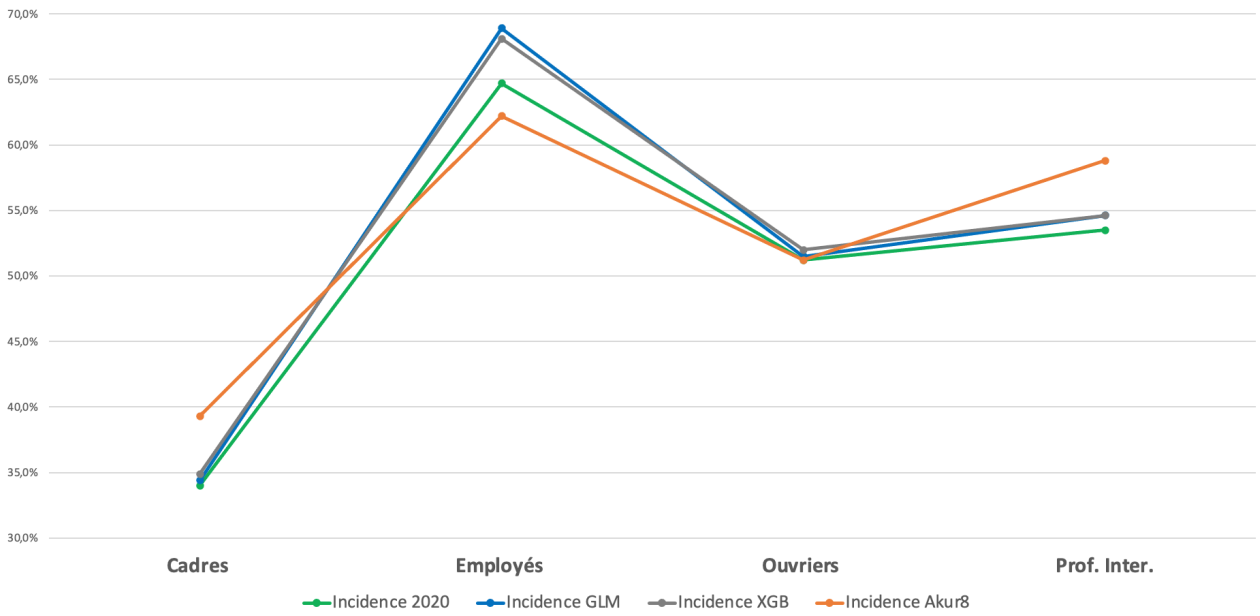


FIGURE 4.6 – Étude de l'incidence en fonction de la CSP

Pour rappel, l'incidence moyenne de ce portefeuille est de 55%. En vert, l'incidence observée pendant la crise sanitaire, elle varie presque du simple au double, en fonction de la CSP du salarié. En effet, la catégorie « Cadres » a la plus faible incidence, 21% inférieure à la moyenne du portefeuille. Par conséquent, lorsqu'en moyenne un individu de ce portefeuille est victime de 0,55 arrêt dans l'année, un salarié cadre n'en fait que 0,34. À l'inverse, la catégorie « Employés » a la plus forte incidence, 10% au-dessus de la moyenne.

En outre, la modalité « Ouvriers » est très bien prédite. Quelque que soit le modèle, l'incidence converge vers le taux observé durant la pandémie. C'est pourquoi les caractéristiques spécifiques qui définissent la classe ouvrière semblent bien prises en compte par les différents modèles.

Comme pour la variable liée au genre, les prédictions des modèles GLM et XGB sont quasiment similaires pour les quatre catégories socio-professionnelles ci-dessus. Par conséquent, les écarts avec l'incidence réelle sont faibles et se regroupent en dessous des 1,5% pour les modalités « Ouvriers », « Cadres » et « Professions Intermédiaires ». Cette différence s'accroît légèrement pour les « Employés » puisque les écarts sont désormais entre 3% et 4%. Pour résumer, ces deux modèles s'ajustent correctement avec la variable explicative sur la CSP. Néanmoins, chacune des modalités de la CSP est surestimée.

En revanche, comme le modèle Akur8 a tendance à prédire en se rapprochant de l'incidence moyenne, les écarts varient davantage. En effet, les catégories « Cadres » et « Prof. Inter. » sont peu repré-

sentées dans notre portefeuille et leur sinistralité apparait bien différente des autres modalités. C'est pourquoi Akur8 est plus prudent concernant les prédictions de ces classes d'individus. Par conséquent, dans Akur8, les professions intermédiaires et les cadres sont surestimés d'environ 5%. De plus, les employés qui ont une forte incidence sont quant à eux sous-estimés de 2,5%.

Cependant, les professions intermédiaires ont une incidence très proche de la moyenne, égale à 53,5%. Ainsi, le résultat des prédictions semble inattendu. Deux raisons expliquent cet écart de modélisation. Premièrement, les professions intermédiaires ne représentent que 10% de notre portefeuille, donc leur poids dans les estimations est faible. Deuxièmement, il y a probablement un effet de corrélation avec une autre variable. Effectivement, la catégorie des professions intermédiaires se compose des métiers de l'enseignement, de la santé ou encore de la propreté qui sont majoritairement représentés par des femmes. Si bien que, la sous-prédiction de l'incidence des femmes pourrait venir compenser la sur-estimation des professions intermédiaires.

Après avoir analysé cette modalité indépendamment des autres, les professions intermédiaires de notre portefeuille sont effectivement composées en majorité de femmes, âgées entre 30 et 45 ans, travaillant dans des entreprises de plus de 250 salariés et avec un salaire moyen.

Pour continuer, l'étude se focalise sur l'âge des salariés.

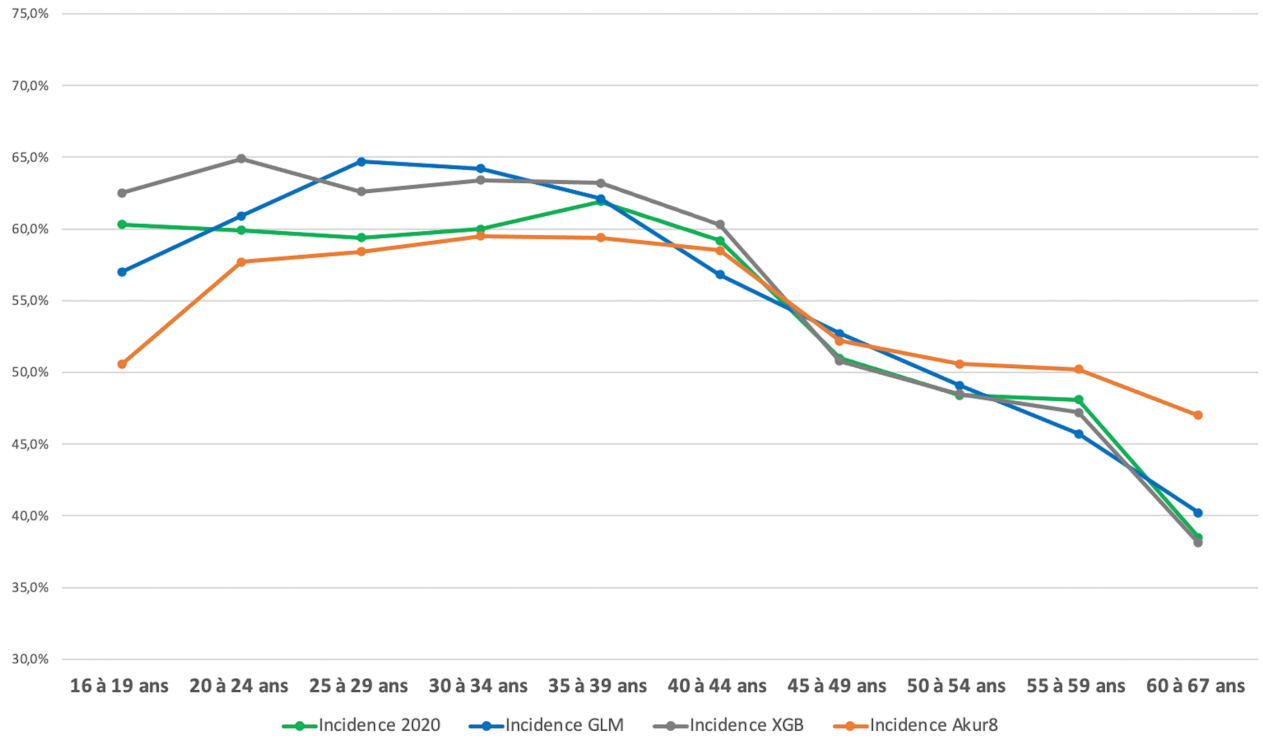


FIGURE 4.7 – Étude de l'incidence en fonction de l'âge

La variable continue contenant l'âge du salarié est transformée en variable catégorielle, définie par tranches de cinq ans d'âge. Dans un premier temps, l'incidence réelle des salariés âgés de 16 à 45 ans est quasi constante et se situe aux alentours de 60%. Au-delà de ce seuil, l'incidence s'avère décroissante en fonction de l'âge. Dans des proportions différentes, les trois modèles captent cet effet de stagnation avant 45 ans pour ensuite observer une légère décroissance.

Les résultats des trois modélisations montrent une baisse d'incidence pour les salariés les plus jeunes. Ce phénomène est intéressant à analyser puisque ce n'est clairement pas le cas pour le

taux d'incidence réel. Pour rappel, cette tranche de population accuse une exposition annuelle très en dessous des autres. Par conséquent, cette faible exposition semble être responsable d'un biais dans l'estimation.

Pour le modèle GLM, une décroissance quasi linéaire est remarquable à partir de la tranche 35/39 ans. Pour les âges compris entre 35 et 67 ans, les prédictions sont très proches de la réalité. Néanmoins, une bosse de surestimation apparaît pour les salariés entre 20 et 35 ans.

Le modèle XGBoost est sûrement celui qui capte le mieux l'influence de chaque tranche d'âge par rapport à la cible. Effectivement, pour tous les âges supérieurs à 35 ans, ce modèle se superpose presque en tout point avec la réalité. Toutefois, les prédictions des salariés âgés entre 20 et 35 ans sont surestimées de près de 3%.

Quant à lui, le modèle Akur8 prédit relativement bien l'ensemble des catégories comprises entre 20 et 60 ans. En revanche, pour les deux tranches d'âges extrêmes (moins de 20 ans et plus de 60 ans) les écarts explosent. Ici aussi, ce modèle souffre du phénomène de moyennage. Cela implique une surestimation de 8% des salariés les plus âgés, qui ont l'incidence réelle la plus faible, et une sous-estimation de 10% pour les plus jeunes. Néanmoins, ces deux classes d'âge extrêmes sont sous-représentées dans notre portefeuille. Elles représentent respectivement 5% et presque 3,5% de notre population totale. Par conséquent, le modèle Akur8 n'accorde pas la même importance aux prédictions de ces sous-ensembles, comparé aux autres catégories.

Désormais, c'est au tour du nombre de salariés dans l'entreprise d'être analysé.

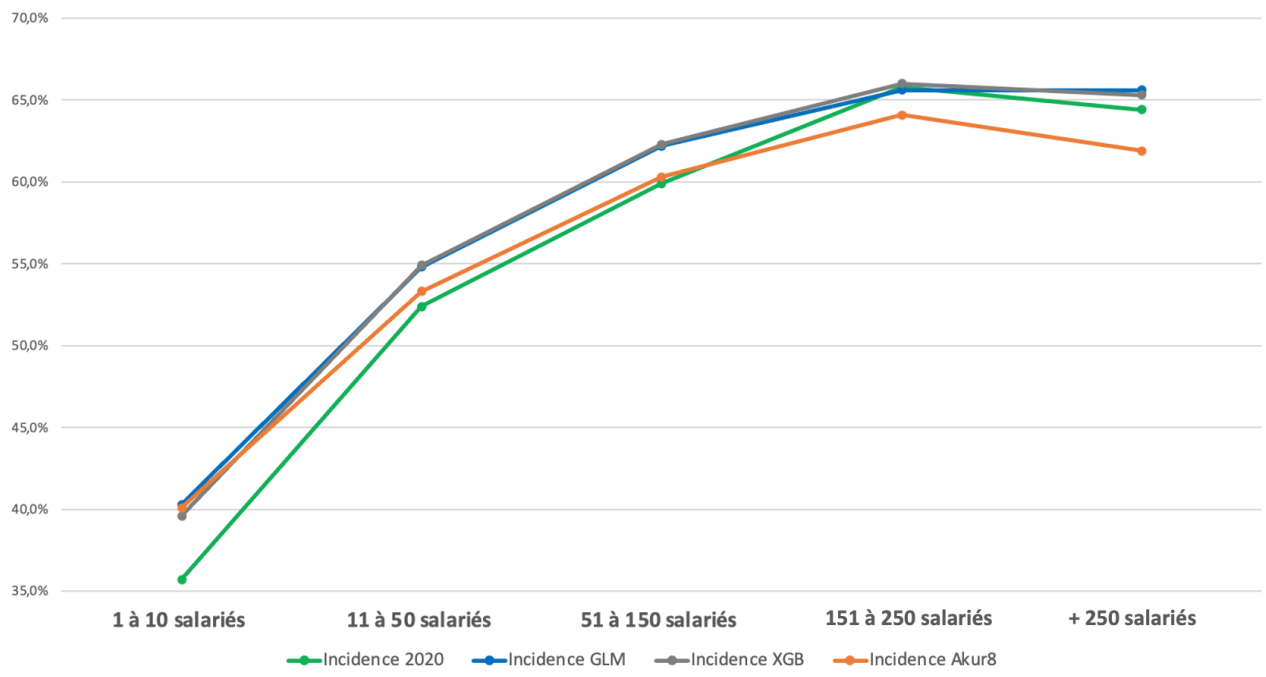


FIGURE 4.8 – Étude de l'incidence en fonction de la taille de l'entreprise

Contrairement à l'âge, plus le nombre de salariés dans l'entreprise augmente et plus le nombre d'arrêts s'intensifie. Ce phénomène n'est pas incohérent puisque les salariés des grandes entreprises sont en général mieux couverts en cas d'arrêt, en comparaison avec les salariés qui travaillent dans des PME par exemple. En effet, dans les grands groupes, grâce au phénomène de mutualisation

du risque, des régimes d'indemnisation complémentaires sont souvent mis en place avec de nombreuses garanties et négociés à des tarifs avantageux. À contrario, dans les petites structures, les complémentaires restent plus onéreuses. Les garanties choisies sont souvent moins intéressantes pour les employés. Par conséquent, cela entraîne une perte de revenu plus importante lorsque le salarié tombe en arrêt maladie.

Pour cette variable, les trois modèles semblent alignés concernant les prédictions. En effet, pour les entreprises qui se composent de plus de 10 salariés, GLM, XGB et Akur8 sont similaires en termes d'écarts de prédictions, entre 1% et 3%.

En revanche, pour la catégorie des entreprises qui emploient entre 1 et 10 salariés, le pourcentage d'erreurs est plus conséquent. Les modèles prennent en compte la sous-incidence de cette modalité mais dans des proportions trop peu importantes par rapport à la réalité. Un écart de presque 5% se dégage.

De la même manière que précédemment, les fortes incidences sont sous-estimées avec le modèle Akur8 alors que les faibles incidences sont surestimées.

Enfin, l'analyse se termine avec la variable relative aux salaires annuels.

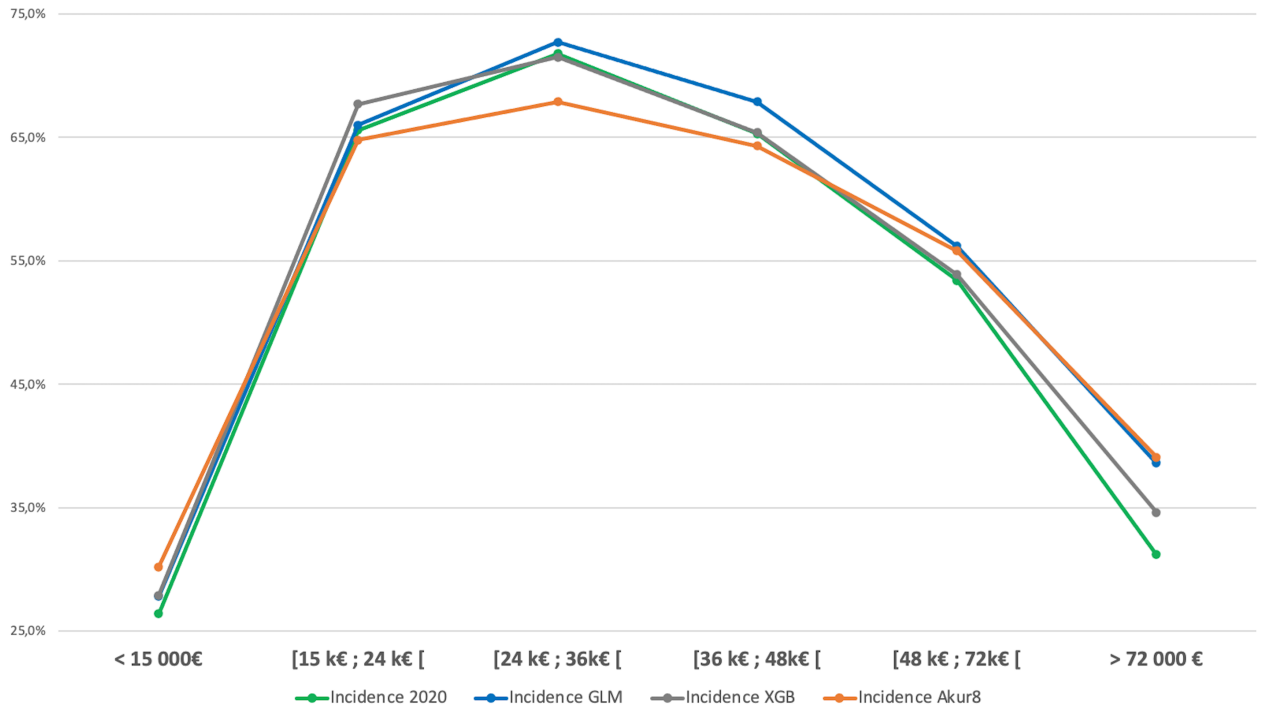


FIGURE 4.9 – Étude de l'incidence en fonction du salaire annuel

Avant tout, les taux d'incidence de cette variable sont très intéressants à analyser. Les salariés, qui ont un salaire annuel compris entre 15 000€ et 48 000€, accusent d'une incidence très élevée, aux alentours de 65%. Cette tranche regroupe plus de 60% de la population étudiée.

L'incidence décroit fortement pour les salaires supérieurs à 48 000€. Effectivement, au-delà de ce seuil, ce sont principalement des assurés qui occupent des postes de bureaux, donc avec une pénibilité moindre. De plus, ils ont généralement des postes clés et leur absence au sein de l'entreprise peut avoir de lourdes conséquences.

En outre, les assurés qui ont moins de 15 000€ de revenu annuel sont ceux qui ont le taux d'incidence le plus faible. On peut supposer que leur situation précaire ne leur permet pas de pouvoir survivre au quotidien avec une baisse de revenu.

Pour rappel, lors d'un arrêt de travail, la sécurité sociale prend en charge 50% du salaire. La complémentaire vient ensuite combler le déficit de revenu. En revanche, pour bénéficier de ces indemnités, il faut justifier de certaines conditions, comme un nombre d'heures travaillées au préalable ou un nombre de jours de cotisations payées. Dans cette tranche de population, nombreux sont les individus en CDD qui ne n'entrent donc pas forcément dans ces conditions.

Ensuite, les résultats de nos différents modèles sont examinées. Pour les salariés qui ont un revenu supérieur à 72 000€, les erreurs de prédictions sont les plus importantes. Elles se situent entre 3% et 8%. En revanche, pour les autres modalités de salaire, les modèles prédisent mieux. Les écarts des modèles GLM et XGBoost sont tous inférieurs à 2%, par rapport à l'incidence réelle.

Par construction, le modèle Akur8 est une fois de plus distant de la réalité pour les incidences éloignées de la moyenne du portefeuille. En outre, par construction de l'outil Akur8, la surestimation des salariés qui ont le plus de revenu semble logique. En effet, au vu de leur faible représentation dans le portefeuille (seulement 6%), le poids affecté à leurs estimations est inférieur aux autres modalités.

4.2.4 Choix du meilleur modèle

La comparaison des modèles est à présent terminée. Plusieurs conclusions sont à prendre en compte et sont résumées dans le tableau récapitulatif suivant.

Critères	GLM	XGBoost	Akur8
Interprétabilité	++	-	+ +
Indicateurs de performance	+	+ +	+ +
Robustesse	+	+ +	+ +
Pouvoir prédictif	+	+ +	+
Facilité de mise en place	+	+	+ +

Dans un premier temps, l'analyse des différentes métriques de performance a permis d'éliminer le modèle GLM. Il présente les moins bons indicateurs. De plus, contrairement à la solution Akur8, les GLM ne captent que des effets linéaires. Pour chaque variable, le modèle ne calcule qu'un seul coefficient et donc les erreurs de prédictions peuvent augmenter.

Ensuite, les courbes de Lift ont mis en évidence le fait que les deux modèles XGB et Akur8 ont tendance à surestimer les faibles incidences et à sous-estimer les fortes incidences. En revanche, dans la plupart des cas, Akur8 présente des écarts de prédiction pour les sous-échantillons qui ne représentent qu'une infime partie de l'ensemble total. Ceci est dû aux poids des estimations réduits à l'encontre de ces catégories. Grâce à l'étude des courbes de Lorenz, qui complète l'information du coefficient de Gini, une préférence pour le modèle XGB se dégage.

Enfin, cette comparaison se finalise en analysant l'incidence de chaque variable, modalité par modalité. Des écarts de prédictions, qui s'expliquent par l'effet de moyennage et des poids affectés aux différentes modalités, est bien présent pour le modèle Akur8.

Grâce à ces comparaisons, les trois modèles étudiés s'avèrent performants. De plus, ils semblent bien s'ajuster aux données du portefeuille. Toutefois, pour l'ensemble des variables, le modèle XGB paraît être le meilleur. Il présente le moins d'écarts de prédictions pour la majeure partie des modalités.

Finalement, l'algorithme XGBoost semble être plus performant qu'Akur8. Il devrait donc être choisi comme étant le meilleur modèle. Cependant, le modèle XGB souffre de sur-apprentissage et sous estime l'incidence des individus les plus risqués du portefeuille. Cela entraîne une hausse du risque dans la mise en place du processus de tarification. De plus, il pâtit principalement de contraintes d'interprétabilité et de mise en place. En effet, les modèles de Machine Learning et spécifiquement XGBoost qui est aujourd'hui désigné comme le meilleur, sont des algorithmes dits « boîte noire » ou *« black box »* en anglais. Les modèles mathématiques sous-jacents et les méthodes utilisées sont complexes à expliciter.

À l'origine, ce terme de *« black box »* est donné puisque seules les données en entrée et les résultats en sortie sont observables, sans forcément en comprendre le fonctionnement interne. Il reste très compliqué de quantifier clairement l'impact de la totalité d'une variable sur la performance du modèle. Ces algorithmes sont donc d'une complexité importante, due aux différentes méthodes utilisées mais surtout à la manière dont elles interagissent. Dans le cas du modèle XGB, la complexité provient de sa composition de plusieurs arbres décisionnels.

De plus, les résultats du modèle XGBoost prédisent uniquement un taux d'incidence, ligne par ligne, pour chaque individu de notre base d'entrée. Dans l'optique de tarifer de nouveaux assurés, cela risque de poser problème. En cas de contrôle ou d'audit, un modèle construit avec cet algorithme reste trop peu explicatif.

En définitive, grâce à la double comparaison sur la performance et l'interprétabilité des modèles, c'est naturellement la solution Akur8 qui devient le meilleur compromis. Ce modèle est donc choisi pour quantifier les impacts de la crise sanitaire.

En outre, pour le modèle qui se base sur les données antérieures à la crise sanitaire, c'est également à l'aide de l'outil Akur8 que les modélisations vont être effectuées.

4.3 Analyse des résultats

4.3.1 Remise en contexte

Pour rappel, l'objectif de ce mémoire est de comparer deux bases de données. La première est relative aux données observées durant les années 2018 et 2019 réunies. La seconde concerne les observations en période de crise sanitaire, soit en 2020.

L'incidence moyenne de la base antérieure au Covid est égale à 43%. C'est-à-dire, qu'un salarié présent toute l'année aura en moyenne 0,43 arrêt par an. Durant la pandémie, l'incidence moyenne explose et atteint 55%. Cette hausse est donc de l'ordre de 28%. L'objectif est d'abord d'identifier les impacts de la crise Covid sur l'incidence des arrêts de travail, spécifiquement pour le risque incapacité. Ensuite, il est de mettre en évidence l'évolution des profils types de sinistrés, avant et pendant la crise sanitaire.

Si bien que, plusieurs questions se posent :

— Est-ce que la crise sanitaire a eu un effet sur le choix des variables explicatives introduites dans les modèles d'incidence ?

— Quels sont les profils types qui sont le plus impactés par la hausse de l'absentéisme ?

— Quels vont être les conséquences de cette sur-incidence pour les organismes d'assurance complémentaires ?

4.3.2 Construction des modèles finaux

4.3.2.1 Base 2020

Dans la suite, l'outil Akur8, retenu comme la meilleure solution d'estimation, est utilisé pour déterminer le modèle optimal. Pour cela, il est nécessaire de faire varier les hypothèses initiales et d'optimiser le rapport performance en minimisant le nombre de variables.

Pour rappel, avant toute création de modèles, un certain nombre d'informations doit être fourni. Tout d'abord, il faut renseigner la variable cible, les variables discriminantes, le type de modèle, la fonction de perte et les variables d'exposition et de stratification. Ensuite, l'outil propose différents modèles qui sont regroupés dans un « GridSearch ». Avant de choisir le meilleur modèle, une analyse préalable de ce « GridSearch » est indispensable.

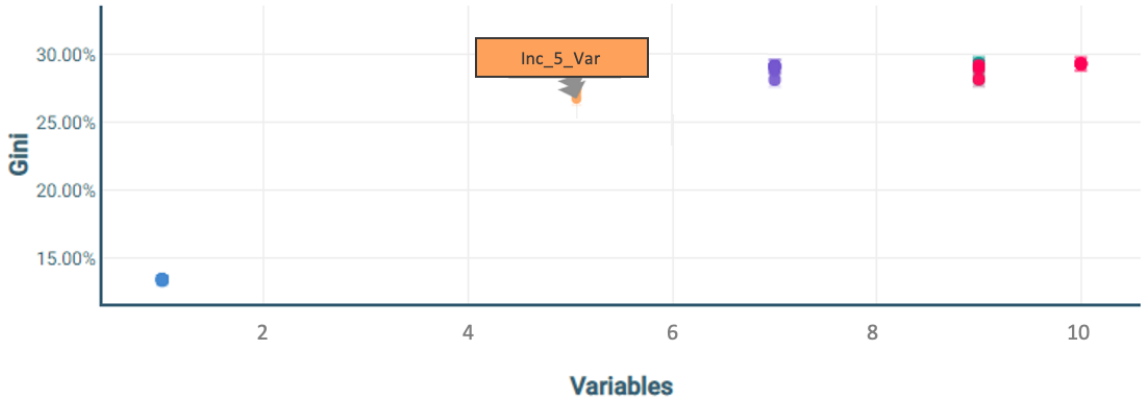


FIGURE 4.10 – « GridSearch » affichant les différents modèles construits sur la base 2020

Ce « GridSearch » présente les performances du modèle, définies par le coefficient de Gini, en fonction du nombre de variables. Une conclusion se dégage : plus le nombre de variables augmente et plus les performances sont bonnes. Cependant, au-delà de cinq variables, l'ajout de variables supplémentaires n'améliore que très peu le Gini. C'est pourquoi la combinaison optimale est donc d'intégrer cinq variables à nos modélisations.

Ces variables sont respectivement la tranche d'établissement, la tranche de salaire annuel, l'âge, la CSP et enfin le genre.

Les différents indicateurs de performance sont regroupés dans le tableau ci-dessous.

	Akur8 avec 5 var
Gini	26,67%
R^2	7,99%
RMSE	1,136
MSE	1,29
MAE	0,738

Par la suite, la courbe de Lift, la courbe de Lorenz et l'importance des variables sont présentées.

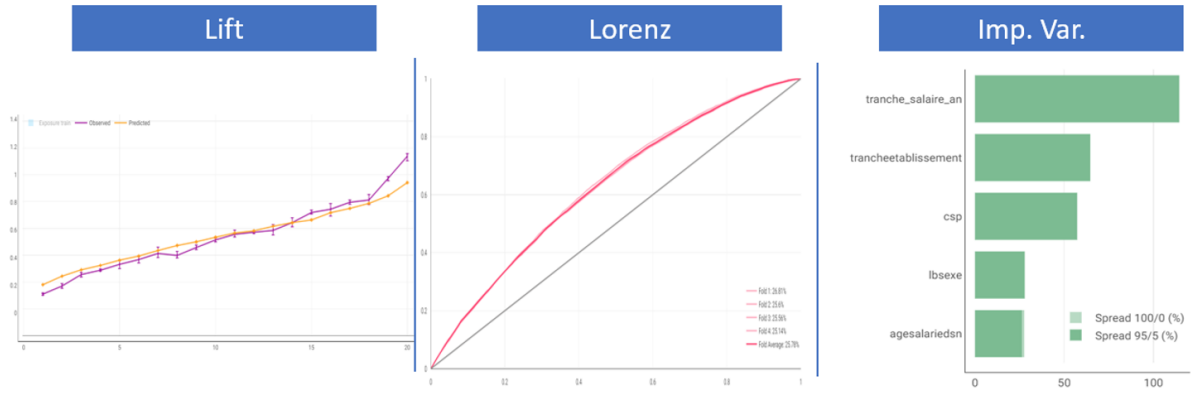


FIGURE 4.11 – Courbe de Lift, de Lorenz et importance des variables pour le modèle Akur8 à 5 variables pour l'année 2020

Tout d'abord, l'analyse se porte sur la courbe de Lift. Comme indiqué précédemment, ce modèle a tendance à ramener les prédictions vers l'incidence moyenne. En effet, il surestime les faibles incidences (les prédictions pour les salariés situés dans le groupe le plus à gauche sont supérieures aux observations) et sous-estime les fortes incidences. En majorité, les écarts de prédictions concernent les groupes de population faiblement représentés.

La courbe de Lorenz moyenne, en rose foncé, est construite en superposant les courbes de Lorenz de chaque étape de la cross validation (rose clair). Le modèle est donc très stable. Les courbes se superposent quasiment en tous points.

Par ordre d'importance, le salaire annuel des salariés influence fortement l'estimation du nombre d'arrêts. Ensuite, vient respectivement la taille de l'entreprise et la CSP, à proportion quasiment égale. Enfin, le genre et l'âge sont les dernières variables discriminantes du modèle. Leur impact est plus de trois fois inférieur par rapport aux tranches de salaire.

4.3.2.2 Base 2018/2019

De manière similaire, différents modèles sur la base de données antérieures à la crise sanitaire sont construits. L'étude du « GridSearch » présente le même dénouement que le modèle précédent. Parmi toutes les vairables retenues, le modèle le plus performant est également composé de cinq variables. De plus, ces cinq variables sont identiques avec celles de la base 2020. Les différents indicateurs de performance sont regroupés dans le tableau ci-dessous.

	Akur8 avec 5 var
Gini	27,59%
R^2	7,12%
RMSE	1,042
MSE	1,086
MAE	0,633

Comparé au modèle construit durant la période Covid, le modèle sur les données 2018 et 2019 a globalement de meilleures métriques. En effet, le coefficient de Gini augmente de presque 1%, pour atteindre 27,59%.

En ce qui concerne le RMSE, il passe de 1,136 en 2020 à 1,042 sur les années d'avant crise sanitaire, soit une amélioration de 9%.

De plus, le MAE est bien meilleur sur ce modèle puisqu'une amélioration de 14% est notable.

En revanche, le coefficient R^2 diminue et passe de 7,99%, pour le modèle retenu en période de crise, à 7,12% pour celui-ci.

Pour conclure, le modèle ci-dessus cumule la sinistralité de deux années différentes. Il semble être plus performant. En effet, se baser sur deux années consécutives permet au modèle de gagner en robustesse et en qualité. Le modèle s'ajuste mieux aux données et les prédictions sont ainsi plus proches de la réalité.

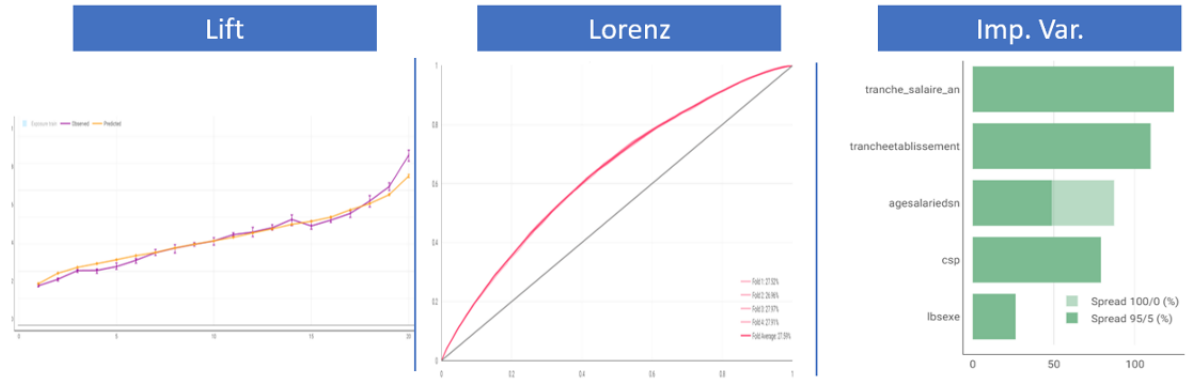


FIGURE 4.12 – Courbe de Lift, de Lorenz et importance des variables pour le modèle Akur8 à 5 variables pour les années 2018 et 2019

À l'aide de la courbe de Lift, située à gauche, les prédictions sont très proches du nombre réel d'arrêts. Des écarts importants sont uniquement présents pour les incidences extrêmes.

La courbe de Lorenz moyenne, très lisse, se superpose pour chacun des folds utilisés pour la base de test. Ce modèle segmente bien la population et s'ajuste aux données fournies en entrée.

Concentrons-nous sur l'ordre d'importance des variables. Comme pour le modèle en période de pandémie, les tranches de salaire annuel et les tranches d'établissement sont les deux variables les plus importantes du modèle. Toutefois, la variable associée au nombre de salariés par établissement est bien plus significative dans ce modèle. Elle a donc plus d'importance dans la qualité globale de ce modèle. En effet, le spread $100/0$ passe de 60% pour le modèle en période de crise à 110% pour celui-ci.

Ensuite, l'ordre d'importance des trois dernières variables varie également. Précédemment, par ordre d'importance, il y avait la CSP, le genre et l'âge avec respectivement pour valeur de spread $100/0$ 55%, 35% et 35%. Désormais, l'ordre et la valeur du spread ont changé. L'âge arrive en tête avec un spread $100/0$ de 80%, puis la CSP avec un spread de 75% et enfin le genre qui a un spread inchangé avec le modèle précédent.

Finalement, sur chacune des deux bases étudiées, le modèle le plus performant est construit. Ensuite, dans le but de répondre à la problématique de ce mémoire, l'analyse se poursuit avec l'évaluation des impacts de la crise sanitaire.

4.3.3 Comportement des variables et des coefficients associés

Malgré une sinistralité particulière durant la crise sanitaire, les deux modèles semblent relativement proches. En effet, ce sont les mêmes variables discriminantes qui impactent l'incidence du nombre d'arrêts. De plus, la comparaison des différentes métriques atteste d'écarts relativement proches.

Néanmoins, chaque variable n'a pas le même comportement vis-à-vis du risque considéré. C'est ce qui va être analysé dans cette partie. Par ordre d'importance, les coefficients associés à chaque variable sont examinés.

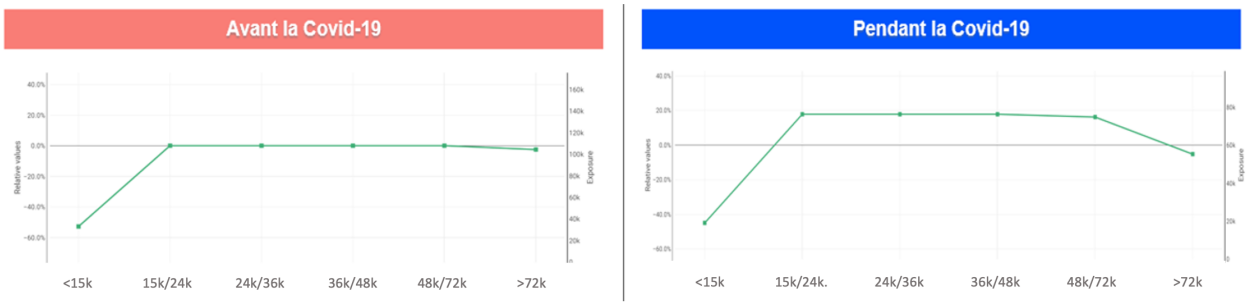


FIGURE 4.13 – Variation des coefficients associés au salaire annuel

Avant la crise sanitaire, seuls les très faibles salaires, inférieurs à 15 000€, et les hauts salaires, supérieurs à 72 000€, ont un effet sur l'incidence. En effet, les coefficients des autres modalités sont nuls.

La ligne horizontale égale à 0% représente le coefficient de base du modèle. Il est désigné par le terme « d'offset » et se caractérise par une constante propre à chaque modèle. Il peut également servir de référence au modèle. Ensuite, en fonction des caractéristiques de l'individu prédit, cet offset va être multiplié par le coefficient associé à la modalité correspondante. Il est sensiblement proche de la moyenne générale de l'incidence du portefeuille.

Attention, la forte sous-incidence des faibles salaires est à prendre avec précautions. Effectivement, pour le modèle de gauche, le coefficient associé à cette modalité est de -53%. C'est à dire que toutes choses étant égales par ailleurs, un individu qui gagne moins de 15 000€ par an fera en moyenne deux fois moins d'arrêts que le reste de la population.

Comme indiqué précédemment, cette catégorie de population est corrélée avec d'autres. Les salariés qui gagnent moins de 15 000€ annuel sont principalement employés à temps partiels. Le profil prédominant de cette catégorie de population est caractérisé par des individus de sexe féminin et

d'âge jeune. De plus, la faible exposition annuelle associée à cette catégorie de population rajoute un biais dans l'estimation.

Par conséquent, l'effet des salariés qui ont les revenus les plus modestes semble pris en compte par les coefficients d'autres variables.

En outre, pour les hauts salaires, le coefficient s'explique logiquement. Effectivement, une majorité sont des cadres. Ils occupent donc des postes qui sont physiquement moins contraignants et qui entraînent une diminution du taux d'absentéisme.

Désormais, toutes les modalités de salaires qui avaient un coefficient nul pour le modèle de gauche, n'en ont plus dans le modèle en situation de crise sanitaire. À présent, les salaires compris entre 15 000€ et 72 000€ jouent positivement sur la probabilité de tomber en arrêt. Cela veut dire que l'absentéisme de tous ces assurés a fortement augmenté et est désormais bien différent de la constante de base, durant cette année particulière.

De plus, les hauts salaires semblent plus épargnés par ce choc sur la sinistralité. En effet, durant cette année et particulièrement durant les périodes de confinement, ils ont pu bénéficier de solutions de protection, comme le télétravail par exemple. Par conséquent, ils ont été moins exposés au virus que les autres catégories de population.

Quant aux bas salaires, l'effet reste le même que pour le modèle précédent. Pour les mêmes raisons, ils ont toujours une probabilité de tomber en arrêt plus faible que les autres.

En résumé, la tendance générale de ces deux modèles est similaire. Les bas et hauts salaires ont une incidence plus faible que les autres. Néanmoins, le reste de la population accuse une forte hausse de l'incidence. Finalement, la variable parait plus impactante dans le modèle en situation de crise sanitaire.

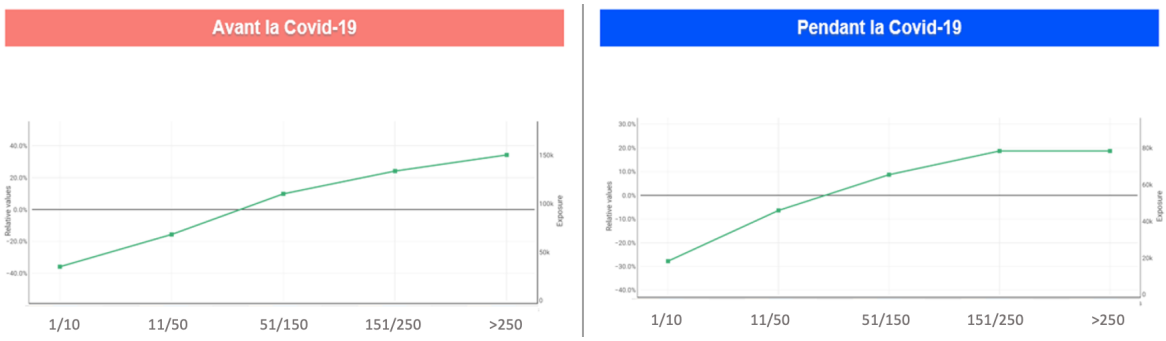


FIGURE 4.14 – Variation des coefficients associés à la taille d'entreprise

Désormais, le comportement des coefficients correspondant au nombre de salariés par entreprise va être étudié. Pour les deux modèles, la tendance est similaire. Plus l'établissement compte de salariés et plus le coefficient associé est élevé. Par conséquent, toutes choses étant égales par ailleurs, le modèle prédit une incidence qui augmente avec la taille de l'entreprise. Ainsi, la crise sanitaire ne semble pas bouleverser le comportement de certaines catégories de la population, mais plutôt impacter le portefeuille de façon homogène.

De plus, une frontière semble se dessiner autour des entreprises de taille moyenne. En effet, pour les entreprises qui comptent moins de 50 salariés, les coefficients reflètent une incidence plus faible que la base du modèle. C'est le contraire pour les entreprises de plus de 51 salariés.

En outre, les grandes entreprises de plus de 250 salariés ont vu leur coefficient se réduire. Par conséquent, elles ont été plus épargnées par l'explosion générale de l'incidence par rapport au reste de la population.

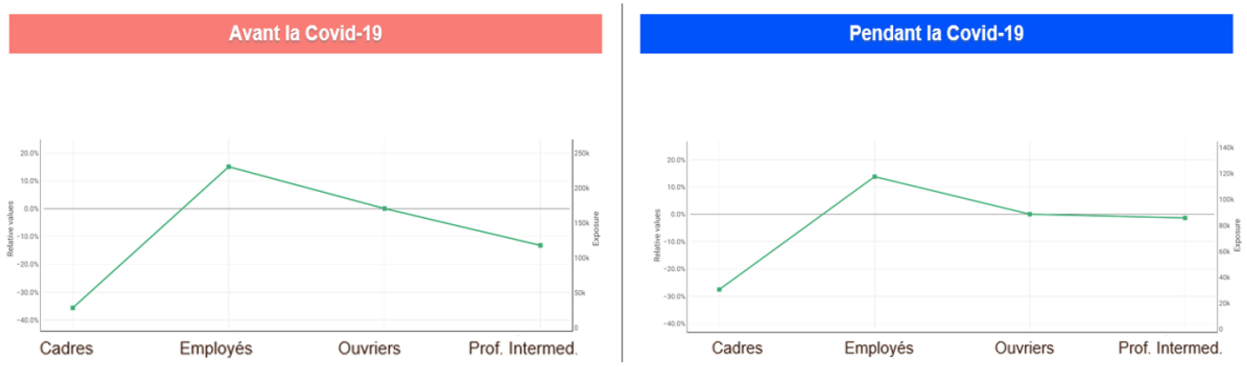


FIGURE 4.15 – Variation des coefficients associés à la CSP

Désormais, ce sont les coefficients rattachés à la CSP des salariés qui sont étudiés. Ici aussi, la tendance générale des coefficients est proche pour les deux modèles. Malgré un coefficient associé aux cadres un peu plus élevé durant la pandémie, seule la catégorie des professions intermédiaires varie nettement. En 2018 et 2019, le coefficient associé était plus faible de 12% par rapport à l'offset. Cette modalité a donc une sous-incidence par rapport à la moyenne. En 2020, cette diminution n'est plus que de 1%. Malgré la hausse moyenne de l'incidence, cette modalité est davantage affectée par la crise sanitaire, du moins plus que les autres.

Dans les deux modèles, le coefficient associé à la modalité employés reflète une forte sur-incidence, d'environ 14%, par rapport à l'incidence de base.

Enfin, la CSP regroupant les ouvriers peut être vue comme la modalité de référence. Les coefficients associés à cette modalité sont nuls dans les deux modèles. Effectivement, le fait d'être un ouvrier n'engendre pas de variation de l'incidence estimée.

Finalement, les tendances sont restées les mêmes, excepté pour les professions intermédiaires. La crise sanitaire est donc responsable de l'augmentation du niveau d'incidence de toutes les modalités.

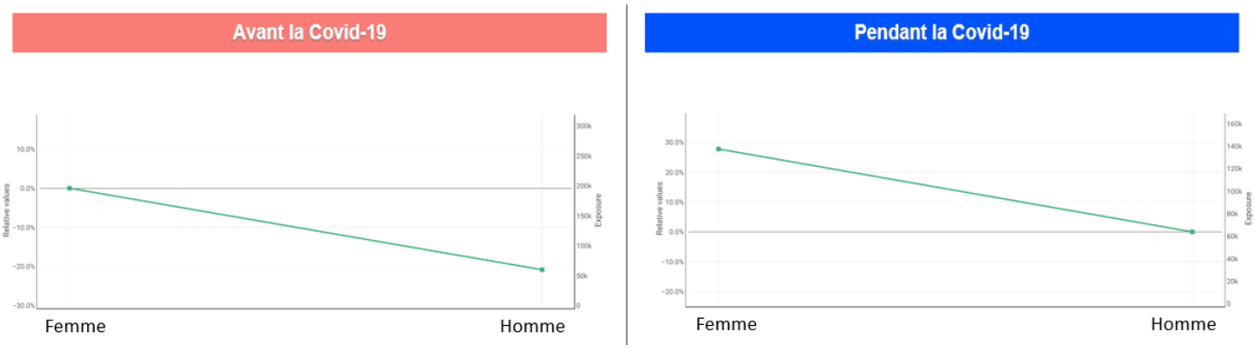


FIGURE 4.16 – Variation des coefficients associés au genre

En ce qui concerne la variable genre, les hommes ont un coefficient inférieur à celui des femmes. Toutefois, l'écart d'incidence entre ces deux modalités n'est pas identique. En effet, avant la Covid-19, la sur-incidence des femmes avoisine les 20% par rapport aux hommes. En ce qui concerne le

modèle sur 2020, la sur-incidence des femmes augmente et atteint désormais presque 30%. Par conséquent et toutes choses étant égales par ailleurs, cela veut dire que la crise est venue impactée dans des proportions plus importantes les salariées de sexe féminin.

À travers cet exemple sur le genre des individus, la différence des offset entre ces deux modèles semble claire et évidente.

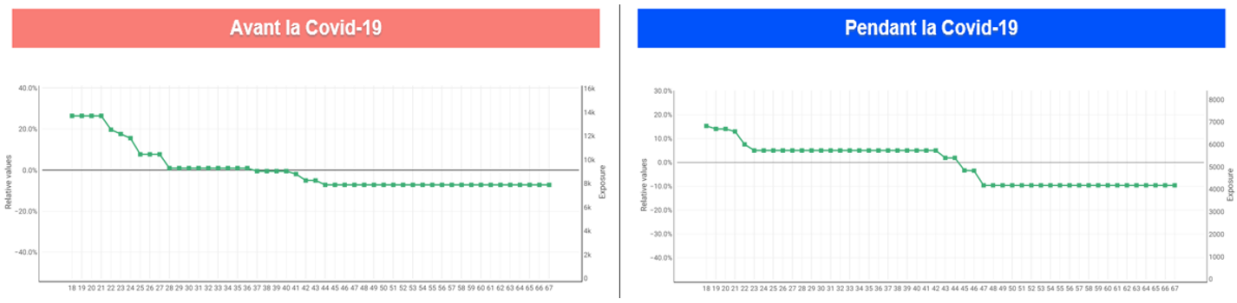


FIGURE 4.17 – Variation des coefficients associés à l'âge

Enfin, pour les coefficients liés à l'âge, une différence entre les deux modèles apparaît. Malgré une tendance générale à la décroissance, les effets ne sont pas les mêmes. Pour le modèle avant crise, les moins de 21 ans ont une sur-incidence qui explose. En effet, lorsqu'en moyenne un salarié lambda est victime d'un arrêt de travail par an, un individu âgé de moins de 21 ans en fera 1,3. Ensuite, les âges compris entre 28 et 40 ans ont des coefficients nuls. Le fait d'appartenir à cette tranche d'âge n'entraîne pas de modification supplémentaire du taux d'incidence. Enfin, pour les âges supérieurs à 40 ans, les coefficients sont négatifs et proches de - 5%.

Les salariés très jeunes ont moins de qualifications et occupent donc des postes qui peuvent être durs physiquement. Cette pénibilité est due aux positions douloureuses, aux charges portées ou encore aux gestes répétitifs. Cela occasionne donc une sur-incidence des arrêts de travail. En revanche, même si les plus âgés ont une santé plus fragile, ils occupent majoritairement des postes de bureaux qui mettent moins en péril leur état physique. De plus, malgré que les plus âgés ont un taux d'incidence plus faible que les plus jeunes, ils s'arrêtent plus longtemps en moyenne.

Le modèle durant la Covid-19 repose sur trois paliers de coefficients. En effet, lui aussi capte la sur-incidence des plus jeunes, mais dans des proportions plus faibles. Désormais, les salariés âgés entre 22 et 42 ans sont associés à un coefficient positif, égal à 5%. Par conséquent, être dans cette tranche d'âge entraîne une hausse du nombre d'arrêts annuel. Le modèle prend bien en compte la bosse de sur-incidence observée durant l'année 2020. Enfin, les salariés âgés de plus de 47 ans ont également un coefficient négatif, mais qui vaut désormais -10%.

Néanmoins, la forte sous-incidence des salariés les plus âgés, plus de 60 ans, observée dans nos deux bases de données n'est pas prise en compte dans ces estimations. La faible proportion d'individus qui compose cette tranche ne permet donc pas de conclure à une sous-incidence significative.

Pour conclure cette partie sur l'étude des coefficients, ces deux modèles segmentent très bien les observations. Pour toutes les variables, la majorité des modalités suit l'évolution de l'incidence réelle, dans des proportions quasi similaires. Ceci confirme donc que les deux modèles retenus sont très performants.

De plus, les valeurs des métriques présentées précédemment caractérisent de bons modèles prédictifs. En effet, on ne savait pas si un coefficient de Gini proche de 27% et un R^2 aux alentours de 7% allaient être suffisants. Désormais, ces résultats affirment que c'est bien le cas.

En revanche, l'étude du comportement de ces coefficients a permis de mettre en évidence quelques limites de ces modèles. Effectivement, certaines modalités peuvent être corrélées les unes aux autres. De plus, l'offset varie d'un modèle à l'autre si bien qu'il semble difficile de percevoir avec exactitude l'évolution des coefficients.

4.3.4 Application de la loi d'incidence

Une fois l'étude du comportement des coefficients effectuée, une application de la loi d'incidence, provenant des modèles Akur8, va être élaborée. En effet, cela va permettre de mettre en évidence les conséquences de la crise sanitaire sur l'incidence. Au préalable, un profil type du portefeuille est choisi pour servir d'exemple dans cette application.

Pour illustrer les résultats des modélisations, les différents coefficients obtenus permettent la mise en place de la loi d'incidence. Ainsi, les conclusions de nos deux modèles, avant et pendant crise sanitaire, sont adaptées sur un profil qui possède les caractéristiques suivantes : il s'agit d'une femme, âgée de 40 ans, qui est employée d'une grande entreprise de plus de 250 salariés et qui perçoit un salaire annuel fixe de 42 000€.

Les résultats des prédictions du premier modèle, basé sur les données antérieures à la Covid-19, sont affichés ci-dessous. Pour chacune des modalités, un coefficient est associé comme suit.

Offset		0,509										
agesalariedsn			CSP			lbsexe			tranche_salaire_an			tranche_entreprise	
18	1,26		Cadres	0,662		Femme	1		<15k€	0,499		1 à 10 salariés	0,666
19	1,251		Employés	1,135		Homme	0,818		[15k€;24k€[	1		11 à 50 salariés	0,848
20	1,271		Ouvriers	1					[24k€;36k€[	1		51 à 150 salariés	1,096
21	1,297		Prof. Inter.	0,884					[36k€;48k€[	1		151 à 250 salariés	1,231
22	1,194								[48k€;72k€[	1		+250 salariés	1,307
23	1,172								>72k€	0,989			
24	1,143												
25 -> 27	1,091												
28 -> 36	1,015												
37 -> 40	0,997												
41	0,982												
42	0,958												
43	0,962												
44	0,925												
45 -> 67	0,919												

FIGURE 4.18 – Présentation des coefficients pour la construction de la loi d'incidence sur 2018/2019

Ensuite, les résultats du modèle en période de crise sanitaire sont exposés.

Offset		0,451									
agesalariedsn			CSP			lbsexe			tranche_salaire_an			tranche_entreprise	
18	1,152		Cadres	0,724		Femme	1,278		<15k€	0,579		1 à 10 salariés	0,721
19	1,14		Employés	1,138		Homme	1		[15k€;24k€[	1,178		11 à 50 salariés	0,936
20	1,14		Ouvriers	1					[24k€;36k€[	1,178		51 à 150 salariés	1,097
21	1,129		Prof. Inter.	0,986					[36k€;48k€[	1,178		151 à 250 salariés	1,188
22	1,075								[48k€;72k€[	1,167		+250 salariés	1,189
23 -> 42	1,049								>72k€	0,947			
43	1,018												
44	1,018												
45	0,965												
46	0,964												
47 -> 67	0,903												

FIGURE 4.19 – Présentation des coefficients pour la construction de la loi d'incidence sur 2020

D'abord, il est important de constater que le coefficient de base, défini par l'offset, est différent dans ces deux modèles. Cet offset s'interprète comme une constante qui vient à être déformée en fonction des modélisations. En effet, il représente un certain individu de référence. Cependant, suivant les modèles, les caractéristiques qui définissent cette référence ne sont pas les mêmes. Pour le modèle avant crise, le taux d'incidence de base est plus important. D'un côté il est égal à 50,9% et de l'autre 45,1%. Par conséquent, les caractéristiques qui définissent cette base ne sont pas les mêmes et entraînent une hausse du nombre de sinistres dans le premier modèle.

Pour récupérer le taux d'incidence tête par tête, représenté par la lettre π, du profil défini précédemment, il faut donc multiplier l'ensemble des coefficients associés à ses caractéristiques, de la manière suivante.

$$\pi = \beta_{offset} \times \beta_{age} \times \beta_{CSP} \times \beta_{genre} \times \beta_{salaire} \times \beta_{entreprise}.$$

Ainsi pour le modèle construit sur les années d'observation 2018 et 2019, le taux d'incidence se calcule de la manière suivante.

$$\pi_{18/19} = 0,509 \times 0,997 \times 1,135 \times 1,000 \times 1,000 \times 1,307 = 75,3\%.$$

Le taux d'incidence de cette assurée est donc de 75,3%. Ainsi, le profil sélectionné présente des caractéristiques qui favorisent la hausse du nombre d'arrêts. En effet, durant cette période d'observation, le taux d'incidence moyen du portefeuille est de 43%. Ce profil particulier accuse une sur-incidence de prêt de $\frac{75,3\%}{43\%} - 1 = 75\%$.

De manière similaire, le calcul du taux d'incidence est effectué sur ce même profil à l'aide des résultats du modèle en période de pandémie.

$$\pi_{20} = 0,451 \times 1,049 \times 1,138 \times 1,278 \times 1,178 \times 1,189 = 96,4\%.$$

Comme pour le modèle précédent, le taux calculé est bien supérieur à l'incidence moyenne du portefeuille, égale à 55% en période de crise. Ici aussi, la sur-incidence constatée gravite autour des $\frac{96,4\%}{55\%} - 1 = 75\%$. Cela confirme donc que le profil type retenu pour cette analyse présente des caractéristiques qui entraînent une forte hausse du nombre d'arrêts. De plus, l'écart avec la moyenne

du portefeuile est identique pour les deux périodes d'études. Par conséquent, l'impact de la crise sanitaire semble être le même entre ce profil particulier et l'ensemble du portefeuille d'assurés.

En outre, pour un individu qui possède exactement les mêmes caractéristiques, le taux d'incidence en période de pandémie est $\frac{96,4\%}{75,3\%} - 1 = 28\%$ supérieur aux années précédentes. À travers cet exemple, les conséquences de ce virus sur la probabilité de tomber en arrêt sont clairement identifiables.

4.3.5 Interprétation générale

Les différentes approches proposées apportent chacune des informations complémentaires. Elles permettent de mieux connaître la structure du portefeuille et l'influence de variables explicatives sur l'entrée en incapacité. De plus, ces informations fournissent des outils décisionnels intéressants pour mieux appréhender ce risque en période de crise sanitaire.

Dans un premier temps, il est important de constater qu'inclure seulement cinq variables permet d'aboutir à des prédictions fidèles à la réalité. Les modèles avec plus de variables n'apportent pas d'informations significativement meilleures. Leurs indicateurs de performance sont quasiment similaires à ceux du modèle retenu.

Ensuite, malgré la sur-incidence de 28% causée par la crise sanitaire, les variables discriminantes des deux modèles sont identiques. Même si leur pouvoir prédictif n'est pas le même, ce sont bien le salaire annuel, la taille de l'entreprise, la CSP, l'âge et le genre des salariés qui influent sur la variable cible.

En définitive, la pandémie a impacté l'ensemble du portefeuille. Elle a accentué la dérive du risque prévoyance collective, notamment avec la mise en place d'arrêts dérogatoires (personnes à risques, cas contacts, garde d'enfants ...). De plus, les individus de tous les secteurs d'activité, de tout âge, et de toutes caractéristiques ont donc vu leur incidence croître durant cette année particulière.

En revanche, même si toute la population a subi cette crise de plein fouet, certaines catégories de population ont été plus exposées que d'autres à ce virus. Leur influence dans les modèles s'est donc renforcée. C'est le cas pour les modalités suivantes :

- Les assurés qui ont un revenu annuel compris entre 15 000€ et 48 000€. Dans le modèle en période de crise sanitaire, les coefficients associés à ces modalités forment une bosse de sur-incidence.

- Comme le montre les résultats des différents modèles, les professions intermédiaires sont également plus touchées que le reste de la population. En effet, cette catégorie se compose principalement de professeurs des écoles, de personnel travaillant dans la santé ou dans le social. Toutes ces personnes sont en contact permanent avec de nombreux individus. Ils sont donc plus enclins à recevoir et transmettre des virus, qui somme toute sont très contagieux.

- Tous les salariés âgés entre 22 et 42 ans sont aussi heurtés par cette crise. Beaucoup d'entre eux ont sous-évalué la gravité de ce virus et ne se sont donc pas assez protégés. Ils ont également continué à rencontrer du monde, ce qui a favorisé la propagation de la Covid-19. De plus, ce sont eux qui ont dû utiliser le plus les arrêts dérogatoires mis en place par la gouvernement.

En revanche, d'autres catégories semblent avoir été plutôt épargnées par cette pandémie. C'est le cas pour les modalités suivantes :

- Les assurés d'entreprise de plus de 250 salariés ont effectivement été moins impactés que le reste de la population. Dans les grosses structures, des solutions de travail à distance ont été plus rapidement déployées. Leur couverture complémentaire est généralement plus intéressante.

- Les cadres ont également été moins touchés par cette hausse d'incidence. Eux aussi ont pu bénéficier de solutions comme le télétravail. Leur métier est vraisemblablement moins éprouvant que d'autres.

- Contre les aprioris initiaux, les salariés âgés de plus de 48 ans ont eux aussi pu échapper un peu plus à cette crise. Malgré un virus plus virulent sur une population âgée ou fragile, les modèles construits prédisent le contraire. Ces salariés ont probablement été plus attentifs et se sont davantage protégés.

Finalement la crise sanitaire a de multiples répercussions sur les salariés. Leur santé physique et psychologique s'est dégradée. Leur engagement au sein de l'entreprise a diminué et le turn-over a quant à lui augmenté. En outre, nombreux sont ceux qui remettent en question leur vie professionnelle. En sortie de crise sanitaire, beaucoup de reconversions ont été constatées.

Finalement, cette crise a accentué la dérive du risque prévoyance et plus particulièrement de l'incapacité, notamment avec la mise en place d'arrêts dérogatoires tels que les arrêts pour garde d'enfant, pour garde de personnes à risque ou les arrêts pour cas contact. En France, la hausse des coûts, en lien avec cette crise sanitaire, avoisine les 20% pour l'ensemble des organismes assureurs. Néanmoins ce n'est pas la seule conséquence, puisqu'en plus de la hausse des prestations versées, le comportement des assurés varie énormément. C'est pourquoi l'émergence de nouveaux risques couplée au boulversement de la sinistralité nécessitent de revoir les modèles de calculs, les provisions faites, etc. En outre, cela provoque un changement du principe de mutualité, qui se concrétise par une hausse des prix des couvertures d'assurance.

Aujourd'hui, les organismes assureurs ont eux aussi modifié leur fonctionnement pour répondre aux besoins des assurés. En effet, diverses innovations autour de la digitalisation et de l'amélioration des systèmes de gestion, qui passe par la dématérialisation des processus, sont constatées.

Malgré tout, cette étude présente quelques limites. La période de crise sanitaire n'est pas représentative d'une année normale en termes de d'arrêts de travail, notamment à cause des conditions d'octroi et des arrêts dérogatoires. Par conséquent, baser un processus de refonte de normes tarifiare sur ces données ne pourrait pas être appliqué dans la réalité. Néanmoins, tous ces travaux ont servi de point de départ pour la construction de nouvelles normes tarifaires, qui se reposeront sur des années de sinistralité post-covid. De plus, afin d'aboutir à de nouvelles normes, la construction d'une loi de maintien en incapacité devrait être modélisée.

Conclusion

L'apparition du virus de la Covid-19 a bouleversé le quotidien des français mais également celui des organismes assureurs. Ils ont dû faire face à des défis économiques et sociaux émergents. En particulier, cette pandémie a engendré une hausse brutale de l'absentéisme. De plus, de nombreux changements sont apparus. En effet, la pandémie a fait évoluer les attentes des assurés concernant leur protection et donc leur couverture de prévoyance complémentaire. C'est donc ce qui nous a motivés à faire une étude approfondie de la sinistralité sur l'année 2020.

Aujourd'hui, les français se soucient davantage de leur couverture face aux différents risques sociaux. Leur besoin de se sentir correctement protégés s'est renforcé. Cependant, le prix reste encore un facteur prédominant dans leur choix de couverture. Fortement poussé par les attentes des jeunes générations, le secteur de l'assurance a su profiter de cette crise pour améliorer son développement et optimiser sa digitalisation. Le but est donc de simplifier l'accès à l'information, de renforcer le suivi et les conseils personnalisés mais surtout de réduire les temps de déclaration ou de gestion des dossiers. Désormais, nombreux sont les assureurs qui proposent des services 100% digitaux. Ces assureurs essayent de mieux comprendre les profils de risques, dans le but de proposer des solutions adaptées aux besoins des assurés. Ceci est rendu possible grâce à la DSN.

En effet, la mise en place de la Déclaration Sociale Nominative a permis de corriger certaines limites des systèmes d'informations utilisés auparavant en prévoyance collective. Avant 2017 et la mise en place de manière obligatoire de la DSN, les compagnies d'assurance ne collectaient que les informations sur les salariés sinistrés. En outre, dans les bases de données, ne sont présentes que les informations sur les sinistrés ayant eu un arrêt d'une durée supérieure à la franchise. Celle-ci est définie à la signature du contrat. Pour construire des modèles ou refaire des normes tarifaires, il était donc nécessaire de se restreindre au périmètre des assurés sinistrés post franchise.

La généralisation de la DSN à toutes les entreprises est donc un grand bouleversement dans le monde de la prévoyance. Désormais, les flux de données sont reçus mensuellement. Ces informations concernent aussi bien les assurés sinistrés, d'une durée inférieure ou supérieure à la franchise contractuelle, que les individus non sinistrés. Désormais, les bases disponibles sont beaucoup plus précises et bien mieux représentatives de la population sous risque. Toutes les modélisations effectuées au cours de ce mémoire sont construites à l'aide des données de la DSN. C'est pourquoi seules les années 2018 et 2019 peuvent être utilisées pour caractériser l'incidence avant crise.

À l'aide d'une étude sur le suivi de portefeuille, réalisée sur la base de plusieurs indicateurs de sinistralité, l'évolution de l'absentéisme est principalement portée par la forte hausse du nombre d'arrêts. Les autres indicateurs comme la durée moyenne des arrêts, l'âge moyen à la survenance ou encore les indemnités journalières versées, sont restés relativement stables. De plus, l'année de sinistralité 2020 s'est affichée bien trop spécifique pour représenter l'absentéisme des années futures. En effet, cette année est plutôt considérée comme une parenthèse dans le temps. Le monde de l'assurance a désormais conscience que toutes les lois construites à partir de données basées sur l'année 2020, ne peuvent servir de référence sur le marché, même dans le cas d'une nouvelle pandémie future.

La conclusion de cette étude a permis d'affiner les objectifs de ce mémoire afin de se focaliser uniquement sur la construction de modèles prédictifs de l'entrée en incapacité, durant la période de crise sanitaire. De plus, toutes les modélisations effectuées serviront à analyser et à justifier les évolutions de ce risque. Le but est de mieux comprendre le virus de la Covid-19, et de savoir

comment il a impacté notre portefeuille.

Une fois que les objectifs étaient bien fixés, différents modèles de Machine Learning, avec des complexités plus ou moins grandes, ont été confrontés. Le but de cette comparaison est de ne retenir que le modèle le plus performant. Des modèles GLM ont donc été développés, suivis des modèles XGBoost et enfin des modèles réalisés grâce à une solution spécialisée pour la tarification, nommée Akur8. Avec l'aide de différentes métriques qui mesurent la performance des modèles, la solution Akur8 s'est montrée la meilleure. En effet, ces modèles utilisent l'intelligence artificielle dans le but de combiner des GLM pénalisés (aussi appelés modèles GAM) avec la théorie de la crédibilité. Les modèles issus de cet outil sont un peu moins performants que ceux basés sur l'algorithme XGBoost. Cependant, ils sont plus interprétables, plus robustes et plus auditables pour un risque complexe comme la prévoyance.

Ensuite, les modèles les plus performants ont été construits sur la base 2020 ainsi que sur les données avant crise sanitaire. Les variables significatives de ces deux modèles et leur importance ont été analysées. Ce sont principalement les mêmes variables qui influencent le nombre d'arrêts de nos assurés. Cela prouve que la pandémie a touché l'intégralité de la population. Bien que ce soient les personnes âgées ou fragiles qui ont le plus souffert de ce virus, sur une population d'actifs la Covid-19 a impacté l'intégralité des salariés du portefeuille. En moyenne sur ce portefeuille, le taux d'incidence est de 55% en 2020, soit une augmentation de plus de 25% comparée aux années précédant la crise. En revanche, au cours d'une analyse modalité par modalité de chaque variable, la crise n'a pas impacté de manière homogène la population. En effet, certaines catégories, comme la CSP « Professions Intermédiaires », ont majoritairement été touchées. Ce n'est pas anodin puisque cette catégorie correspond aux salariés qui ont le plus été exposés au virus. Elle se compose notamment de professeurs des écoles, de personnel travaillant dans le médical ou dans le social.

Finalement, ces travaux présentent quelques limites. Au cours de l'avancée de ce projet, l'absentéisme en période de Covid s'est révélé bien trop atypique. En effet, refaire une norme tarifaire en se basant sur cette année n'a pas de sens, même pour prévenir d'une future crise sanitaire ou autres chocs de la sinistralité. De plus, la sinistralité des années post pandémie ne semble pas revenir à l'identique comparée à celle des années avant la crise sanitaire. En effet, de sérieux changements de comportement sont constatés. Aujourd'hui, l'absentéisme des années après crise sanitaire est stable mais il est bien plus élevé que celui des années 2018 et 2019. Par conséquent, les travaux effectués sur l'année 2020 vont être dupliqués sur les années 2021, voire 2022. L'objectif est donc de refaire des normes tarifaires des différents risques en prévoyance, en se basant sur une sinistralité stable.

Bibliographie

Akur8 [2021] : The IA Pricing Company : "Tutorial for AG2R".

Ayming / AG2R La Mondiale [2021] : "13eme baromètre de l'absentéisme et de l'engagement".

C. Bentéjac et al. [2021] : "A Comparative Analysis of XGBoost". In : College of Science and Technology, University of Bordeaux.

M. Casotto et al. [2019] : "Credibility and Penalized Regression", pour Akur8 France.

J. CAUX [2017] : "Modélisation de l'entrée en incapacité de travail en prévoyance collective". Mémoire d'actuaire, ISFA

T. Chen et C. Guestrin [2016] : "XGBoost : A Scalable Tree Boosting System". in : KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 785-794.

P. De Jong et G. Z. Heller [2008] : "Generalized linear models for insurance data". In : Cambridge University Press.

J. Garrido et al. [2016] : "Generalized linear models for dependent frequency and severity of insurance claims". In : Insurance : Mathematics and Economics, vol. 70, p. 205-215.

B. Harouna Abassi [2020] : "Modélisation de l'incidence en incapacité sur un portefeuille de prévoyance individuelle". Mémoire d'actuaire, ISUP.

D. Henoc Akaffou [2020] : "Méthode alternative de tarification santé : GLM/XGBoost". Mémoire d'actuaire, EURIA.

Roel Henckaerts et al. [2020] : "Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods". In : North American Actuarial Journal, 25(2), p. 255–285.

P. M. Mbow [2020] : "Calibration de lois d'incidence en incapacité". Mémoire d'actuaire, ISUP.

J. A. Nelder et R. W. M. Wedderburn [1972] : "Generalized Linear Models". In : Journal of the Royal Statistical Society, Series A (General), Vol. 135, No. 3, p. 370-384.

Site d'information *"économie.gouv.fr"* :
`https://www.economie.gouv.fr/entreprises/declaration-sociale-nominative-dsn#`.

Site d'information *"Vie-publique.fr"* :
`https://www.vie-publique.fr/fiches/protection-sociale`.

Table des figures

Annexe A

Complément de l'étude sur les indicateurs de sinistralités :

Dans cette partie, la fin de l'étude présentée au cours de la partie 1.4, concernant les cinq indicateurs de sinistralité, est détaillée. Tout d'abord, l'évolution des indemnités journalières depuis 2015 est introduite :

IJ	fin Sept N	fin Sept N+1	fin Sept N+2	fin Sept N+3	fin Sept N+4	fin Sept N+5	fin Sept N+6
2015	16,26 €	14,51 €	14,26 €	14,27 €	14,30 €	14,30 €	14,31 €
2016	15,22 €	14,37 €	14,08 €	14,05 €	14,07 €	14,09 €	
2017	15,32 €	14,53 €	14,22 €	14,22 €	14,22 €		
2018	15,58 €	14,72 €	14,33 €	14,27 €			
2019	15,96 €	15,20 €	14,81 €				
2020	17,79 €	15,80 €					
2021	17,08 €						

FIGURE 4.20 – Triangle de développement des IJ, (en €)

De manière similaire à l'étude des prestations versées, du nombre de sinistres ou encore de la durée moyenne des arrêts, la première année de développement n'est pas assez consolidée et ne permet pas une analyse significative du risque arrêt de travail. Cependant, pour les années de développement suivantes, les IJ sont relativement constantes par année de survenance. En effet, pour les arrêts survenus en 2015, la moyenne des IJ versées est d'environ 14,3€ par jour et par salarié. En 2019, cette moyenne se situe plus autour de 15€ par jour et par salarié. En effet, comme expliqué dans la partie 1.4 et selon des études de l'INSEE, les salaires sont revalorisés chaque année, pour faire face notamment à l'inflation.

Au cours des six dernières années, les IJ versées ont nettement augmenté, de quasiment 2€ par jour et par salarié, mais cela ne semble pas être une conséquence de la crise sanitaire, qui frappe notre pays depuis mars 2020.

Ensuite, l'âge moyen des salariés lors de la survenance d'un sinistre est analysé :

Age	fin Sept N	fin Sept N+1	fin Sept N+2	fin Sept N+3	fin Sept N+4	fin Sept N+5	fin Sept N+6
2015	39,6	39,4	39,5	39,6	39,6	39,6	39,6
2016	40,6	39,9	40,0	40,0	40,0	40,0	
2017	40,4	40,2	40,2	40,2	40,2		
2018	40,7	40,0	40,0	40,0			
2019	40,3	39,8	39,8				
2020	39,3	39,8					
2021	39,7						

FIGURE 4.21 – Triangle de développement de l'âge moyen à la survenance

L'âge moyen à la survenance semble être un indicateur relativement stable au cours du temps, puisque depuis 2015, il ne cesse d'osciller entre 39 et 40 ans. Cette évolution s'explique majoritairement par la composition du portefeuille et ne semble avoir aucun lien avec la crise sanitaire.

Annexe B

Complément de l'étude des statistiques descriptives faites variable par variable :

Répartition des salariés par genre

Le graphe suivant analyse la répartition entre les hommes et les femmes dans le portefeuille considéré.

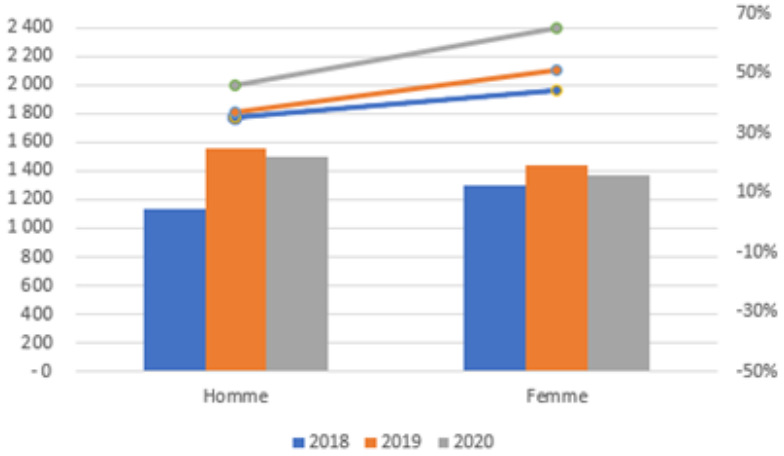


FIGURE 4.22 – Répartition des assurés en fonction de leur genre

S'agissant de l'année 2018, le portefeuille se compose d'une majorité de femmes. Toutefois, cette majorité s'inverse à partir de 2019. La forte augmentation du nombre d'hommes entre 2018 et 2019 s'explique par l'apport d'affaires nouvelles, avec notamment l'intégration d'une nouvelle CCN principalement composée de salariés masculins. Par ailleurs, le nombre d'assurés présents dans le portefeuille diminue entre 2019 et 2020, dans des proportions identiques entre les femmes et les hommes.

Pour rappel, lors des modélisations nous allons utiliser deux bases. Une première sur les années 2018 et 2019 et une seconde sur l'année 2020. Au global, dans la première base, la répartition entre les sexes s'équilibre, 50% pour chaque modalité. Dans la seconde, il y a 52,3% d'hommes et 47,7% de femmes.

D'un point de vue réglementaire, il n'est pas possible de différencier les tarifs d'un contrat de prévoyance en fonction du genre de l'assuré. En revanche, force est de constater que les femmes ont une incidence plus élevée que celle des hommes. En effet, en 2018 comme en 2019 le taux d'incidence des hommes est de 36%. Par opposition, celui des femmes est respectivement égal à 44% en 2018 et à 51% en 2019. En nombre, les femmes ont donc tendance à être victimes de plus d'arrêts de travail.

De plus, la pandémie entraîne une hausse significative de l'incidence. Néanmoins, cette augmentation semble similaire en fonction du genre. Chez les hommes, le nombre de nouveaux arrêts moyen durant l'année atteint 46%. Pour les femmes, ce nombre est désormais proche de 65%. En effet, la hausse de l'incidence est de l'ordre de 28%, quel que soit le genre du salarié.

Pour rappel, la probabilité qu'un salarié du portefeuille tombe en arrêt durant l'année 2020 est de 26%. Par conséquent, d'après la définition de l'incidence, cela revient à dire que les femmes qui ont au moins un arrêt durant cette année en font en moyenne 2,5. En ce qui concerne les hommes, leur nombre d'arrêts n'est que de 0,8 par an.

Une analyse de l'exposition est désormais présentée.

	Homme	Femme	**Moyenne**
2018 + 2019	73%	76%	**75%**
2020	74%	77%	**75%**

En regroupant les années 2018 et 2019, l'exposition moyenne est de 75%. Pour rappel, elle était respectivement de 78% pour 2018 et de 72% pour 2019. Une exposition de 75% veut donc dire que les salariés sont présents en moyenne durant les trois quarts de l'année, soit pendant 9 mois par an. Durant la crise sanitaire, l'exposition moyenne du portefeuille reste identique.

De plus, pour les deux bases considérées, les hommes ont un taux d'exposition légèrement plus faible que celui des femmes. En résumé, les mouvements de personnel au sein des entreprises touchent plus fréquemment les hommes.

Répartition des salariés en fonction de leur statut

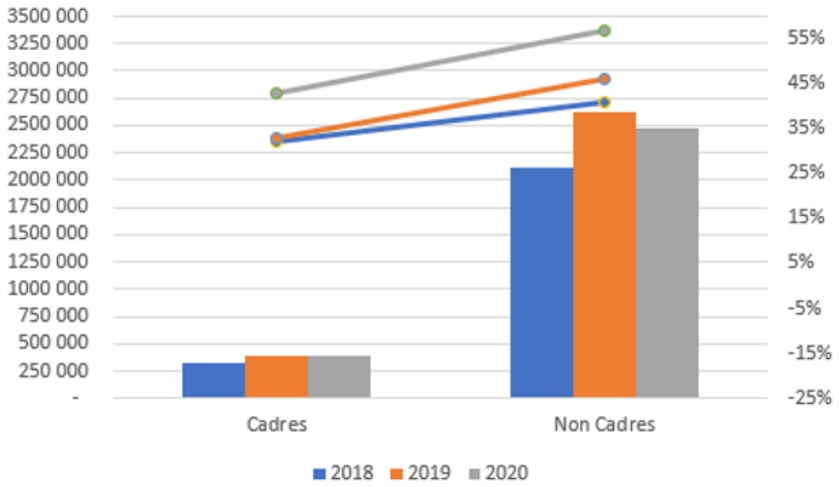


FIGURE 4.23 – Répartition des assurés en fonction de leur statut Cadre / Non-Cadre

La répartition des salariés en fonction du statut de leur contrat de travail est très hétérogène. Une extrême majorité des salariés sont « Non-Cadres ». Ils représentent environ 90% du portefeuille, contre 10% pour les salariés « Cadres ». L'évolution du nombre de cadres au cours du temps est plutôt stable, malgré une légère augmentation en 2019. En revanche, le nombre de salariés non-cadres varie dans des proportions plus importantes. Une augmentation de 20% est observable en 2019, et une réduction de 6% à partir de 2020.

Ensuite, l'incidence des cadres est nettement plus faible que celle des non-cadres. Il y a environ 10% d'écart entre ces deux modalités. Comme pour la CSP « Cadre », les caractéristiques et les métiers associés à cette modalité ne permettent pas aux salariés d'avoir une forte incidence (emploi moins contraignant, poste clé de l'entreprise, gestion d'équipes, ...).

	Cadres	Non-Cadres	**Moyenne**
2018 + 2019	92%	72%	**75%**
2020	93%	71%	**75%**

La présence annuelle en entreprise des cadres est bien supérieure à celle des non-cadres, quelle que soit l'année d'observation. De manière similaire avec la CSP « Cadre », l'exposition annuelle est très élevée, plus de 11 mois par an en moyenne. Néanmoins, les salariés non-cadres font diminuer l'exposition, puisqu'ils ne sont présents en entreprise que 8 mois et demi par an.

Répartition des salariés par libellé du contrat

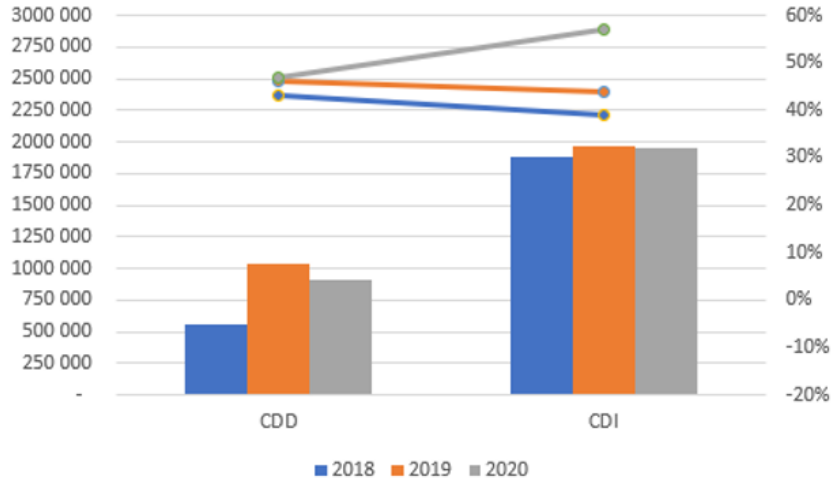


FIGURE 4.24 – Répartition des assurés en fonction du libellé de leur contrat

Ensuite, la répartition du nombre de salariés en fonction des modalités CCD ou CDI est étudiée. Premièrement, les individus en contrat à durée indéterminée représentent environ 70% de ce portefeuille. Deuxièmement, le nombre de CDI est relativement stable au fil des années. Entre 2018 et 2020, le nombre de CDI est proche de 1 850 000 individus. En revanche, le nombre de CDD varie davantage. En effet, il y a 550 000 salariés en CDD en 2018 contre plus de 1 050 000 en 2019, soit une augmentation de 90%.

De plus, l'incidence n'a pas la même allure entre les années avant et pendant crise sanitaire. Quelle que soit l'année d'observation, l'incidence des CDD est stable et proche de 45%. Il semble donc que la crise sanitaire n'ait pas clairement impacté la probabilité de tomber en arrêt des salariés en CDD. Pour les années 2018 et 2019, l'incidence des CDI est inférieure à celle des CDD et se situe aux alentours de 41%. Néanmoins, pour l'année 2020, leur incidence est supérieure, avec un taux qui approche des 57%. La hausse de l'incidence en période de crise sanitaire est donc principalement portée par les salariés en CDI.

	CDD	CDI	**Moyenne**
2018 + 2019	42%	88%	**75%**
2020	44%	90%	**75%**

Il est évident que l'exposition annuelle des salariés en CDD est bien inférieure à celle des autres. Les salariés en contrat à durée déterminée ont une exposition moyenne proche de 43%. En effet, ce chiffre est plutôt stable entre 2018/2019 et 2020. Il représente environ cinq mois de présence par an.

Depuis une vingtaine d'années, la durée moyenne des CDD semble diminuer en France. En 2000 cette durée était de 120 jours. Aujourd'hui, elle avoisine les 50 jours par contrat, soit quasiment trois mois. La présence des salariés en CDD est proche de cinq mois. Il semble donc qu'une grande partie de nos assurés voie leur premier contrat en CDD se renouveler. Ce chiffre vient contrebalancer la forte présence des CDI, puisque sur les trois années d'observation, l'exposition moyenne est de 89%, soit presque onze mois durant l'année.

Répartition des salariés par type de gestion

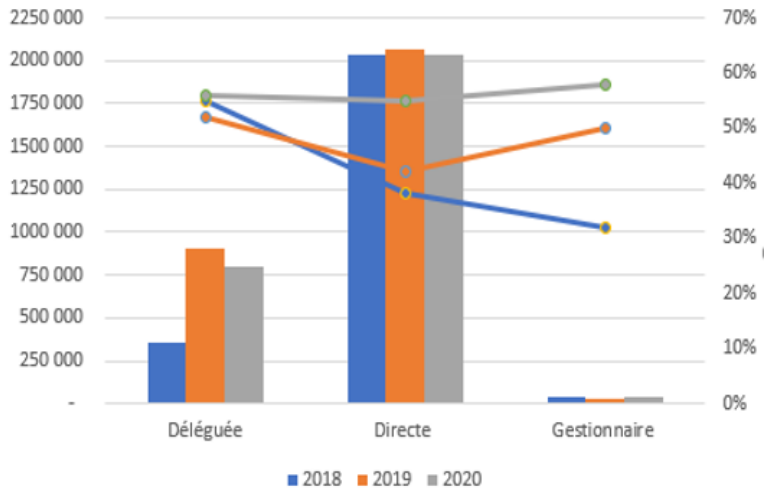


FIGURE 4.25 – Répartition des assurés en fonction du type de gestion

Une extrême majorité de salariés a un contrat de prévoyance en gestion directe, c'est-à-dire géré par AG2R La Mondiale. Pour chaque année, il y a environ 2 000 000 d'individus dans cette catégorie, soit entre 70% et 87% du portefeuille. Leur évolution est constante sur la période étudiée. La modalité « Déléguée » correspond aux contrats dont la gestion est léguée majoritairement à des courtiers. Elle représente entre 12% et 29% de ce portefeuille. En 2019, le nombre de contrats en gestion déléguée explose. Une hausse de 140% est constatée par rapport à l'année précédente. Tous les nouveaux assurés qui intègrent ce portefeuille au début de l'année 2019 semblent donc être géré par des organismes extérieurs.

Ensuite, pour certains contrats, AG2R La Mondiale n'est pas assureur mais seulement gestionnaire. Cela ne concerne qu'une infime partie de notre portefeuille, entre 1 et 2% suivant les années. Leur évolution est stable au cours du temps.

Compte tenu de la faible proportion de données concernant le type de gestion "gestionnaire", l'incidence observée est très volatile, donc pas significative, ni interprétable. En revanche, l'analyse peut être faite en se focalisant sur les gestions "directe" et "déléguée".

L'incidence de la gestion déléguée est constante depuis 2018, autour de 55%. En revanche, celle de la gestion directe a bien évolué durant la crise sanitaire. Avant la pandémie, l'incidence tournait autour de 45% alors qu'en 2020 elle est désormais égale à 56%.

	Déléguée	Directe	Gestionnaire	**Moyenne**
2018 + 2019	57%	80%	81%	**75%**
2020	59%	81%	81%	**75%**

Finalement, attardons-nous sur l'évolution de l'exposition concernant le type de gestion du contrat de prévoyance. De manière similaire à l'incidence, le type "gestionnaire" est laissé de côté. En revanche, en ce qui concerne les types de gestion déléguée et directe, l'exposition moyenne ne varie pas au cours du temps. Les contrats en gestion directe ont une exposition plus forte que la moyenne, avec quasiment cinq points de pourcentage en plus. Tandis que, les contrats en gestion déléguée ont une moyenne proche de 58%, soit une sous-exposition d'environ dix-sept pour cent par rapport à la moyenne.

Annexe C

Comparaison de modèles avec différents seuils de rejet :

Voici une capture d'écran des coefficients estimés de deux modèles. Le premier avec un seuil de rejet faible et le second avec un seuil de rejet élevé :

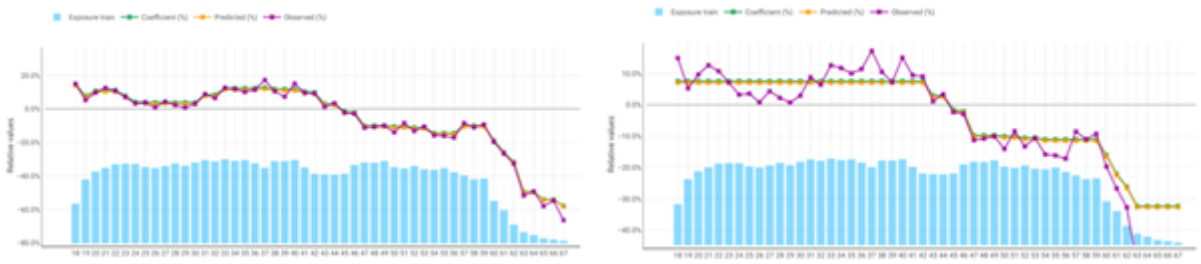


FIGURE 4.26 – Comparaison d'un modèle avec une forte importance pour les observations (à gauche) et d'un modèle avec une forte importance pour les coefficients (à droite)

À gauche, les prédictions représentées par la courbe jaune suivent quasi parfaitement les observations. Cependant, dans le graphe de droite, les hypothèses données aux coefficients sont plus fortes. Les prédictions sont désormais plus éloignées des observations et elles se basent principalement sur l'évolution des coefficients.

En bref, le GLM avertit l'utilisateur sur l'instabilité de l'estimation du paramètre en lui fournissant une erreur standard importante. Cependant, il n'ajuste pas intrinsèquement le coefficient pour prendre en compte cette volatilité. Il laisse le soin à l'utilisateur d'effectuer des ajustements ad-hoc pour prendre en compte le manque d'exposition dans un segment spécifique. Ce problème est résolu avec l'intégration de la théorie de la crédibilité dans les modèles utilisés sous Akur8. Toutefois, pour les utilisateurs, le mélange des deux modèles est directement intégré et l'outil propose un algorithme d'apprentissage automatique unifié.

Dans le but d'illustrer la différence entre ces deux modèles, prenons l'exemple d'une variable spécifique, définissant l'âge de nos salariés.

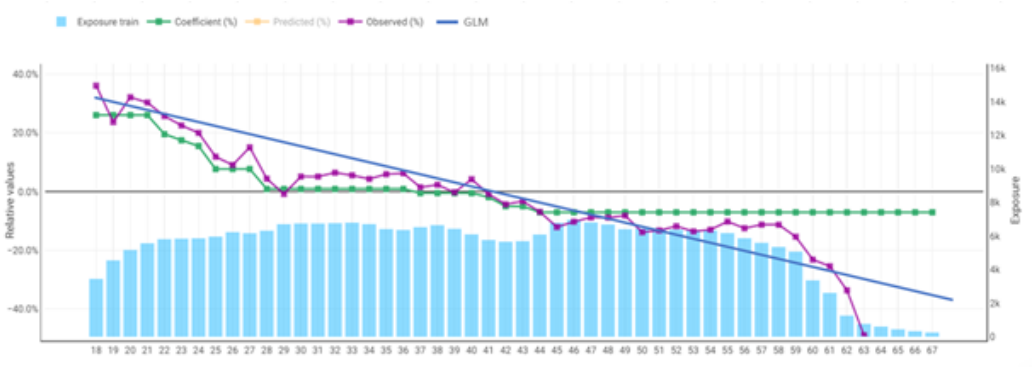


FIGURE 4.27 – Exemple illustratif sur les différences au niveau des coefficients

En violet, le nombre d'arrêts observés, découpés en fonction de l'âge des salariés de notre population. La ligne bleue représente l'effet linéaire de l'âge sur le nombre d'arrêts de travail. C'est ce qu'on observe avec les modèles GLM. En effet, le coefficient lié à l'âge est négatif, donc l'effet de la variable décroît de manière monotone au fur et à mesure que l'âge augmente. Enfin, la courbe verte correspond aux coefficients liés à l'âge pour le modèle sous Akur8. La différence avec les GLM est flagrante puisque différents coefficients sont associés en fonction de l'âge du salarié. Pour les jeunes salariés, les coefficients sont bien plus importants, alors que pour les individus âgés de plus de 45 ans le coefficient est bien inférieur aux autres. Cela veut donc dire que les individus les plus jeunes ont un impact plus important sur la variable cible, mais avec une tendance non linéaire.

Finalement, la tendance globale reste la même, plus l'âge augmente et moins les salariés influencent la variable cible. Néanmoins, avec les modèles GAM, l'information perçue est beaucoup plus précise. Des effets relatifs à une ou plusieurs tranches d'âge apparaissent.