Mémoire présenté le :

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

Par : Alkhoury Perla Maria

Titre : Analyse du lien entre la consommation médicale et le risque arrêt de travail

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

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

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Entreprise : Crédit Agricole Assurances

Nom :

Signature :

Directeur de mémoire en entreprise :

Nom : LETAILLEUR Charles-Henri

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

La protection sociale est un univers en constante évolution. Le pilotage et le suivi du risque est un enjeu majeur pour un assureur. Afin d'éviter la dérive des portefeuilles, il est essentiel de bien analyser la sinistralité et de considérer le risque le plus finement possible.

Bien que le risque Santé soit souvent associé à celui de Prévoyance, les deux risques ont jusqu'à présent été étudiés de façon indépendante. Toutefois, l'arrêt de travail est directement lié à la santé de l'assuré. De ce fait, on peut penser à un potentiel lien entre les prestations santé et la survenance ainsi que la durée d'un arrêt de travail. Inversement, l'arrêt de travail peut avoir un impact sur la consommation santé de l'assuré.

Le but de ce mémoire est d'apporter des connaissances en analysant le lien qui peut exister entre ces deux risques. L'étude s'articule autour de trois grandes parties.

La première consistera à analyser l'impact de la consommation en santé sur l'exposition au risque d'arrêt de travail. Un modèle de prédiction d'arrêt de travail sera élaboré à l'aide d'un Modèle Linéaire Généralisé classique. Ce modèle sera étudié, optimisé et raisonné afin de pouvoir l'utiliser opérationnellement.

La deuxième concernera l'impact de cette consommation en santé sur la durée de l'arrêt. Un modèle de durée sera utilisé pour tenter de prédire au mieux la durée des arrêts.

La troisième s'intéresse à l'influence d'un arrêt sur le comportement et les consommations santé des assurés. Un modèle sera élaboré dans le but de prédire si un assuré est en arrêt ou non sur la base de sa consommation santé.

Mots clés : Arrêt de travail, Incapacité, Prévoyance, Santé, Modèles Linéaires Géneralisés, Régression Logistique

Abstract

The social protection sector is constantly evolving. Effective risk management and monitoring are crucial for insurers. To prevent portfolio drift, it is essential to analyze claims data thoroughly and consider risk as finely as possible.

Although Health and Welfare risks are often associated, they have traditionally been studied independently. However, work stoppages are directly related to the insured's health. Therefore, there may be a potential link between health consumption and the occurrence and duration of a sick or disability leave. Conversely, a sick leave can impact the insured's healthcare consumption.

The purpose of this thesis is to provide insights by analyzing the link that may exist between these two risks. The study is divided into three main parts.

The first part is to analyze the impact of healthcare consumption on the risk of sick leave. A predictive model of sick or disability leave will be developed using a classic Generalized Linear Model. This model will be studied, optimized, and rationalized in order to be operational.

The second part will examine the impact of healthcare consumption on the duration of a sick or disability leave. A duration model will be used to predict the length of a sick or disability leave.

The third part focuses on the influence of a sick or disability leave on insured individuals behavior and healthcare consumption. A model will be developed to predict whether an insured person is on a work stoppage based on their healthcare consumption.

Key words : Sick leave, Disability, Health, Generalized Linear Models, Logistic Regression.

Remerciements

Je tiens à d'abord à remercier chaleureusement mon tuteur en entreprise Charles-Henri LE-TAILLEUR, tout d'abord, pour la confiance qu'il m'a accordé mais aussi pour son aide au choix du sujet et son accompagnement.

Je tiens aussi à remercier l'ensemble des membres de l'équipe Etude et Inventaire pour leur accueil et les réponses qu'ils ont pu m'apporter. Travailler à leurs côtés durant ces deux années fût un plaisir. Je souhaite tout particulièrement remercier mon collègue Dimitri NIERAT, qui a toujours su se montrer disponible afin de répondre à mes questions et m'accompagner dans mes travaux.

Je remercie l'équipe pédagogique et administrative de l'ISFA et plus particulièrement mon tuteur académique Fréderic PLANCHET pour son accompagnement et sa disponibilité tout au long de l'année.

Le mot de la fin est adressé aux personnes qui me supportent et soutiennent au quotidien, je pense à ma mère Noha ALKHOURY, mes soeurs Catherina et Sandrine ALKHOURY.

Table des matières

Introduction

Le marché de l'assurance Santé Prévoyance a fortement évolué ces dernières années et plus particulièrement la concurrence au sein de la branche. Les revalorisations tarifaires suite à la hausse des dépenses de santé ainsi que les possibilités de résiliation infra-annuelle favorisent une accélération inédite de la mobilité des clients. Parallèlement, l'ouverture à de nouveaux marchés, tel que celui des agents de la Fonction Publique avec la réforme PSC, engendrera un net durcissement de la compétition sur le marché. L'explosion exponentielle des données et des moyens de collecte et d'exploitation permet une connaissance et donc une anticipation des besoins plus précise afin d'éviter de subir les consommations.

Dans cet environnement concurrentiel accru, il est crucial pour un assureur de bien suivre et analyser le risque le plus finement possible. Premièrement, le bon pilotage et le suivi assurent la maîtrise du risque et protègent d'une éventuelle dérive du portefeuille. Deuxièmement, ils permettent de mettre en évidence les besoins d'ajustements tarifaires. Le Crédit Agricole Assurances, "nouvel acteur" sur le marché de la Santé Prévoyance Collective, est en constante évolution et gagne de plus en plus en part de marché. Sa stratégie peut être qualifiée d'offensive, ce qui rend l'étude et la mise en place de nouveaux indicateurs intéressantes.

Même si le risque Santé est souvent associé à celui de Prévoyance, les deux risques ont jusqu'à présent été étudiés de façon indépendante. Toutefois, l'arrêt de travail est directement lié à la santé de l'assuré ; la dégradation de la santé d'un individu pouvant mener à un arrêt. Cette dégradation peut notamment se mesurer par une tendance de consommation médicale à la hausse. Intuitivement, une personne en "mauvaise santé" consommera plus en santé et aura plus de chances d'être en arrêt de travail.

Inversement, un individu ne consommant pas "beaucoup" en santé, peut être interprété comme ne prenant pas soin de sa santé. En cas d'arrêt, il peut avoir besoin de plus de temps pour récupérer, et donc pour reprendre son activité professionnelle. Ainsi, on peut penser à un potentiel lien entre les prestations santé et la durée de l'arrêt de travail.

Enfin, l'arrêt de travail peut se traduire par une fragilisation de l'état de santé d'un individu et s'accompagner ainsi d'une consommation santé plus élevée. Ainsi, l'arrêt de travail peut avoir un impact sur la sinistralité en Santé.

Le but de ce mémoire est d'analyser le lien existant entre ces deux risques. Dans un premier temps, nous étudierons l'impact de la consommation en santé sur l'exposition au risque d'arrêt de travail. Une deuxième partie concernera l'impact de cette consommation sur la durée de l'arrêt. Enfin, nous regarderons l'influence d'un arrêt sur le comportement et les consommations santé des assurés.

Chapitre 1

Contexte de l'étude

Sommaire

Ce mémoire a pour objectif d'étudier le lien potentiel entre l'assurance prévoyance et l'assurance santé. Dans ce qui suit, nous allons présenter les notions utilisées ainsi que l'intérêt d'un point de vue métier.

1.1 Assurance Santé

L'assurance santé a pour objectif de se prémunir du risque pour un assuré (ou un bénéficiaire) d'être malade et ainsi de faire face à des dépenses pour soigner, prévenir, ou guérir cette maladie ou ce syndrome. Les différents sinistres couverts par un contrat de Santé sont souvent classés suivant leur nature :

— Hospitalisation : il s'agit par exemple des frais de séjour, les honoraires de chirurgie, une anesthésie, des prestations dites "confort" telles que la chambre particulière ou les frais d'accompagnant.
— Soins courants : il s'agit des dépenses de consultations de médecins généralistes et spécialistes, des frais de laboratoires, d'analyses.
— Pharmacie : il s'agit des prescriptions de médicaments (remboursés par le régime obligatoire).
— Optique : il s'agit de remboursements de lunettes (verres et monture), lentilles, ou opérations optiques.
— Dentaire : il s'agit de soins dentaires, prothèses ou encore l'orthodontie.
— Autres : il s'agit de toutes les autres prestations telles que les cures thermales prescrites et remboursées par le régime obligatoire, le remboursement de séances de médecines douces.

Nous discuterons par la suite de la pertinence de cette répartition et notamment la répartition retenue pour notre étude. Dans ce mémoire, nous ne tenons compte que de la consommation de l'assuré et donc excluant de l'étude la consommation médicale des éventuels ayants droits parce que seul l'assuré est concerné par l'assurance prévoyance.

1.2 Assurance Prévoyance

L'assurance prévoyance est une protection financière contre le risque de perte partielle ou totale des revenus d'un individu. Il est commun de distinguer les actes par les faits générateurs : accident de travail ou maladie (dans un cadre professionnel ou privé). Trois risques sont couverts :

— L'incapacité : l'état d'incapacité désigne l'incapacité physique, constatée par le médecin traitant de l'individu, à continuer ou à reprendre à temps plein l'activité professionnelle, à la suite à d'un accident de travail ou d'une maladie (art. L.321-1 du Code de la Sécurité Sociale). La durée maximale de maintien en incapacité est de 36 mois, au-delà de cette période, l'individu passe en invalidité au sens de la Sécurité Sociale.
— L'invalidité : l'invalidité désigne l'incapacité permanente.
— Le décès : les prestations de la Sécurité Sociale ne remplaçant pas la totalité de la perte de revenus liée à un décès, la complémentaire prévoyance permet d'ajouter des revenus aux assurés et bénéficiaires.

Nous nous intéressons au risque d'incapacité et plus particulièrement à l'occurrence ou pas d'un arrêt de travail ainsi qu'à la durée de ce dernier.

1.3 L'intérêt d'un point de vue métier

Même si ces deux risques sont souvent associés l'un à l'autre, leurs natures et approches diffèrent. En santé, les sinistres sont fréquents et leurs coûts très variables. En prévoyance, les fréquences sont moindres, mais leurs coûts peuvent s'avérer élevés. Les deux risques ont jusqu'à présent été étudiés de façon indépendante. Néanmoins, on peut penser que la santé d'un assuré peut influencer son exposition au risque d'arrêt de travail. En effet, intuitivement, une personne en "mauvaise santé", consommant plus de certains types de soins de santé, a plus de chances de tomber en arrêt. Par ailleurs, une personne n'ayant pas "beaucoup" consommé en santé, et donc ne prenant pas "soin" de sa santé, peut prendre plus de temps pour guérir et donc reprendre son activité professionnelle, si elle s'avère être en arrêt.

D'un point de vue métier, il peut s'avérer intéressant d'étudier la relation existante entre ces deux risques, notamment dans le but de mieux prévenir l'arrêt de travail. Effectivement, les arrêts de travail ne faiblissent pas et l'absentéisme dévoile un coût économique de plus en plus important pour les assureurs. Cette étude a pour but de mieux comprendre et prévenir ce risque et présente un double enjeu :

— Un en termes de pilotage technique ;

— Un en termes de souscription ;

1.3.1 Pilotage Technique

La spécificité d'une compagnie d'assurance réside dans l'inversion de son cycle. En conséquence, un décalage peut avoir lieu entre la date de survenance, le générateur du paiement et la date effective du règlement. L'un des enjeux majeurs d'un assureur est de suivre et de gérer son risque. Cela permet :

— De surveiller les indicateurs et d'éviter toute dérive ;
— De mettre en évidence les besoins de tarification et d'ajustement tarifaire ;
— De s'assurer de la cohérence de la stratégie ;

Le pilotage et le suivi du risque constitue ainsi un enjeu majeur pour un assureur. Il est essentiel de bien analyser la sinistralité afin de limiter la dérive des portefeuilles. Dans un contexte concurrentiel, il est important de considérer le plus finement possible le risque. Les données santé peuvent constituer un indicateur supplémentaire et innovateur pour le suivi du risque arrêt de travail et son provisionnement. L'exploitation de la data Santé pourra ainsi être utilisée pour anticiper une éventuelle dérive de l'absentéisme. Elle pourra également contribuer à piloter les indexations tarifaires pour les entreprises disposant des deux assurances.

1.3.2 Souscription

Par ailleurs, le portefeuille est majoritairement composé de contrats santé. Le groupe cherche à développer son portefeuille prévoyance et notamment en "multi-équipement". Dans le cas où une entreprise souhaiterait la souscription d'une couverture Prévoyance, l'analyse de sa sinistralité santé selon la méthodologie décrite permettrait d'affiner le positionnement tarifaire sur le risque Arrêt de Travail. Enfin, le groupe se veut actif notamment en termes de Prévention. Un véritable diagnostic est établi pour guider le client dans le choix de la solution la plus adaptée en fonction du contexte de son entreprise, de ses enjeux et de ses objectifs. Cette étude pourra être utilisée en guise de prévention dans le but d'atténuer l'effet d'une éventuelle hausse de la sinistralité en prévoyance. Elle pourra compléter les programmes déjà mis en place en proposant des services de prévention aux entreprises les plus exposées au risque arrêt de travail.

Chapitre 2

Environnement et base de données

Sommaire

2.1 Construction de la base

La construction de la base est une étape cruciale de notre étude. Des données fiables et en quantité suffisante assurent la robustesse de la modélisation.

Pour cette étude, nous nous intéressons aux individus ayant à la fois une couverture Santé et une couverture Prévoyance incapacité. Pour ce faire, nous utiliserons quatre bases :

— La base EFFECTIFS Santé comportant l'ensemble des effectifs bénéficiant d'une couverture Santé, leurs numéros de contrat, leurs identifiants assurés, leurs dates de naissance, leurs dates et durées d'affiliation ;
— La base EFFECTIFS Prévoyance comportant l'ensemble des effectifs bénéficiant d'une couverture Prévoyance, ainsi que les différentes garanties souscrites ;
— La base SINISTRES Prévoyance comportant l'ensemble des sinistres de type incapacité, invalidité et Décès (Y compris les sinistres non indemnisés) ;
— La base SINISTRES Santé comportant toutes les consommations en santé et les informations telles que le nombre d'actes, le montant remboursé...

La première étape est de sélectionner les individus "double équipés". Pour ce faire, nous croisons les tables Effectifs Santé et Effectifs Prévoyance et sélectionnons les individus présents dans les deux bases. De plus, il est important de s'assurer que ces assurés disposent d'une garantie incapacité. En effet, une personne peut avoir un contrat de prévoyance sans garantie incapacité. L'inclusion de cette personne biaiserait notre étude.

Concernant la période d'observation, nous sélectionnons les exercices 2017, 2018 et 2019. En effet, le Crédit Agricole s'étant lancé sur le marché Santé Prévoyance collectives en 2015, le portefeuille de 2016 ne comporte pas suffisamment d'individus. Par ailleurs, l'année 2020 est fortement influencée par la pandémie COVID-19. Les constats du marché affirment que la consommation santé a diminué tandis que les arrêts se sont multipliés. D'une part, plusieurs études montrent que les confinements ont engendré une baisse importante des prestations santé. Ce phénomène peut être expliqué par le report des rendez-vous jugés "non urgents" de soins courants, d'optiques ou encore de dentaires, mais aussi des opérations chirurgicales. D'autre part, la pandémie a eu un impact significatif sur l'absentéisme : la suspicion ou la confirmation de cas de Covid-19 constituent une partie importante des motifs d'arrêt de travail de l'année. Ainsi, l'intégration de l'année 2020 risque de créer un biais important quant au comportement des assurés. Enfin, le décalage temporel entre la survenance d'un sinistre et son enregistrement dans les bases de données de l'assureur, nous empêche de retenir l'année 2021. En effet, nous ne disposons pas de recul comptable suffisant pour l'inclure dans cette étude.

Ensuite, nous rajoutons la condition de 12 mois consécutifs d'observation. Les assurés doivent être présents dans nos bases pendant une période suffisamment longue pour nous permettre d'avoir un recul conséquent sur leurs prestations.

Par conséquence, les différentes sélections nous mènent à retenir 34 500 individus. Notre base à ce stade se compose de l'identifiant de l'assuré, sa date de naissance, sa date d'ancienneté, son sexe et sa catégorie socioprofessionnelle.

La seconde étape est la sélection des prestations santé. Le montant des prestations étant très influencé par le niveau de garantie, nous optons pour le nombre d'actes par poste. Nous décidons d'affiner la répartition présentée précédemment et de regrouper les sinistres sous 15 familles d'actes :

— Chirurgie Yeux. On décide d'isoler les actes de chirurgie des yeux. En effet, on suppose que l'impact sur l'arrêt de travail n'est pas le même que le reste des actes d'optique ;
— Optique Autres ;
— Dentaire. On regroupe l'ensemble des consommations dentaires en une seule famille ;
— Forfait Journalier ;
— Chambre Particulière ;
— Séjour Hospitalier. On regroupe l'ensemble des frais liés à un séjour hospitalier ;
— Maternité ;
— Forfait Journalier Psychiatrique. On regroupe les actes de psychiatrie et de psychologie ;
— Consultation Généraliste ;
— Consultation Spécialiste ;
— Pharmacie ;
— Médecine Douce ;
— Soins Autres ;
— Dispositifs médicaux ;
— Cures ;

Par ailleurs, nous décidons de supprimer la variable "Optique_Autres" car nous jugeons qu'elle n'est pas discriminante dans notre modèle et qu'elle apportera du bruit à notre étude. Intuitivement, le seul acte optique qui peut mener à un arrêt est la chirurgie, acte que nous avons isolé dans la variable "Optique_chirurgie".

FIGURE 2.1 – Sélection temporelle des données santé

Enfin, via la table des prestations prévoyance, nous identifions la présence ou non d'un arrêt survenu entre 2017 et 2019. Nous récupérons la date de survenance et la date de fin de l'arrêt si arrêt il y a.

Dans le cas où l'assuré a eu un arrêt, nous décidons ainsi de retenir le nombre de prestations santé durant les 12 mois, 6 mois et 3 mois précédant le premier arrêt. Dans le cas où l'assuré n'a eu aucun arrêt et afin de faciliter l'interprétation, nous retenons la moyenne du nombre d'actes santé sur 12, 6 et 3 mois. De plus, nous calculons la durée de l'arrêt en jour. La date de sortie de l'arrêt maximale étant le 31/12/2019, on note que nos données ont une censure à droite.

Il est à noter que notre base comporte l'intégralité des arrêts, y compris ceux non indemnisés (grâce à la franchise). En effet, la franchise, en assurance Prévoyance, correspond à la période pendant laquelle un assuré ne reçoit pas d'indemnités journalières de son assureur. Cette période s'exprime généralement en jours et dépend des termes et des conditions du contrat. Cette dernière est souvent source de censure de données, et est importante à considérer car elle peut limiter la quantité et dégradé la qualité des données disponibles pour les assureurs. Dans notre étude, les données ne présentent pas de censure liée à la franchise. En effet, nous disposons des sinistres indemnisés et non indemnisés.

Les deux variables d'intérêt sont "arret" (qui vaut 1 si arrêt il y a, 0 sinon) et "nb_jour_arret" qui correspond à la durée de l'arrêt en jours. Ci-dessous un aperçu de la base de données :

Date de survenance

Consommation Santé sur 12, 6 et 3 mois

id_assure	dat_nais	Age_début	GENRE	college	DT_SURV	annee_surv	ib_jours_arre	arret	Duree_affil	Optique_chirurgie_AVT_12	Chambre_Part_AVT_12
10001	[redacted]	33,1	F	ENSEMBLEDUPERSONNEL				0	12	0	0,0
100014	[redacted]	19,3	F	ENSEMBLEDUPERSONNEL	02/09/2017	2017	4	1	32	0	0,0
100017	[redacted]	29,6	F	ENSEMBLEDUPERSONNEL				0	36	0	0,0

FIGURE 2.2 – Aperçu Base de données

2.2 Statistiques descriptives

2.2.1 Variable : Sexe

Le sexe est un paramètre intéressant à exploiter. Il peut constituer une variable influente dans notre modèle. En effet, on observe généralement des comportements différents entraînant des tendances et des taux d'absentéisme différents selon le sexe.

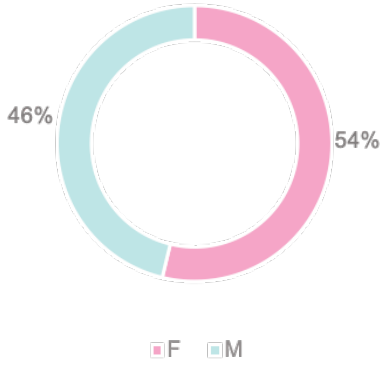


FIGURE 2.3 – Répartition des assurés par sexe

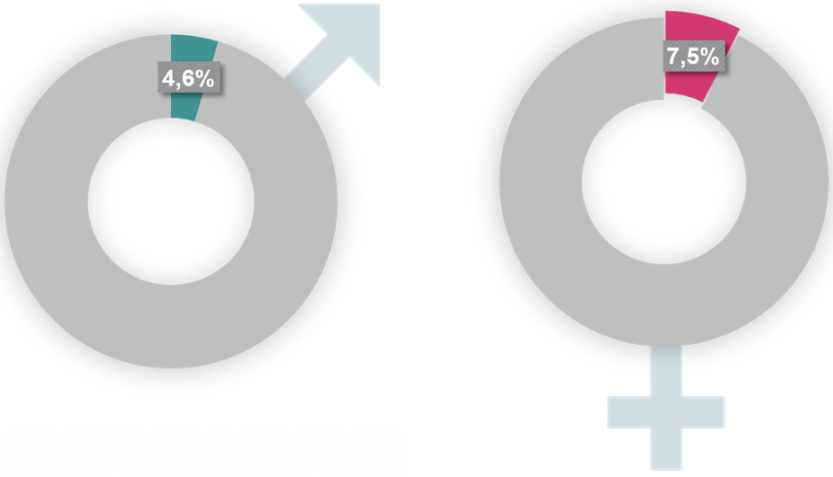


FIGURE 2.4 – Probabilité de tomber en AT selon le sexe

On observe dans notre base de données une majorité de femmes. Cette particularité est liée au périmètre de l'étude. Par ailleurs, le taux de femmes ayant eu au moins un arrêt durant la période d'étude est de 7,5%, soit 63% de plus que chez les hommes. Ce constat est observable ailleurs sur le marché et peut être expliqué en partie par les arrêts liés à la maternité.

2.2.2 Variable : Age

Nous nous intéressons à présent à l'âge de début de la période d'observation : la date de début de période d'observation à savoir, le 01/01/2017, ou la date de début de couverture si le salarié n'était pas couvert en début de période d'observation.

Nous observons l'histogramme de la variable "Age" avant traitement. S'agissant de contrats collectifs, la majorité des assurés ont entre 20 et 65/66 ans. En effet, les contrats de santé prévoyance collectifs couvrent les salariés. En supposant que le salarié partira à la retraite dès qu'il sera en mesure d'obtenir une retraite à taux plein à la Sécurité Sociale, la durée de cotisation nécessaire varie en fonction des générations, des secteurs et des catégories socio-professionnelles. On peut toutefois, supposer que les valeurs au-delà de 66 sont "aberrantes".

On décide de remplacer ces valeurs, jugées aberrantes. Afin d'éviter une altération sensible de la distribution, on choisit de remplacer toutes les valeurs supérieures à 66 ans par la moyenne des âges observés qui est de 43,7.

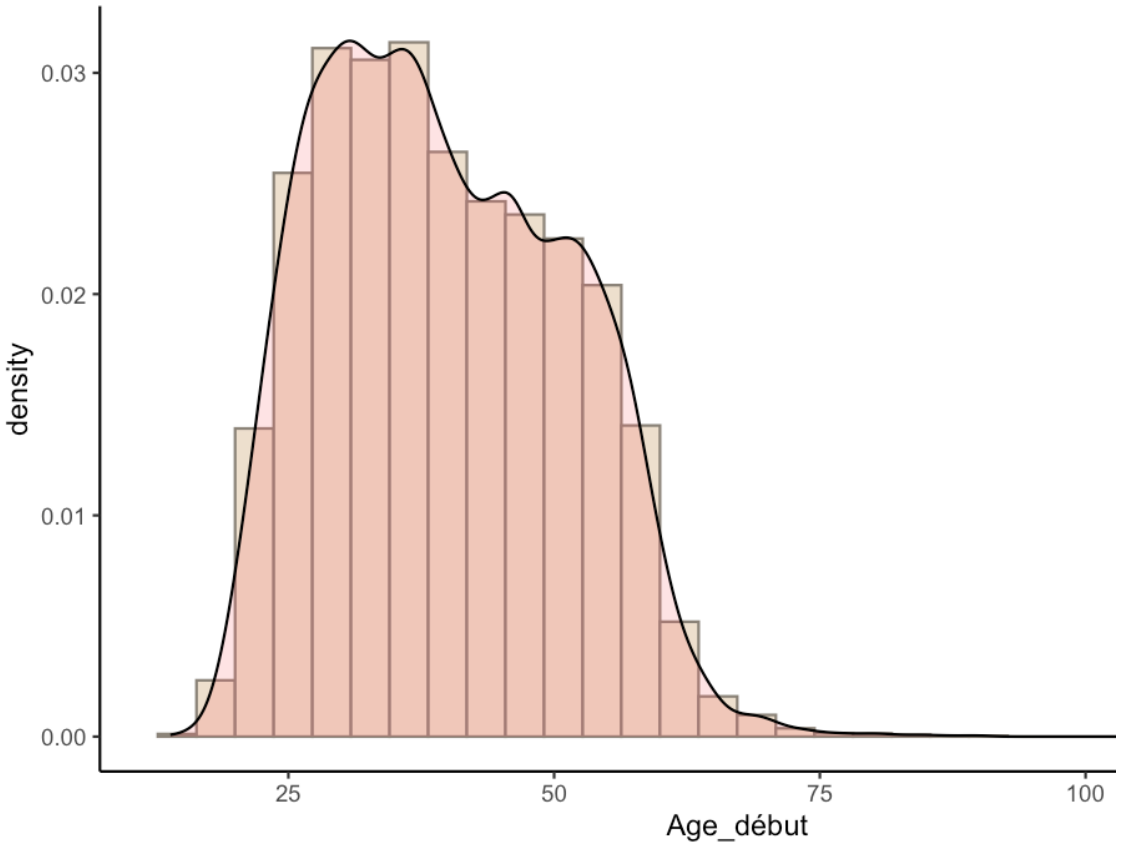


FIGURE 2.5 – Histogramme de la variable Age_début

Nous décidons de regarder de plus près le *boxplot* de la variable Age_début.

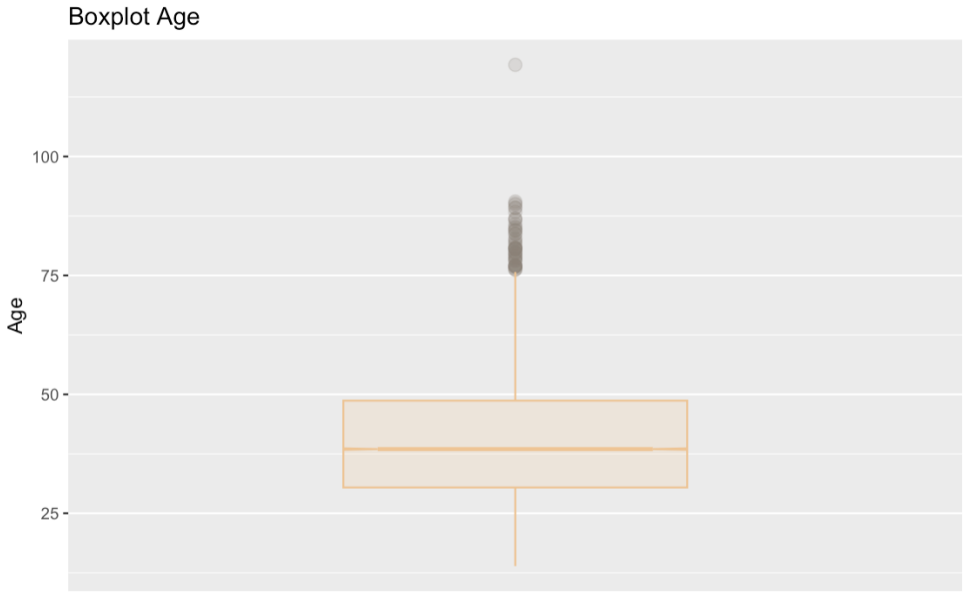


FIGURE 2.6 – Boxplot de la variable Age

Nous observons l'histogramme ci-dessous de la variable après traitement. La majorité des âges sont compris entre 25 et 60 ans, ce qui est cohérent vu qu'il s'agit de salariés.

FIGURE 2.7 – Histogramme de la variable Age

Par ailleurs, nous décidons de regarder la répartition de la variable par sexe. Nous observons une répartition par âge relativement homogène entre les hommes et les femmes.

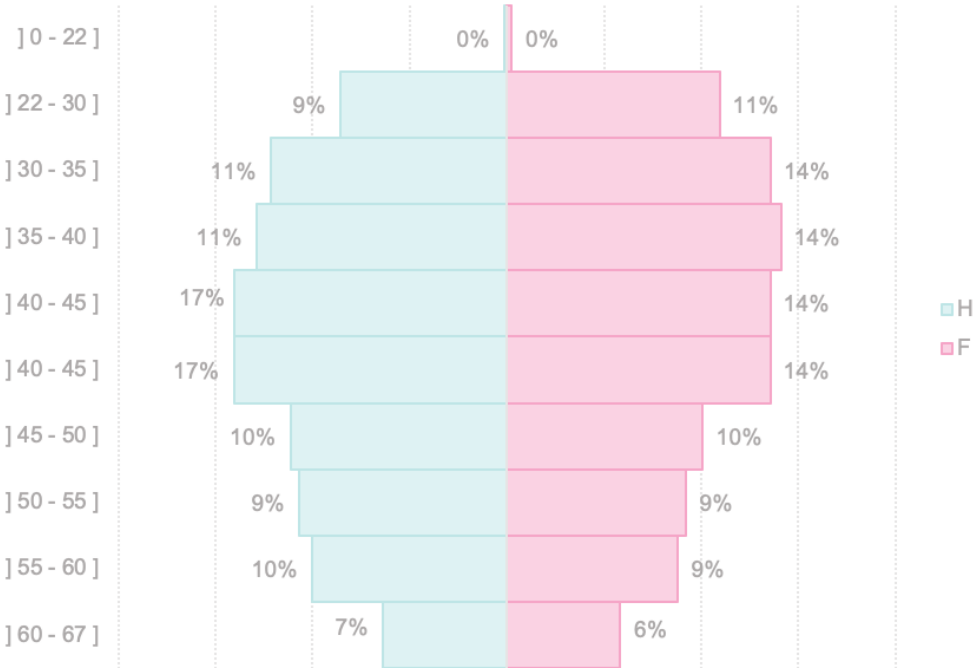


FIGURE 2.8 – Pyramides des âges des assurés

Nous nous intéressons à présent aux taux d'entrée en arrêt de travail par âge et par sexe. Nous remarquons que les taux sont très irréguliers. Cela s'explique par la faible volumétrie des données. Nous notons tout de même qu'à tout âge la courbe des femmes est supérieure à celle des hommes et que les tendances à la hausse ou à la baisse semblent s'accorder. Nous décidons par la suite de les lisser avec la méthode de Whittaker Henderson.

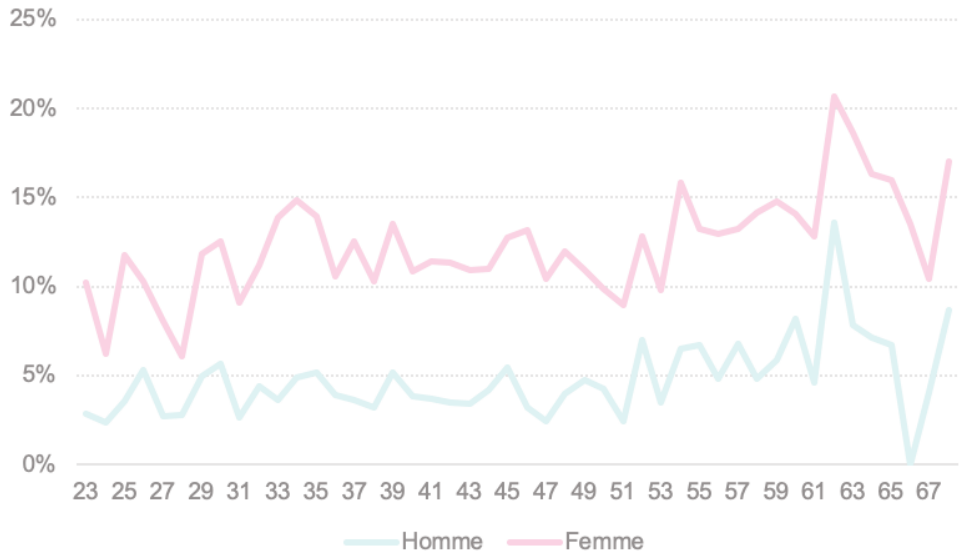


FIGURE 2.9 – Taux d'entrée en AT par âge et sexe

2.2.3 Variable : Collège

L'ensemble du personnel est une catégorie qui regroupe les cadres et les non-cadres. L'information n'est ainsi pas disponible sur 68% du périmètre. Cela risque de biaiser les résultats et de ainsi rendre la variable non-exploitable. Nous décidons à ce stade de garder cette dernière. Sur les informations observées, on note une majorité de non-cadres.

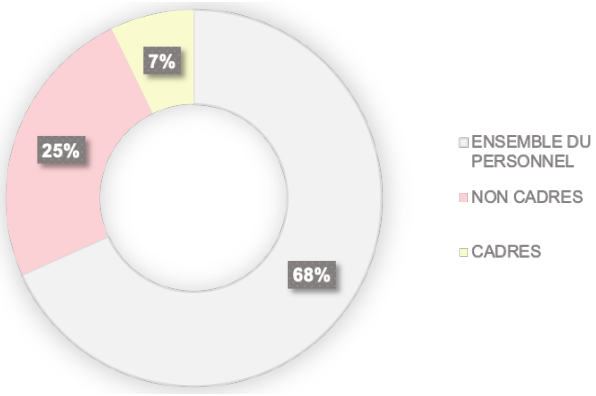


FIGURE 2.10 – Répartition des assurés par collège

En s'intéressant à la probabilité d'entrée en arrêt de travail par catégorie socioprofessionnelle. On remarque que la probabilité la plus haute est observée chez les non-cadres – ce qui est en accord avec les constats du marché. En effet, les différences dans les comportements d'absentéisme entre les cadres et les non-cadres peuvent en partie s'expliquer par certains facteurs socio-économiques : en particulier, des tâches parfois physiquement plus exigeantes, mais aussi des salaires et des conditions d'emploi moins avantageux. D'une part, les non-cadres sont souvent plus susceptibles de souffrir de problèmes de santé liés au travail, tels que des blessures ou des maladies professionnelles. D'autre part, selon les observations sur le marché, les cadres ont tendance à plus consommer en santé et donc de prendre plus soin de leur santé.

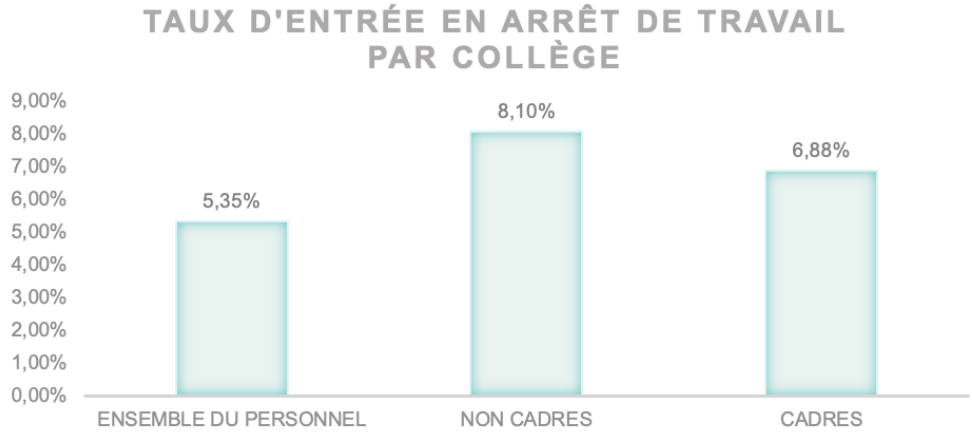


FIGURE 2.11 – Taux d'entrée en arrêt de travail par collège

2.2.4 Variable : Secteur d'activité

Grâce aux systèmes d'information du Crédit Agricole Assurances, nous avons pu déterminer les branches d'activité pour chaque assuré. Cette information sur la branche d'activité n'est pas disponible à l'origine. Nous pensons que cette variable peut être discriminante dans notre modèle : la probabilité d'avoir un arrêt, les natures des arrêts et potentiellement la consommation médicale peuvent être différentes en fonction des secteurs d'activité.

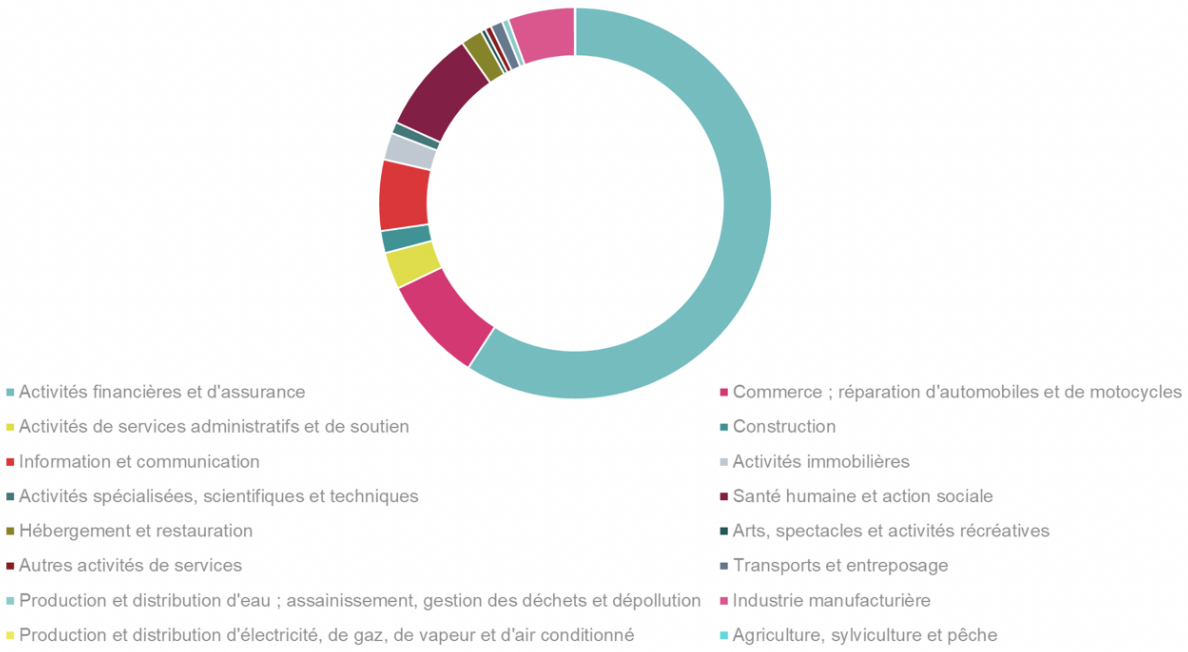


FIGURE 2.12 – Répartition des assurés par secteur d'activité

Toutefois, nous constatons une répartition non-uniforme et fortement déséquilibrée. Cette répartition est essentiellement due au périmètre de l'étude : le portefeuille dit "double équipé". Nous tentons dans un premier temps de retraiter la variable et de catégoriser les secteurs d'activité en deux catégories : Services et Industrie. Nous obtenons la répartition suivante : 90% de Service et 10% Industrie. Ce fort déséquilibre nous empêche à ce stade de garder cette variable, qui a priori aurait pu être discriminante. Dans ce qui suit, nous tenterons de l'ajouter aux modélisations et la retirerons dans le cas où cette dernière ne serait pas significative.

2.3 Données déséquilibrées

Nous observons un taux global d'entrée en incapacité de 6,14%. Ce taux est dans la fourchette de ce que l'on observe généralement sur le marché.

Toutefois, notre base de données possède beaucoup plus de lignes avec arret=0 que de lignes où arret=1. Ainsi, elle est dite déséquilibrée. On parle de données déséquilibrées dès lors que les deux classes ne présentent pas les mêmes fréquences. Dans notre cas, nous sommes loin d'une représentation 50% - 50%. Ce phénomène est fréquent dans les modèles de classification et peut

s'avérer problématique lorsqu'il n'est pas traité. En effet, on risque de biaiser notre modèle si l'on ne prend pas en compte ce déséquilibre. En particulier, un modèle linéaire parviendra difficilement à capter le signal de la classe minoritaire si cette dernière est très peu représentée. La maximisation de la vraisemblance conduira généralement à prédire une probabilité relativement uniforme pour tous les individus.

Pour remédier à ce problème, deux solutions se présentent à nous :

— Sous-échantillonnage : il s'agit ici de retirer aléatoirement des observations de la classe majoritaire. En d'autres termes, on supprime des lignes d'individus n'ayant pas eu d'arrêt pendant la période d'observation.

— Sur-échantillonnage : il s'agit ici de répliquer des observations de la classe minoritaire, tirées aléatoirement. Cette dernière verra son poids augmenter lors d'une modélisation linéaire.

Ne disposant pas de beaucoup de données, nous optons pour la deuxième option. Pour ce faire, nous utilisons la fonction "*mwmote*" du package R "*Imbalance*" et décidons de doubler le taux d'arrêt et de rajouter 2 117 lignes de la classe minoritaire. Cela est uniquement utilisé pour améliorer la performance de la régression logistique et ces lignes seront supprimées par la suite. Par ailleurs, il est impératif d'accorder une attention particulière aux mesures de performance. Nous expliciterons ce point d'attention par la suite.

2.4 Lissage avec Whittaker Henderson

2.4.1 Rappels théoriques

Nous nous intéressons à présent au lissage de la courbe des taux d'entrée en incapacité par sexe. Nous optons pour un modèle de lissage non-paramétrique car nous ne connaissons pas la loi théorique qui pourrait être adaptée aux taux. Whittaker-Henderson est une des méthodes non-paramétriques les plus connues. Elle présente l'avantage de pouvoir modifier facilement des paramètres de régularité et de fidélité afin de s'adapter au mieux aux données brutes.

Deux critères principaux sur lesquels repose le lissage sont la fidélité et la régularité :

— Les taux lissés doivent être proches des taux originaux, ceci est contrôlé par le critère de fidélité.

— La courbe des taux ajustés doit être la plus lisse possible, ceci est contrôlé par le critère de régularité.

Le principe du modèle de Whittaker-Henderson est de combiner linéairement la régularité et la fidélité de manière à minimiser la somme de ces deux critères afin d'obtenir le meilleur rapport "fidélité/régularité" pour le lissage. L'importance du critère de régularité par rapport au critère de fidélité peut se paramétrer via un paramètre de poids. Soit :

$$min(M) = F + hR \tag{2.1}$$

$$\text{Où} \begin{cases} F \text{ est la mesure de fidélité} \\ R \text{ est la mesure de régularité} \\ h \text{ est le poids affecté à la régularité} \end{cases}$$

On note que si h est nul, $M = F$. Il n'y a pas lissage. Le critère de fidélité s'exprime de la façon suivante :

$$F = \sum_{i=1}^{n} w_i \times (q_i - \hat{q}_i)^2 \tag{2.2}$$

$$\text{Où} \begin{cases} n \text{ représente le nombre de taux sur lesquelles le lissage doit être effectué} \\ i \text{ est l' indice des taux bruts} \\ q_i \text{ est le taux brut} \\ \hat{q}_i \text{ est le taux lissé} \\ w_i \text{ est le facteur poids} \end{cases}$$

Généralement, la pondération peut se faire de deux manières :

— Attribuer un poids de 1 pour chaque âge ;

— Définir le poids en fonction des effectifs de chaque âge ;

Nous optons pour la deuxième option et pondérons la série des taux bruts avec les poids des effectifs pour chaque âge. Le critère de régularité s'exprime comme suit :

$$R = \sum_{i=1}^{n-z} (\Delta^z q_i)^2 \tag{2.3}$$

Où $D\Delta_i$ représente la différence des taux bruts

$$\begin{cases} \Delta q_i = q_{i+1} - q_i \\ \Delta^2 q_i = \Delta(\Delta q_i) = \Delta(q_{i+1} - q_i) = (q_{i+2} - q_{i+1})(q_{i+1} - q_i) \end{cases}$$

$\Delta^z q_i$ est Δq_i composé z fois. Le paramètre z du modèle permet de contrôler la régularité du lissage.

2.4.2 Application

Le lissage a été effectué sur R à l'aide des codes de lissage mis à disposition sur le site de Ressources-Actuarielles pour les modèles de durée. Nous testons plusieurs paramètres z et h. Dans le graphique ci-dessous, le facteur h représente la régularité et z la fidélité.

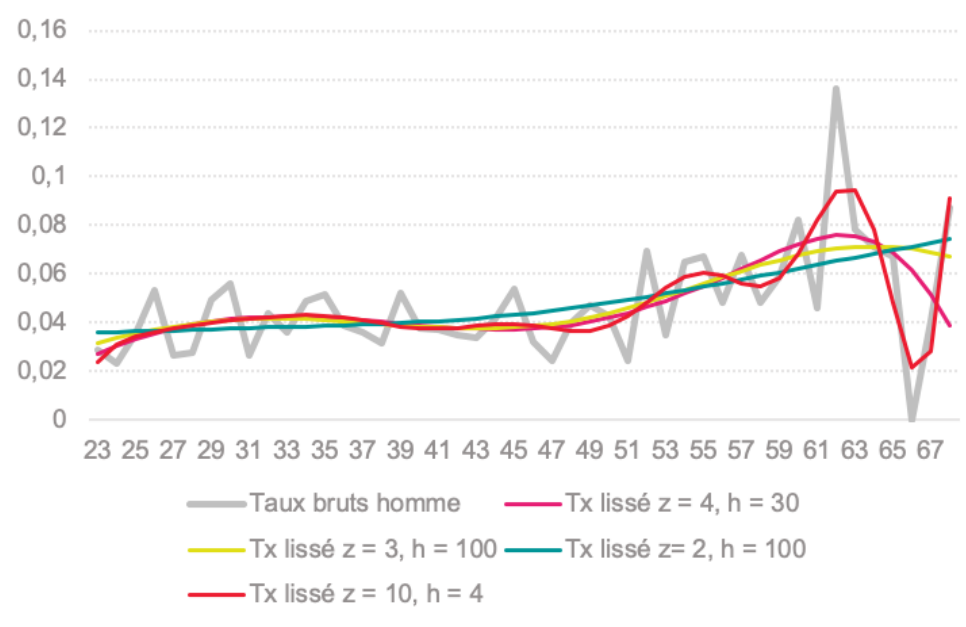


FIGURE 2.13 – Comparaison différents lissages pour les taux bruts hommes

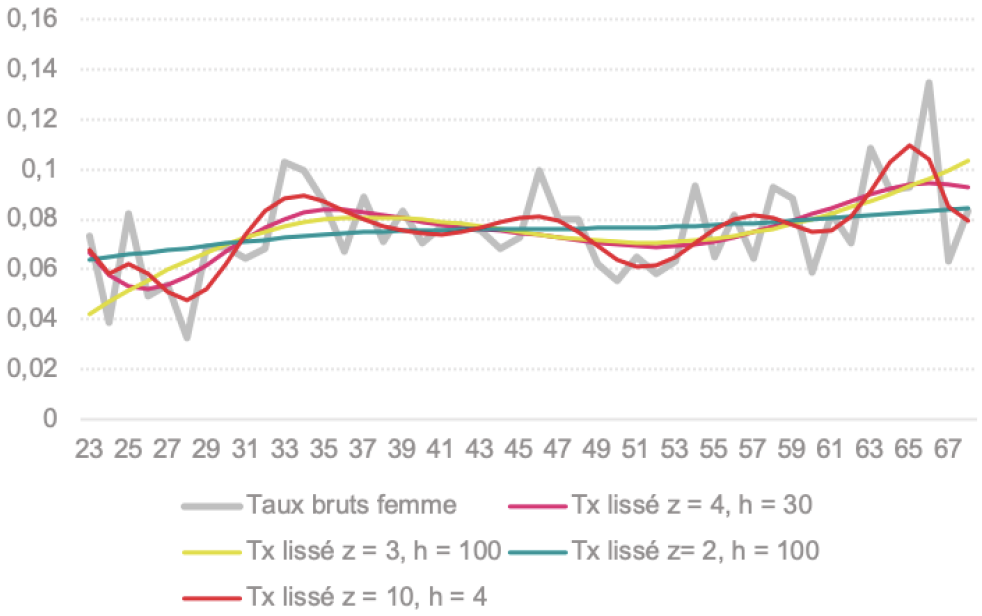


FIGURE 2.14 – Comparaison différents lissages pour les taux bruts femmes

2.4.3 Test d'adéquation de l'ajustement et validation du lissage

Dans l'objectif de valider les taux lissés précédemment, nous allons mettre en place des tests afin de nous assurer que les taux lissés ne soient pas trop éloignés des taux bruts et choisir le lissage le plus adéquat.

Le critère SMR (*Standardized Mortality Ratio*)

Le SMR est défini comme le rapport du nombre d'entrées à l'état prédit par la courbe des taux lissés et la courbe brute initiale. Ce ratio permet d'évaluer si le comportement de la courbe lissée

est proche de la courbe brute : si le ratio vaut 1, la courbe lisse est parfaitement identique à la courbe brute.

$$SMR = \frac{\sum_{i=1}^{n} q_i N_i}{\sum_{i=1}^{n} \hat{q}_i N_i} \tag{2.4}$$

Où $\begin{cases} q_i \text{ représente la probabilité brute à l'âge i d'entrer en AT} \\ \hat{q}_i \text{ représente la probabilité ajustée à l'âge i d'entrer en AT} \\ N_i \text{ représente le nombre de personnes à l'âge i} \end{cases}$

Ci-dessous les résultats, on voit qu'en moyenne les taux lissés sont très proches des taux bruts.

	Tx lissé z = 4, h = 30	Tx lissé z = 3, h = 100	Tx lissé z= 2, h = 100	Tx lissé z = 10, h = 4
Femme	99,99999997%	99,99999998%	99,99999997%	99,99999992%
Homme	99,99999998%	99,99999997%	99,99999998%	99,99999996%

FIGURE 2.15 – Tableau de résultat des lissages

Le critère de SMR est validé pour tous les lissages. Toutefois, il ne nous permet pas de sélectionner les paramètres optimaux. On se propose de sélectionner les courbes selon la cohérence de leurs tendances.

2.4.4 Validation et choix du lissage

Pour la courbe des hommes, nous sélectionnons la courbe avec z=4 et h=30. La tendance est plutôt constante, l'augmentation observée entre 50 et 63 ans est lissée. Enfin, la forte baisse peut se justifier par le départ à la retraite.

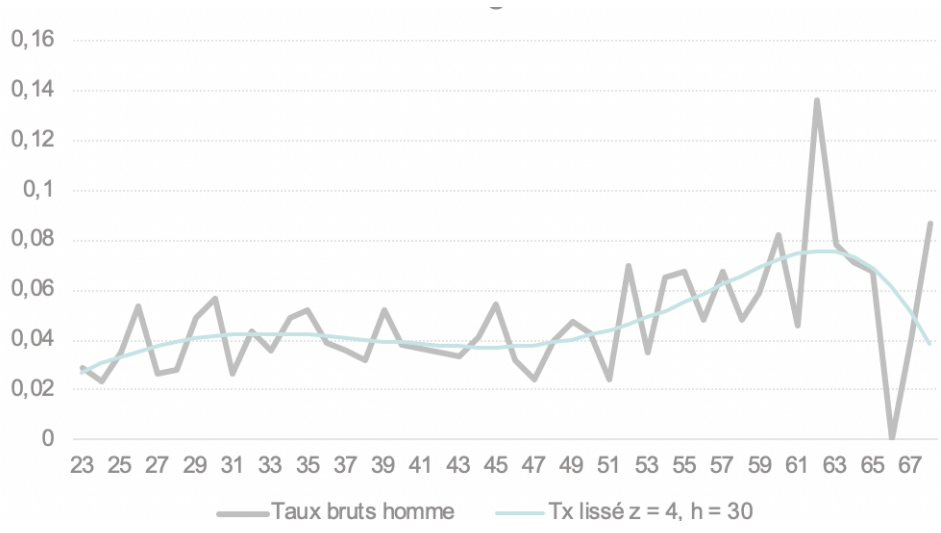


FIGURE 2.16 – Taux lissés des hommes

Concernant les taux d'entrée en AT des femmes, nous décidons de sélectionner le taux lissé avec z=10 et h=4. En effet, l'augmentation entre 30 et 45 ans peut s'expliquer par des arrêts liés à la maternité et est donc à conserver. L'augmentation suivie d'une baisse à partir de 65 ans est également présente chez les femmes. Nous optons pour le lissage de l'augmentation entre 62 et 64 ans et la conservation de la baisse à partir de 65 ans, cette dernière se justifiant par le départ à la retraite.

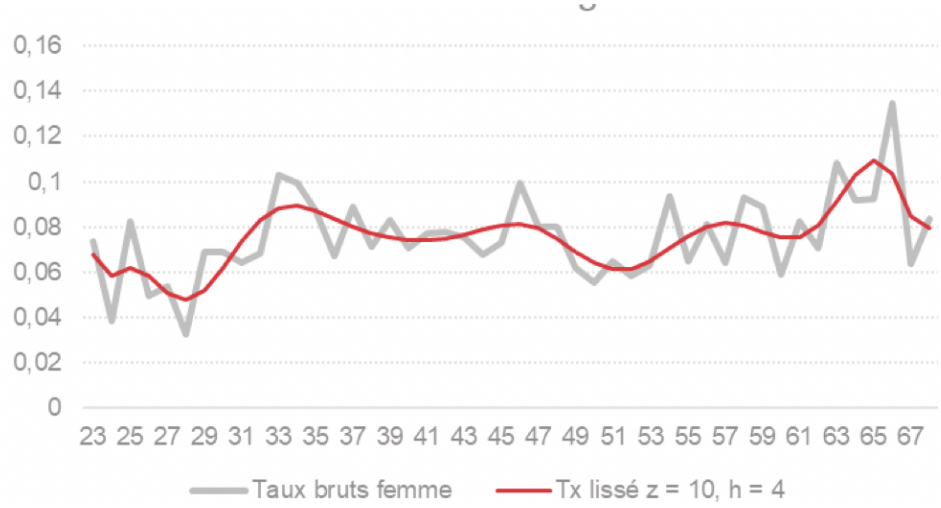


FIGURE 2.17 – Taux lissés des femmes

2.5 Premières analyses

2.5.1 Analyses statistiques

Avant toute modélisation, nous effectuons des analyses statistiques. Premièrement, nous comparons la consommation santé des individus ayant eu un arrêt avec celle des individus n'ayant pas eu d'arrêt. Ensuite, nous nous intéresserons aux taux d'entrée d'arrêt de travail en fonction des comportements de la population en termes de consommation médicale.

Comparaison de la consommation Santé moyenne

Nous nous sommes intéressés, dans un premier temps, à la comparaison des consommations santé durant les 12, 6 et 3 mois précédant un arrêt et la moyenne sur 12 mois de la consommation santé des assurés n'ayant eu aucun arrêt durant la période. Nous observions des différences significatives sur plusieurs postes.

En prenant l'exemple de l'hospitalisation, nous constatons qu'un individu ayant eu un arrêt de travail consomme 14% de plus sur le poste forfait journalier hospitalier qu'un individu n'en n'ayant pas eu un. Ce taux atteint les 177% lorsqu'on s'intéresse aux trois mois qui précèdent l'arrêt. Cette observation nous motive à nous intéresser aux périodes "plus proches" de l'arrêt : 6 et 3 mois. En effet, nous pensons que le lien est d'autant plus fort durant ces périodes.

FIGURE 2.18 – Moyenne de consommation Forfait Journalier sur 12, 6, et 3 mois

Comparaison du taux d'entrée en arrêt de travail

Dans un second temps, nous nous intéressons à la segmentation du portefeuille selon le comportement des assurés en termes de consommation médicale. La période d'observation choisie est celle de 12 mois. Nous évaluons le nombre d'individus ayant eu un arrêt dans la population sur le nombre total d'individus de la population.

Consultation spécialiste : Nous décidons de comparer le taux d'entrée eu arrêt des individus ayant eu entre 0 et 5 actes de consultation chez un spécialiste et ceux ayant eu plus de 5 actes. Nous constatons que le taux est de 5,5% chez la population peu consommatrice en consultation d'un spécialiste et de 9,9% chez la population plus consommatrice.

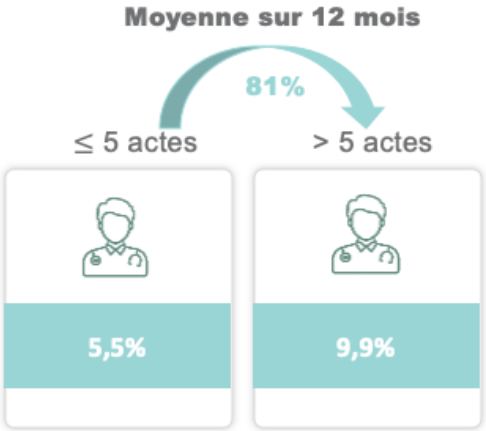


FIGURE 2.19 – Taux d'entrée en AT : Consultation spécialiste

Hospitalisation : En nous intéressant aux consommations d'hospitalisation, nous constatons que l'écart est d'autant plus fort sur des actes tels que l'hospitalisation. Le taux est de 3,7% chez la population peu consommatrice en hospitalisation et de 10,6% chez la population plus consommatrice. En effet, nous pensons que les variables liées à l'hospitalisation peuvent être plus discriminantes dans nos modèles ; une personne ayant eu plusieurs actes en hospitalisation peut

être plus susceptible d'avoir un arrêt de travail.

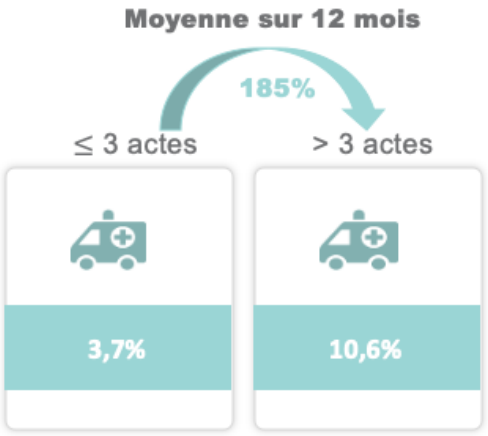


FIGURE 2.20 – Taux d'entrée en AT : Hospitalisation

Maternité : Le taux d'entrée en arrêt passe de 6,1% à 10,0% pour les assurés ayant plus d'un acte de maternité. Cela implique qu'un individu ayant plus d'un acte sur l'année aura plus de chances d'être en arrêt de travail, potentiellement dû à la maternité.

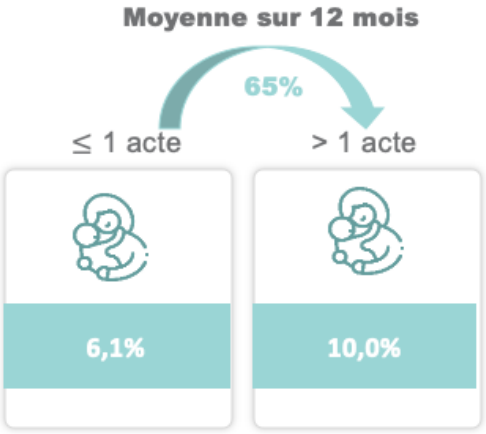


FIGURE 2.21 – Taux d'entrée en AT : Maternité

Chirurgie Optique : En nous intéressant aux actes de chirurgie optique, nous constatons que le taux d'entrée en arrêt passe de 6,1% à 14,3% pour les assurés ayant au moins un acte sur le poste. Un acte de chirurgie augmente considérablement la probabilité pour un individu d'être en arrêt. Toutefois, un arrêt lié à une chirurgie optique dure généralement entre 3 à 5 jours, le coût est ainsi limité pour un assureur.

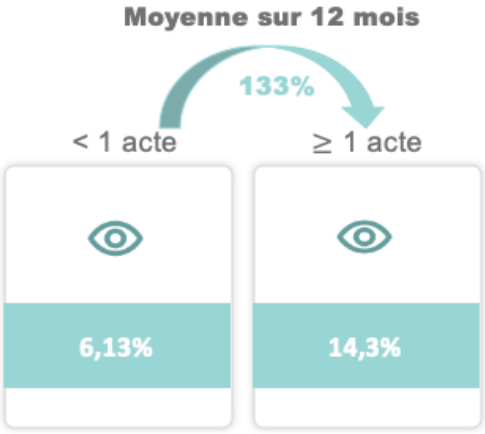


FIGURE 2.22 – Taux d'entrée en AT : Chirurgie Optique

Les statistiques révèlent des écarts de taux d'entrée en arrêt de travail entre différentes populations du portefeuille ayant des comportements différents en termes de consommation Santé.

2.5.2 Test statistique : Kolmogorov-Smirnov

En complément des premières analyses, nous décidons de réaliser un test statistique : le test de Kolmogorov-Smirnov. Ce test est à la base un test d'ajustement à une loi continue et suppose donc de connaître la fonction de répartition de la variable. Toutefois, il s'étend à la comparaison de deux fonctions de répartition empiriques et permet ainsi de tester si ces échantillons suivent la même loi. Soit $X_1, X_2, ..., X_n$ *iid* de fonction de répartition F et $Y_1, Y_2, \ldots, Y_m$ *iid* de fonction de répartition G. Les hypothèses à tester sont :

$$H0 : F = G$$

$$H1 : F \neq G$$

La statistique du test est définie comme suit :

$$D_{n,m} = \sup_{x \in \mathbb{R}} \mid F_n(x) - G_m(x) \mid \tag{2.5}$$

$$\text{Où} \begin{cases} F_n \text{ est la fonction de répartition de } X_1, X_2, ..., X_n \\ G_m \text{ est la fonction de répartition de } Y_1, Y_2, \ldots, Y_m \end{cases}$$

On rejette H_0 si $D_{n,m} \geq d_{n,m,1-\alpha}$ où $d_{n,m,1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi uniforme entre 0 et 1. Si l'hypothèse est rejetée, la *p-value* est inférieure à 0.05 et on peut alors dire que les deux distributions testées sont différentes. Sinon, on ne peut rien dire.

Dans notre cas, nous voulons comparer la consommation médicale des personnes ayant eu un arrêt

de travail avec celle des personnes n'ayant pas eu un arrêt. Nous nous intéressons notamment à la consommation sur les 12, 6 et 3 mois précédant l'arrêt et comparons poste par poste à la moyenne de la consommation des individus n'en ayant pas eu.

Le tableau ci-dessous regroupe les *p-values* des différents tests réalisés. Nous repérons 3 variables dont les *p-values* sont bien supérieures à 5% ; elles sont égales à 1.

	12 mois	6 mois	3 mois
Optique_chirurgie_AVT	1	1	1
Chambre_Part_AVT	1,82E-07	9,27E-07	3,43E-10
Maternite_AVT	4,8E-06	4,63E-08	1,18E-09
Sejour_Hospi_AVT	0	0	0
Forfait_Jour_Psy_AVT	1	1	1
Forfait_Jour_Hospi_AVT	1,85E-11	3,4E-08	3,89E-05
Consult_Gen_AVT	0	0	0
Consult_Spe_AVT	0	0	0
Cure_AVT	1	1	1
SOINS_AUTRES_AVT	0	0	0
Dispositifs_med_AVT	0	0	0
Dentaire_AVT	0	0	0
PHARMACIE_AVT	0	0	0
MEDECINE_DOUCE_AVT	0	0	0

FIGURE 2.23 – Test statistique : Kolmogorov -Smirnov

On peut supposer que la consommation de cures n'impacte pas le fait ou non d'avoir un arrêt. En ce qui concerne la chirurgie optique et le forfait journalier psychiatrique, nous pensons que cela est lié à notre périmètre ou est dû au nombre limité d'actes dont nous disposons dans notre base de données. En effet, nous observons des taux d'entrée en arrêt de travail différents en fonction des consommations sur ces postes. Nous supposons que le manque de volumétrie de données de consommations de ces actes justifie les résultats observés dans cette partie. Néanmoins, nous décidons de garder l'intégralité des variables pour le moment.

Chapitre 3

Modélisation de l'entrée en arrêt de travail

Sommaire

Il convient avant toute régression d'étudier les interactions entre les variables explicatives. En effet, la colinéarité de ces dernières peut s'avérer être un problème. Quand cette dernière est conséquente, elle risque d'augmenter la variance des coefficients de régression, les rendant ainsi instables et difficiles à interpréter. La colinéarité n'aura pas d'impact direct sur les prédictions du modèle, mais plutôt sur les coefficients individuels associés à chaque variable explicative. Or, la grande force des régressions logistiques est l'utilisation et la facilité d'interprétation de ces coefficients individuels.

3.1 Coefficient de corrélation

Nous nous intéressons dans cette partie à la corrélation des variables et non à leur colinéarité. Toutefois, des variables colinéaires sont nécessairement corrélées. Les coefficients de corrélation permettent de mesurer l'intensité et le sens d'une relation quand cette dernière est monotone. Dans ce qui suit, nous présenterons les mesures de corrélations les plus utilisées : Pearson et Spearman.

On rappelle que ces derniers varient entre -1 et 1.

— Si la valeur est proche de 0, pas de relation détectée ;

— Si la valeur est proche de -1, une "forte" relation négative est détectée ;
— Si la valeur est proche de +1, une "forte" relation positive est détectée ;

3.1.1 Pearson

Le coefficient de Pearson permet de détecter la présence ou non d'une relation linéaire entre deux variables quantitatives continues.

Ce coefficient est défini comme suit avec X et Y deux variables quantitatives continues :

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{3.1}$$

$$\text{Avec} \begin{cases} Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \\ \text{Ou} \\ Cov(X,Y) = (\frac{1}{N}\sum_{i=1}^{n}(X_i Y_i)) - (\bar{X}\bar{Y}) \end{cases}$$

3.1.2 Spearman

Le coefficient de Spearman permet de détecter la présence ou non d'une relation monotone (linéaire, exponentielle, puissance ...). Il est défini comme suit :

$$\rho = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}} \tag{3.2}$$

$$\text{Avec} \begin{cases} R_i \text{ le rang de l'observation de } X_i \\ S_i \text{ le rang de l'observation de } Y_i \end{cases}$$

Pour la suite des travaux, nous choisissons de retenir le rho de Spearman comme mesure de corrélation.

3.2 Matrice de corrélation

Une matrice de corrélation regroupe les corrélations entre les variables d'une base de données, elle constitue un bon outil pour détecter et ainsi pouvoir analyser et comprendre les relations entre ces variables. Dans un premier temps, nous expliciterons les critères d'exploitation d'une corrélation. Nous nous intéresserons aux matrices de corrélation de nos différentes bases de données et sélectionnerons les interactions à exploiter.

Nous choisissons d'appliquer les règles suivantes :

— Lorsque deux variables ont une corrélation jugée raisonnable, les deux variables sont conservées pour la modélisation ;

— Lorsqu'elles ont une corrélation supérieure à 0.6, une corrélation jugée forte, un choix entre ces deux variables doit être effectué ;

— Lorsqu'elles sont corrélées, mais pas fortement, nous étudions au cas par cas. En effet, certaines variables explicatives sont importantes sur le plan opérationnel, il est difficile de les retirer de l'étude ;

rho de spearman : ρ	Niveau de corrélation	Variable jugée importante	Variable jugée importante
ρ<0,4	raisonnable	Conservation	Conservation
0,4<ρ<0,6	modérée	Conservation	Suppression ou retraitement
ρ≥0,6	forte	Suppression ou retraitement	Suppression ou retraitement

FIGURE 3.1 – Règles de décision : matrice de corrélation

On observe ci-dessous la matrice de corrélation de la base de données avec le recul de 12 mois, 6 mois et 3 mois. On remarque que de nombreuses variables sont corrélées entre elles. En effet, des consommations santé peuvent entraîner d'autres consommations santé. Par exemple, une consultation chez un médecin généraliste est souvent suivie par des consommations en pharmacie ou de dispositifs médicaux ou d'autres soins ou encore d'une consultation chez un spécialiste.

On se propose d'étudier ces corrélations de deux manières différentes :

— Supprimer les variables qui ne nous semblent pas discriminantes et rajouter au modèle les interactions du reste des variables corrélées ;

— Ajouter aux modèles l'intégralité des interactions et appliquer une régression pénalisée de type Lasso ;

Nous décidons de ne pas supprimer de variables et d'étudier les interactions de l'ensemble des variables possédant une corrélation supérieure à 0,4 dans un premier temps.

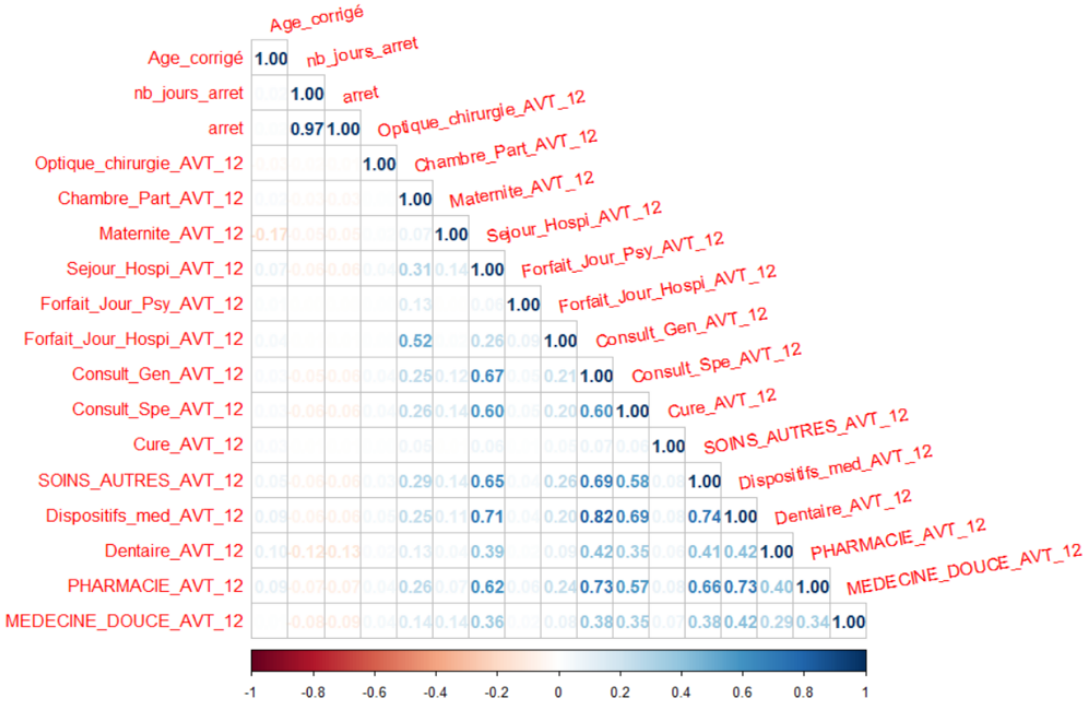


FIGURE 3.2 – Matrice de corrélation sur la base de 12 mois

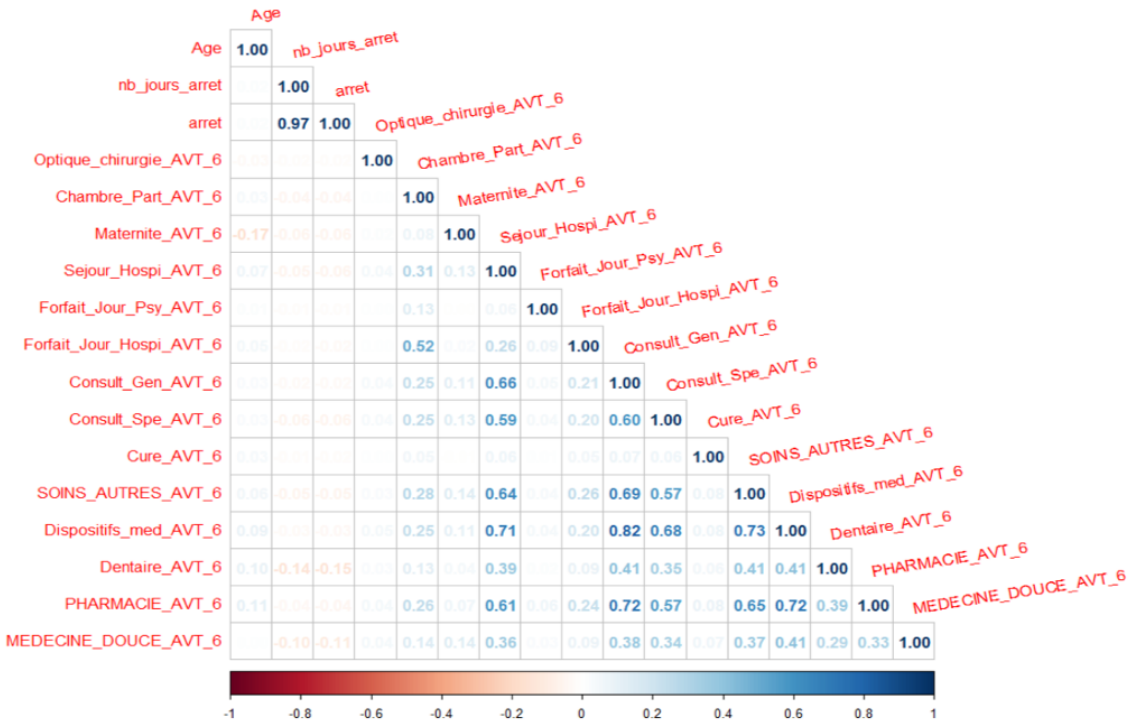


FIGURE 3.3 – Matrice de corrélation sur la base de 6 mois

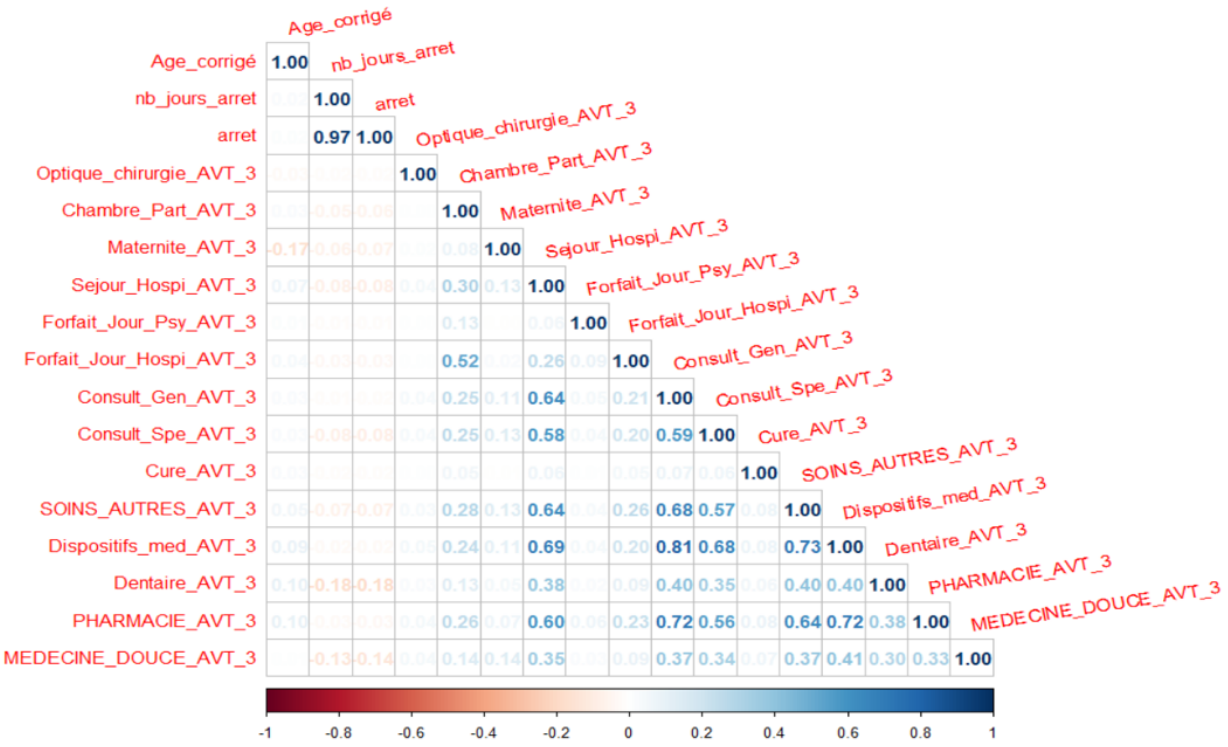


FIGURE 3.4 – Matrice de corrélation sur la base de 3 mois

3.3 Métriques de validation de modèle

Le critère de validation de modèle le plus connu est le critère des moindres carrés ordinaires (MCO). Comme son nom l'indique il s'agit de la moyenne des erreurs ou écart entre la prédiction du modèle et la valeur réelle.

$$MCO = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{yi})^2 \tag{3.3}$$

Avec $\begin{cases} N \text{ le nombre total d'observations} \\ y_i \text{ l'observation réelle} \\ \hat{y}_i \text{ la valeur prédite} \end{cases}$

Ce critère n'est pas très adéquat pour les modèles logistiques binaires. La valeur réelle peut prendre deux valeurs : 0 ou 1 ; alors que la valeur prédite est une probabilité comprise entre 0 et 1. Dans la suite, nous présenterons d'autres métriques et outils plus adéquats :

- Matrice de confusion ;
- Accuracy ;
- Spécificité ;

— Précision ;
— Sensibilité ;
— F1-Score ;
— Courbe ROC.

3.3.1 Matrice de confusion

La matrice de confusion regroupe des indicateurs de performance du modèle : les vrais positifs, les faux positifs, les vrais négatifs et les faux négatifs. Les vrais positifs correspondent aux cas où le modèle prédit correctement l'événement, dans notre cas l'arrêt de travail. On parle de faux positifs lorsque le modèle prédit un arrêt de travail alors que l'assuré n'en a pas eu. Les vrais négatifs correspondent aux cas où le modèle prédit correctement que l'individu n'a pas eu d'arrêt. Enfin, les faux négatifs sont les cas où le modèle prédit qu'il n'y a pas d'arrêt lorsqu'il y en a eu un. Cette matrice est la base de tous les calculs qui vont suivre.

		Valeur Prédite	
		0	1
Valeur réelle	0	Vrai Négatif	Faux Positif
	1	Faux Négatif	Vrai Positif

FIGURE 3.5 – Matrice de confusion théorique

Il est important de déterminer les paramètres qu'on vise à optimiser. Dans notre cas et pour des raisons de prudences, on vise à :

— Avoir le minimum de faux négatifs ;
— Avoir un maximum de vrais positifs ;

3.3.2 *Accuracy*

L'*Accuracy* calcule la proportion de bonnes prédictions sur l'ensemble de prédictions.

$$accuracy = \frac{VN + VP}{VN + VP + FN + FP} \tag{3.4}$$

$$\text{Avec} \begin{cases} VN \text{ les Vrais Négatifs} \\ VP \text{ les Vrais Positifs} \\ FN \text{ les Faux Négatifs} \\ FP \text{ les Faux Positifs} \end{cases}$$

La limite de cette métrique est qu'elle n'indique pas les forces et les faiblesses du modèle. De plus, dans le cas de données déséquilibrées, cette mesure n'est pas la plus adéquate pour mesurer la performance d'un modèle car elle est fortement influencée par le poids de la classe majoritaire, les non-arrêts.

Spécificité

La spécificité est la proportion de négatifs sur le nombre réel de négatifs. Cette métrique nous intéresse davantage et sera utilisée par la suite.

$$specificite = \frac{VN}{VN + FP} \tag{3.5}$$

Précision

La précision est la proportion de vrais positifs sur le nombre total de positifs prédits. Cette dernière permet de mesurer le coût des faux positifs.

$$precision = \frac{VP}{VP + FP} \tag{3.6}$$

Sensibilité

La sensibilité correspond au nombre de vrais positifs sur le nombre réel de positifs. Cette métrique nous intéresse particulièrement et sera utilisée par la suite.

$$sensibilite = \frac{VP}{VP + FN} \tag{3.7}$$

3.3.3 F1 - Score

Le F1-score est une mesure globale de la performance du modèle. Il correspond à la combinaison de la précision et de la sensibilité. Dans notre cas de base déséquilibrée, il peut être plus intéressant que l'*Accuracy* car le nombre de vrais négatifs n'est pas pris en compte.

$$F1Score = 2 \times \frac{precision \times sensibilite}{precision + sensibilite} = 2 \times \frac{\frac{VP}{VP+FP} \times \frac{VP}{VP+VF}}{\frac{VP}{VP+FP} + \frac{VP}{VP+FN}} \tag{3.8}$$

3.3.4 Courbe ROC

La courbe ROC (*Receiver Operating Characteristic*) est une représentation graphique de la performance du modèle. Elle met en relation la sensibilité et la spécificité. L'aire sous la courbe est comprise entre 0.5 et 1 : 0.5 étant le cas aléatoire et 1 la performance parfaite. Plus l'aire sous la courbe est élevée, meilleure est la performance de classification.

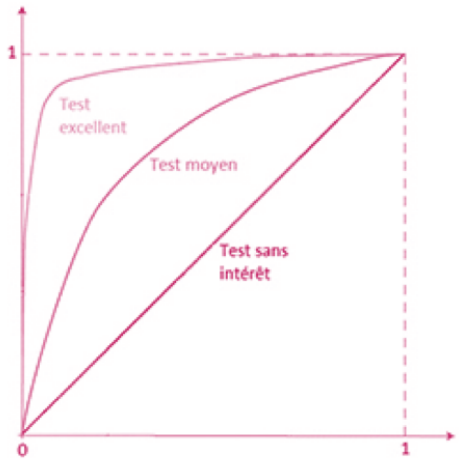


FIGURE 3.6 – Courbe ROC théorique

Dans la suite des travaux, et au vu de nos données déséquilibrées, nous décidons d'utiliser les métriques de sensibilité, de spécificité et de F1-Score pour comparer nos modèles.

3.4 Régression logistique

3.4.1 Rappels théoriques

Notre but est de prédire la probabilité d'entrée en arrêt par les différentes variables présentées précédemment. Lors d'une régression logistique binaire, Y, la variable cible, possède deux modalités 0 et 1.

$$Y = f(X, \alpha) \tag{3.9}$$

$$\text{Avec} \begin{cases} f \text{ le modèle de prédiction} \\ X \text{ la matrice composée des variables explicatives} \\ \alpha \text{ le vecteur de paramètres de la fonction f} \end{cases}$$

Dans le cadre bayésien, on s'intéresse notamment aux probabilités conditionnelles pour chaque modalité de la variable Y :

$$P[Y(\omega) = y_k \mid X(\omega)] \tag{3.10}$$

Dans notre cas binaire :

$$\begin{aligned} P[Y = 0 \mid X] &= \frac{P(Y = 0) \times P[X \mid Y = 0]}{P(X)} \\ &= \frac{P(Y = 0) \times P[X \mid Y = 0]}{P(Y = 0) \times P[X \mid Y = 0] + P(Y = 1) \times P[X \mid Y = 1]} \end{aligned} \tag{3.11}$$

De nombreux modèles existent, nous décidons d'appliquer le modèle Logit parce que ce dernier est adapté au cadre binaire et ses coefficients α des variables explicatives sont faciles à interpréter.

Pour un individu donné ω, la transformation Logit est la suivante :

$$ln\left[\frac{\pi(\omega)}{1-\pi(\omega}\right] = \alpha_0 + \alpha_1 X_1 + ... + \alpha_n X_n \tag{3.12}$$

$$\text{Avec} \begin{cases} \pi \text{ la probabilité d'entrée en AT} \\ \alpha_0, \alpha_1, ...\alpha_n \text{ les paramètres du modèle} \\ X_0, X_1, ...X_n \text{ les variables explicatives} \end{cases}$$

L'estimation des paramètres est réalisée par la méthode de maximum de vraisemblance. Dans le cas binaire, on peut utiliser une loi binomiale $B(1, \pi)$:

$$P[Y(\omega) = y_k \mid X(\omega)] = \pi(\omega)^{y(\omega)} \times (1 - \pi(\omega))^{1-y(\omega)} \tag{3.13}$$

La vraisemblance s'écrit alors :

$$L = \prod_{\omega} \pi(\omega)^{y(\omega)} \times (1 - \pi(\omega))^{1-y(\omega)} \tag{3.14}$$

$$\text{Pour} \begin{cases} y(\omega) = 1, P[Y(\omega) \mid X(\omega)] = \pi \\ y(\omega) = 0, P[Y(\omega) \mid X(\omega)] = 1 - \pi \end{cases}$$

Le problème s'écrit donc sous la forme suivante :

$$\max_{C(X)} \frac{1}{n} \sum_{i=1}^{n} [y_i \times C(X) - ln(1 + C(X)] \tag{3.15}$$

De plus, la vraisemblance étant une fonction convexe, le problème admet une unique solution.

Enfin, l'estimateur $\hat{\alpha} = \hat{\alpha_1}, \hat{\alpha_2}, ..., \hat{\alpha_n}$ possède les caractéristiques suivantes :

- Asymptotiquement sans biais ;
- Asymptotiquement gaussien ;
- Asymptotiquement minimal.

3.4.2 Classification binaire

Toute la difficulté du GLM réside dans la classification binaire de la prédiction du score. En effet, le modèle prédit une probabilité d'entrée en arrêt. Tandis que la variable d'intérêt, elle, est égale à 1 ou à 0. Pour y remédier, il faut fixer un seuil à partir duquel on prédit un arrêt. Soit π le seuil et p la probabilité prédite par le modèle.

$$\begin{cases} \text{Si } \pi \leq p \text{ Alors arret=0} \\ \text{Si } \pi > p \text{ Alors arret=1} \end{cases}$$

Par défaut, il est commun de considérer un $seuil = 0.5$. Toutefois, nos données sont fortement déséquilibrées, rendant ainsi l'utilisation du seuil par défaut inutilisable. Par ailleurs, le choix du seuil dépend des préférences et des objectifs de l'étude. En effet, ce choix affecte la précision du test statistique. Dans le cas où l'on souhaite minimiser les faux positifs, il est préférable de choisir un seuil plus élevé afin d'augmenter la probabilité de prédire correctement les événements positifs. En revanche, dans le cas où l'on souhaite minimiser les faux négatifs, il est préférable de choisir un seuil plus bas afin d'augmenter la probabilité de prédire correctement les événements négatifs.

Par la suite, nous chercherons à déterminer pour chaque modèle le seuil optimal.

3.4.3 Base référence : sans données de consommation santé

Nous tentons un premier modèle avec l'intégralité des variables avec recul de 12 mois ainsi que les interactions citées précédemment. Pour cela, nous divisons notre base en deux échantillons :

— La base d'apprentissage : base que nous utiliserons afin d'estimer les paramètres du modèle. (80% de la base initiale) ;
— La base de validation : base qui nous permettra d'évaluer la performance de notre modèle. (20% de la base initiale) ;

Nous tentons une première modélisation sans les données sur la consommation en santé. Les données utilisées sont les suivantes : âge, sexe et catégorie socioprofessionnelle. Le but est de constituer un modèle de base afin d'évaluer par la suite l'effet de l'ajout de nouvelles variables, dans notre cas la consommation en santé.

Nos données étant fortement déséquilibrées, nous déterminons le seuil optimal dans un premier temps. La démarche suivie sera détaillée par la suite.

Les résultats obtenus sont présentés ci-dessous :

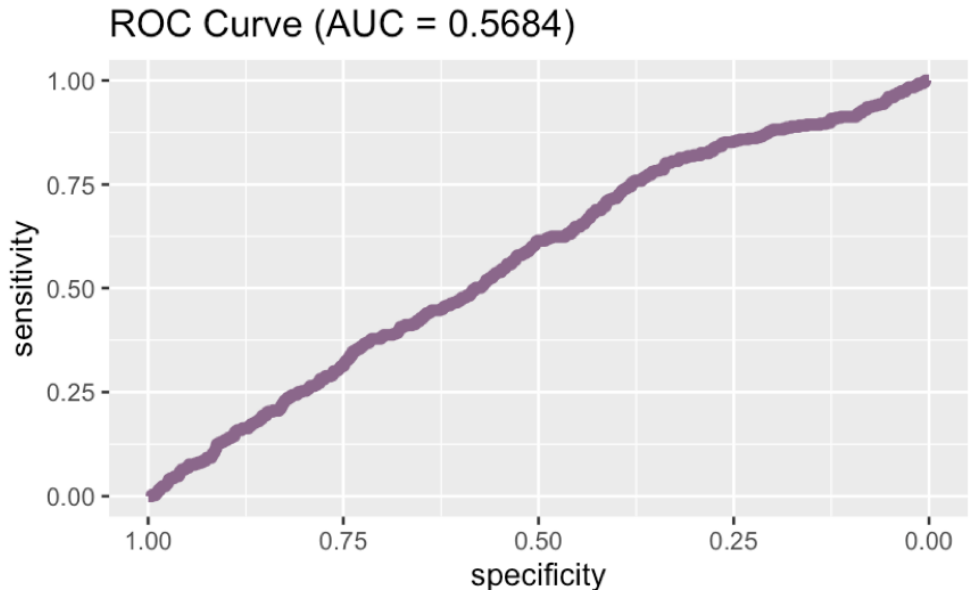


FIGURE 3.7 – Courbe ROC : base sans les données santé

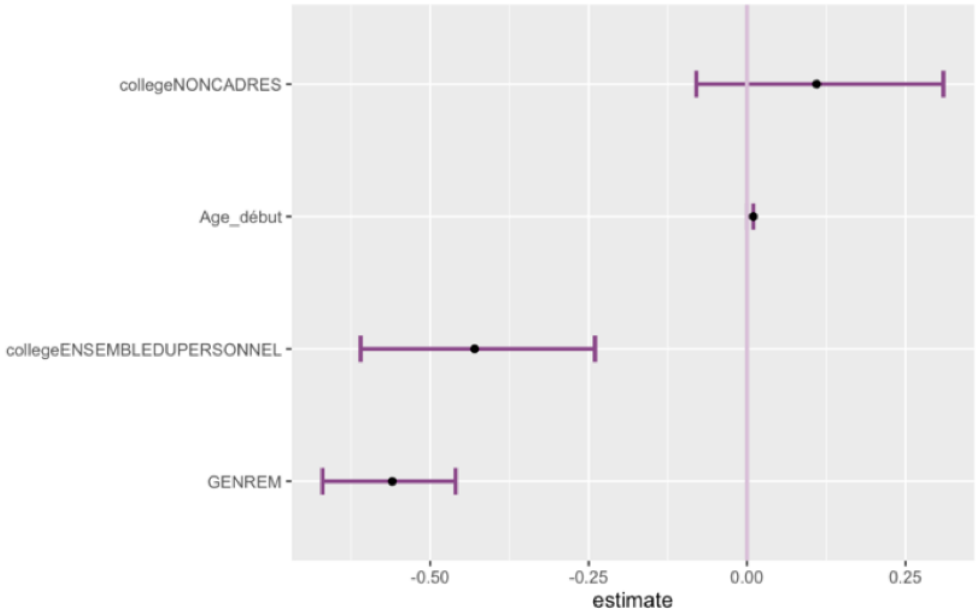


FIGURE 3.8 – Influence des variables : base sans les données santé

FIGURE 3.9 – Matrice de confusion : base sans les données santé

On remarque que la variable "college" semble être la plus déterminante dans ce modèle test. Toutefois, la longueur des intervalles de confiance rend difficile l'interprétation. L'AUC obtenu est faible ; de 0.568 et la matrice de confusion obtenue montrent que le pouvoir prédictif est pauvre : 92% des arrêts prédits n'en sont pas en réalité. Cela constitue un point d'attention et d'amélioration pour la suite. En effet, nous souhaitons adopter une démarche prudente dans cette étude mais sans pour autant surévaluer le risque.

Dans la suite, nous nous proposons d'ajouter les variables de consommation Santé afin d'évaluer l'apport de ces dernières.

3.4.4 Base : avec 12 mois de recul

Sélection des variables

Dans un premier temps, nous tentons une modélisation avec l'intégralité des variables et les croisements de variables et nous intéressons à leurs *Odds-Ratio*, l'intervalle de confiance et la *p-value* correspondante.

Nous rappelons qu'un *Odds-Ratio* (OR) mesure la relation existante entre une variable et la probabilité qu'un événement survienne, dans notre cas l'arrêt de travail. L'*Odds-Ratio* de la variable X_1 s'exprime comme suit :

$$OR = \frac{\frac{P(arret=1|X_1=1,\widetilde{X}_{N-1})}{P(arret=0|X_1=1,\widetilde{X}_{N-1})}}{\frac{P(arret=1|X_1=0,\widetilde{X}_{N-1})}{P(arret=0|X_1=0,\widetilde{X}_{N-1})}} \tag{3.16}$$

Avec $\widetilde{X}$ correspondant au reste des variables

Nous rappelons également que dans une régression logistique, les coefficients de la régression sont des *log odds ratio*. Un OR supérieur à 1 indique que la variable a un effet "positif" sur la variable cible. En d'autres termes, il y a une influence d'autant plus forte que la valeur de la variable est grande.

Dans le tableau ci-dessous, nous affichons uniquement les OR des variables, les croisements de variables ayant des *Odds-Ratio* égaux ou très proches de 1. Ainsi, on voit qu'un acte de séjour hospitalier augmente les chances de tomber en arrêt. Des consommations sur le poste de maternité ou encore une consultation chez un spécialiste semblent avoir le même effet, avec des intensités différentes. Tandis que les consommations en dentaire ou encore en médecine douce n'ont pas cet effet-là. Les coefficients observés correspondent à peu près aux intuitions que nous avions. Nous remarquons un effet "négatif" des postes dentaire, pharmacie ou encore médecine douce. On suppose que ces effets sont liés aux particularités de notre base et n'ont pas un lien direct avec la probabilité d'être en arrêt de travail. Nous décidons de les garder.

On rappelle que les *p-values* associées aux coefficients indiquent si la variable est significativement associée au risque de tomber en arrêt. Si la *p-value* est inférieure à un niveau de seuil prédéfini (souvent fixé à 5%), alors on peut rejeter l'hypothèse nulle selon laquelle le coefficient est égal à zéro ; la variable n'est pas significativement associée à la probabilité d'avoir un arrêt.

En fixant un seuil de 5%, trois variables sont "à supprimer" : Cures, Forfait jour psy et chambre particulière. Intuitivement, les cures n'ont pas de lien avec la probable occurrence d'un arrêt. Nous supposons que les deux dernières variables n'ont pas d'effets significatifs à cause du peu d'actes observés. Néanmoins, les intervalles de confiance étant larges, nous décidons de supprimer ces 3 variables.

Variables	Odds-Ratio	95% IC	p-value
Age_début	1.01	1.01, 1.02	<0.001
GENRE			
F	—	—	
M	0.63	0.57, 0.69	<0.001
college			
CADRES	—	—	
EDP	0.86	0.73, 1.01	0.061
NONCADRES	1.23	1.04, 1.46	0.017
Optique_chirurgie_AVT_12	1.61	1.16, 2.23	0.004
Chambre_Part_AVT_12	1.04	0.96, 1.12	0.3
Maternite_AVT_12	1.13	1.02, 1.25	0.015
Sejour_Hospi_AVT_12	1.33	1.31, 1.35	<0.001
Forfait_Jour_Psy_AVT_12	1.01	0.89, 1.15	0.8
Forfait_Jour_Hospi_AVT_12	1.23	1.18, 1.29	<0.001
Consult_Gen_AVT_12	0.95	0.93, 0.96	<0.001
Consult_Spe_AVT_12	1.11	1.09, 1.12	<0.001
Cure_AVT_12	1.20	0.89, 1.60	0.2
SOINS_AUTRES_AVT_12	0.98	0.97, 0.98	<0.001
Dispositifs_med_AVT_12	0.92	0.91, 0.93	<0.001
Dentaire_AVT_12	0.77	0.75, 0.80	<0.001
PHARMACIE_AVT_12	0.98	0.98, 0.99	<0.001
MEDECINE_DOUCE_AVT_12	0.81	0.77, 0.86	<0.001

FIGURE 3.10 – Odds-Ratio base 12 mois

Détermination du seuil

Afin de déterminer le seuil et dans le but d'optimiser le pouvoir prédictif de notre modèle, nous décidons de suivre la démarche que nous présenterons juste après. Comme expliqué précédemment, les résultats du modèle seront sous forme de probabilités d'entrée en arrêt de travail. Si la probabilité est supérieure à π, nous considérons que la personne aura un arrêt.

Dans un premier temps, nous appliquons le modèle à la base d'apprentissage afin de confronter les probabilités prédites avec les valeurs réelles de la variable "arret" en analysant les distributions de ces dernières.

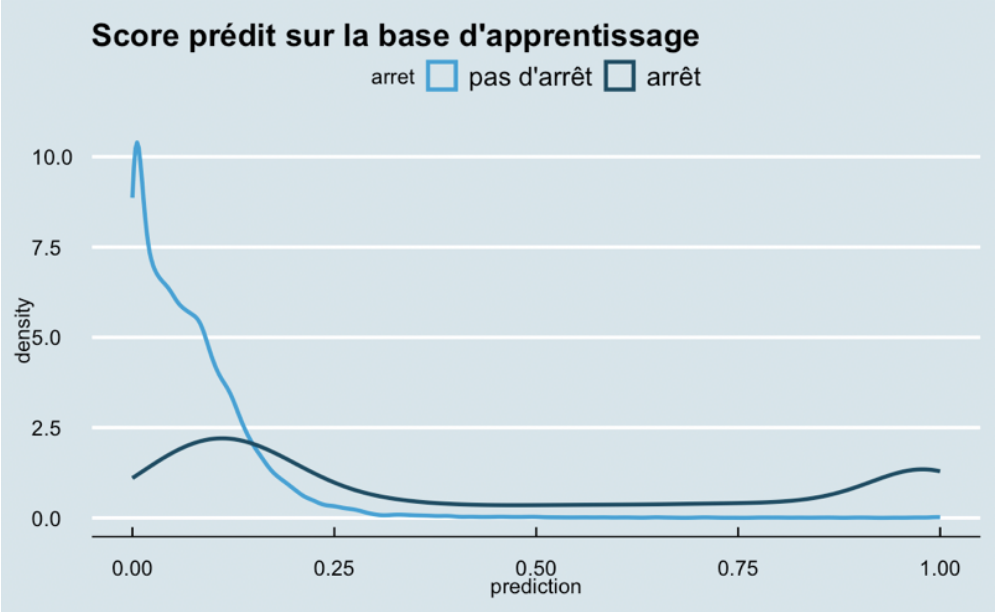


FIGURE 3.11 – Score prédit base 12 mois

Très peu d'individus n'ayant pas eu d'arrêt sur la période d'étude ont des probabilités prédites d'entrée en arrêt supérieures à 0.25 par notre modèle, ce qui est rassurant. Nous supposons ainsi que notre modèle est en mesure d'avoir une bonne spécificité. En revanche, la distribution de la modalité 1 de la variable "arret" (en bleu foncé sur le graphique) est plus étalée. En effet, idéalement, nous espérions qu'elle soit plus concentrée à droite du graphique, afin de faciliter la détermination du seuil et d'optimiser la sensibilité de notre modèle (*cf.* Annexe B). Or on observe une forte concentration entre 0.10 et 0.5 et une autre à partir de 0.85.

D'une part, ce graphique montre bien qu'un seuil de 0.5 aurait très bien prédit les 0 mais non les 1 ; aurait une bonne spécificité mais une mauvaise sensibilité. D'autre part, nous déduisons que la détermination du seuil n'est pas une tâche à négliger.

Nous choisissons de nous aider du package "*cutpointr*" disponible sur R. Comme son nom l'indique, ce package est conçu pour évaluer les *cutpoints* ou points de coupure optimaux. La fonction "*cutpointr*" permet d'utiliser différentes mesures que l'on peut maximiser ou minimiser dont :

- *Accuracy* : Mesure qu'on ne privilégie pas pour les raisons évoquées précédemment ;
- Roc01 : Mesure de la distance au point (0,1) de la courbe ROC, nous décidons de privilégier d'autres métriques ;
- F1- Score : Mesure que l'on teste dans ce qui suit ;
- Sum_Sens_Spec : Maximum de la somme de la spécificité et la sensibilité - mesure que l'on teste dans ce qui suit ;
- Prod_Sens_Spec : Maximum du produit de la spécificité et la sensibilité - mesure que l'on teste dans ce qui suit ;

Nous comparons les *cutpoints* obtenus et les différents résultats obtenus afin de décider de la méthode à appliquer.

Métrique utilisée : F1 Score

En sélectionnant le F1-Score comme mesure à maximiser, nous obtenons un point de coupure de 0.289. Nous nous intéressons, d'une part, à la visualisation du point de coupure sur la distribution des classes respectives de notre variable d'intérêt et d'autre part, à la matrice de confusion obtenue.

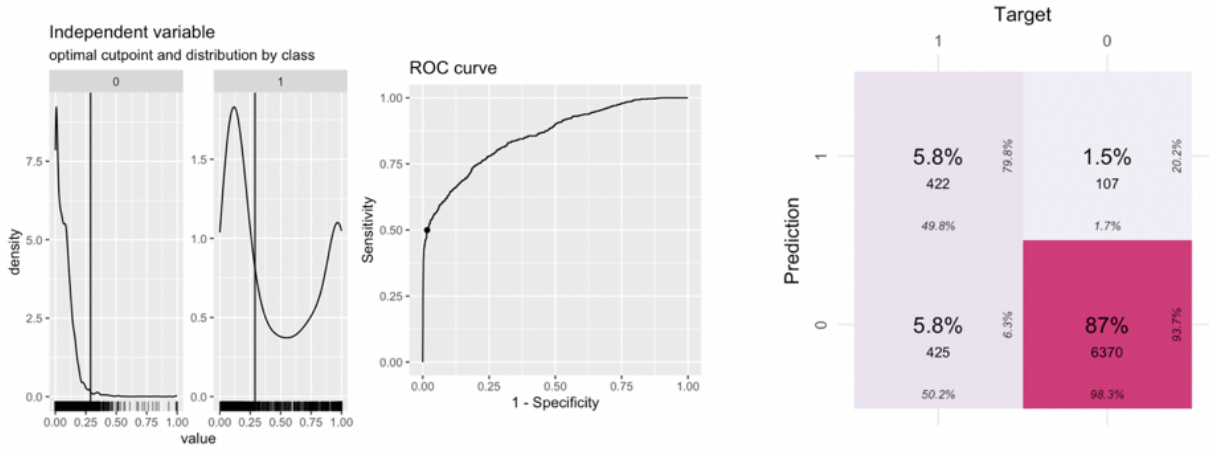


FIGURE 3.12 – Résultats : maximisation de la métrique F1-Score

En nous intéressant aux graphiques de la distribution par classe de la variable arrêt, nous comprenons que le modèle aura plus de mal à prédire les 1 ; une partie de la distribution avant le *cutpoint* ne sera pas détectée. Tandis qu'une "majorité" du signal 0 pourra être détectée ; une relativement plus petite partie de la distribution est à droite du *cutpoint*.

Le F1 Score obtenu et donc le meilleur possible à ce stade est de 0.613. D'après les graphiques ci-dessus, nous comprenons que la précision est bonne : 79.8% des arrêt prédits sont correctes. Mais aussi 93.7% des non-arrêts prédits sont corrects. Néanmoins, notre modèle ne prédit à ce stade que 50% des arrêts. On le juge "pas suffisamment prudent" et nous tenterons de l'améliorer par la suite.

Métrique utilisée : Somme de la sensibilité et de la spécificité

En sélectionnant la somme de la sensibilité et de la spécificité comme mesure à maximiser, nous obtenons un *cutpoint* de 0.124 ; bien inférieur à celui obtenu précédemment.

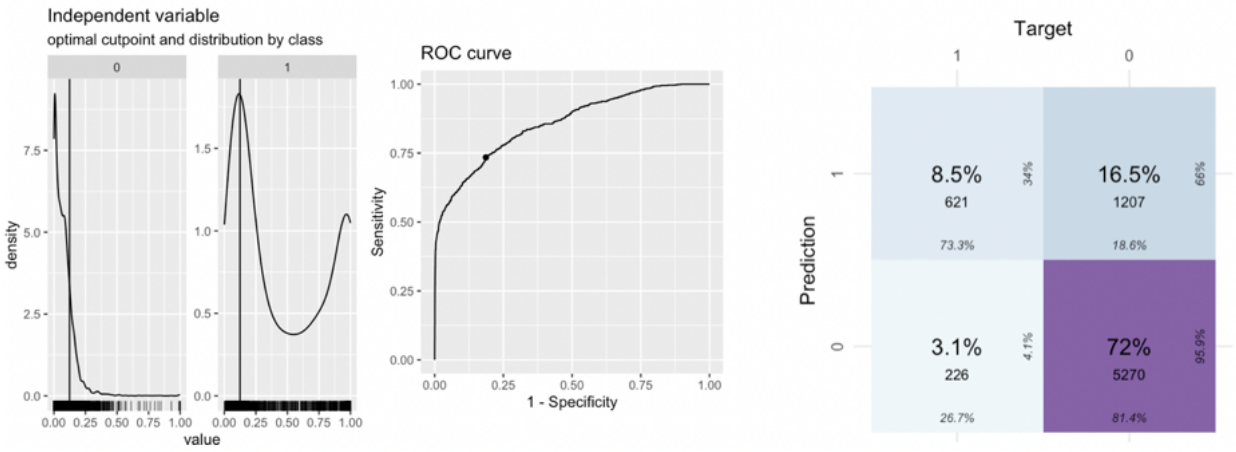


FIGURE 3.13 – Résultats : maximisation de la métrique de la somme de la sensibilité et de la spécificité

Le F1-Score de 0.464 est inférieur à celui obtenu lors de notre précédent essai. Cela se traduit par une forte dégradation de la précision qui est de 34% avec cette métrique ; seuls 34% des arrêts prédits sont en réalité des arrêts. Par ailleurs, ce seuil permet de prédire 73.3% des arrêts et est ainsi bien plus prudent.

Métrique utilisée : Produit de la sensibilité et de la spécificité

En sélectionnant le produit de la sensibilité et de la spécificité comme mesure à maximiser, nous obtenons un *cutpoint* de 0.122, très proche de celui obtenu avec la somme de la sensibilité et de la spécificité.

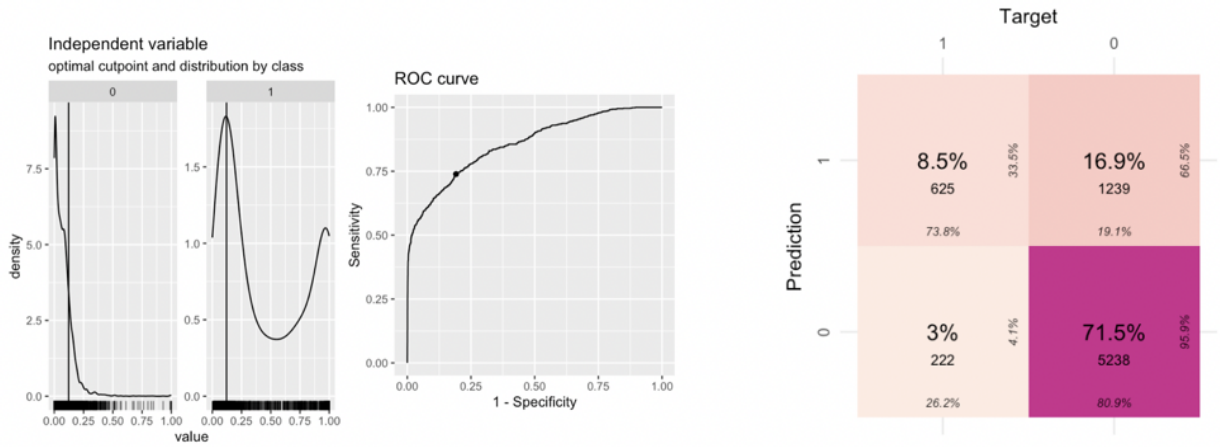


FIGURE 3.14 – Résultats : maximisation de la métrique du produit de la sensibilité et de la spécificité

Le F1-Score est de 0.461. La précision est de 33.5%, toujours faible. En outre, ce seuil permet de prédire 73.8% des arrêts et est ainsi bien plus prudent.

Analyse des résultats obtenus avec le seuil choisi

Nous décidons d'opter pour le seuil qui maximise le F1-Score à ce stade. En effet, nous désirons une démarche prudente mais sans pour autant surévaluer le risque. Nous tenterons par la suite d'améliorer la sensibilité du modèle.

Par ailleurs et à titre comparatif, nous nous intéressons à la courbe ROC ainsi qu'à l'AUC. Une amélioration flagrante est à noter par rapport au modèle dit "de base". En effet, la courbe est bien plus incurvée et l'AUC passe de 0.568 à 0.856. Le F1-Score de 0.133 à 0.613. Les données Santé apportent une réelle plus-value à notre modèle.

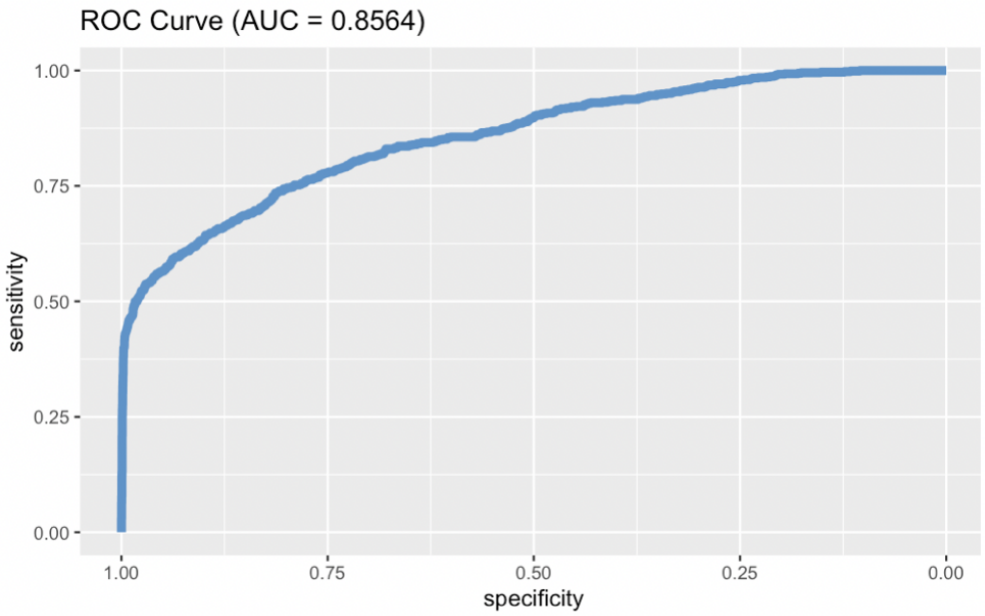


FIGURE 3.15 – Courbe ROC du modèle optimisé 12 mois

Nous décidons de nous intéresser par la suite à la consommation santé 6 mois et 3 mois avant l'arrêt. En effet, nous supposons que le lien existant entre ces deux risques apparaît peu de temps avant l'occurrence de l'arrêt et qu'il serait ainsi intéressant d'affiner l'étude.

3.4.5 Base : avec 6 mois de recul

Nous appliquons la démarche présentée précédemment à la base contenant les données santé avec le recul de 6 mois.

Premièrement, nous tentons une modélisation avec l'intégralité des variables et nous intéressons à leurs *Odds-Ratio*, l'intervalle de confiance et la *p-value* correspondante.

Variables	Odds-Ratio	95% IC	p-value
Age_début	1.01	1.01, 1.02	<0.001
GENRE			
F	—	—	
M	0.68	0.62, 0.74	<0.001
college			
CADRES	—	—	
EDP	0.88	0.75, 1.04	0.14
NONCADRES	1.38	1.17, 1.63	<0.001
Optique_chirurgie_AVT_6	1.41	0.98, 2.03	0.062
Chambre_Part_AVT_6	1.01	0.90, 1.13	0.9
Maternite_AVT_6	1.35	1.21, 1.50	<0.001
Sejour_Hospi_AVT_6	1.46	1.44, 1.49	<0.001
Forfait_Jour_Psy_AVT_6	1.03	0.87, 1.21	0.8
Forfait_Jour_Hospi_AVT_6	1.36	1.27, 1.45	<0.001
Consult_Gen_AVT_6	0.98	0.96, 1.00	0.049
Consult_Spe_AVT_6	1.21	1.18, 1.23	<0.001
Cure_AVT_6	0.66	0.26, 1.66	0.4
SOINS_AUTRES_AVT_6	0.97	0.97, 0.98	<0.001
Dispositifs_med_AVT_6	0.89	0.87, 0.90	<0.001
Dentaire_AVT_6	0.70	0.66, 0.73	<0.001
PHARMACIE_AVT_6	0.98	0.98, 0.98	<0.001
MEDECINE_DOUCE_AVT_6	0.73	0.68, 0.80	<0.00

FIGURE 3.16 – Odds-Ratio base 6 mois

Dans notre étude, les croisements de variables n'apportent pas de plus-value significative, on décide de les omettre de l'affichage.

En fixant un seuil de 5%, quatre variables sont "à supprimer" : Cures, Forfait jour psy, chambre particulière et chirurgie optique. Même si nous pensons que cela est une spécificité de notre périmètre et qu'une chirurgie optique ou une chambre particulière impactent le risque d'incapacité, les intervalles de confiance associés sont larges. Nous décidons de supprimer ces variables.

Nous obtenons les résultats suivants :

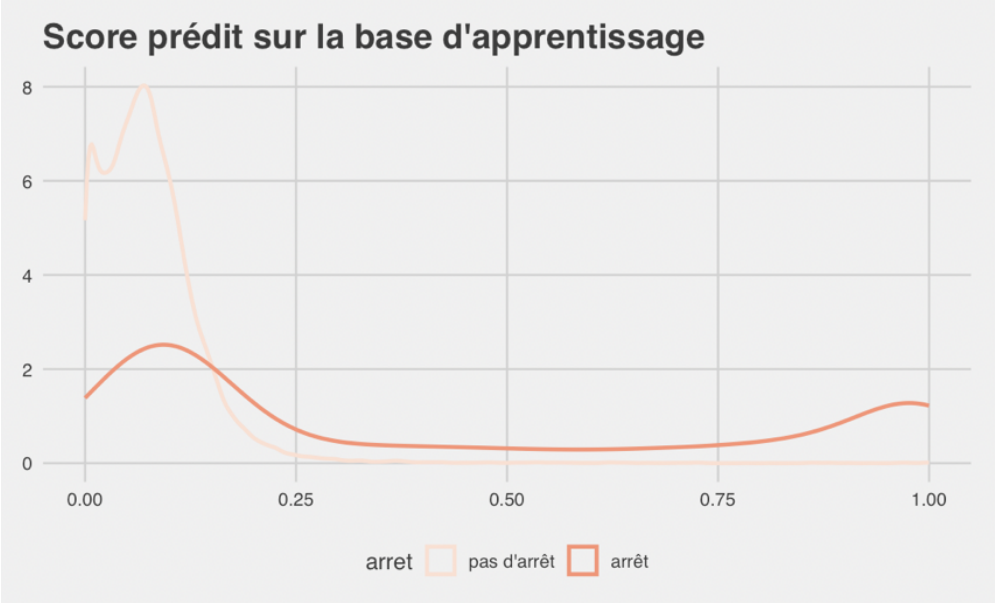


FIGURE 3.17 – Score prédit base 6 mois

La distribution ressemble à la distribution obtenue avec les consommations sur une durée de 12 mois. La distribution de la modalité 1 de la variable "arret" (en orange foncé sur le graphique) présente des concentrations aux mêmes endroits. Celle de la modalité 1 semble présenter une concentration encore plus forte avant 0.25.

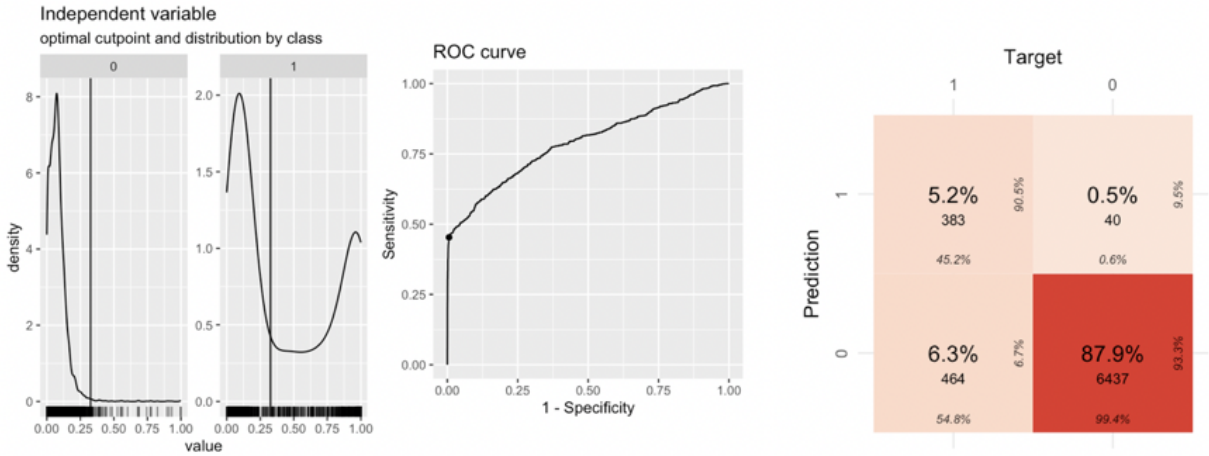


FIGURE 3.18 – Résultats : base 6 mois

Le F1-Score est 0.603. Nous notons une amélioration considérable de la précision : plus de 90% des arrêt prédits en sont réellement ; 99% des non-arrêt prédits sont corrects. Toutefois, nous rencontrons le même problème au niveau de la sensibilité.

3.4.6 Base : avec 3 mois de recul

Nous appliquons la démarche présentée précédemment à la base contenant les données santé avec le recul de 3 mois.

Dans un premier temps, nous tentons une modélisation avec l'intégralité des variables et nous intéressons à leurs *Odds-Ratio*, l'intervalle de confiance et la *p-value* correspondante.

Variables	Odds-Ratio	95% IC	p-value
Age_début	1.01	(1.01 to 1.02)	<0.001
GENRE			
F	—		
M	0.67	(0.61 to 0.73)	<0.001
college			
CADRES	—		
EDP	0.83	(0.70 to 0.97)	0.021
NONCADRES	1.35	(1.14 to 1.60)	<0.001
Optique_chirurgie_AVT_3	0.00	(0.00 to 0.00)	0.91
Chambre_Part_AVT_3	1.27	(1.09 to 1.48)	0.002
Maternite_AVT_3	1.42	(1.23 to 1.68)	<0.001
Sejour_Hospi_AVT_3	1.59	(1.55 to 1.63)	<0.001
Forfait_Jour_Psy_AVT_3	1.05	(0.86 to 1.37)	0.67
Forfait_Jour_Hospi_AVT_3	1.49	(1.37 to 1.64)	<0.001
Consult_Gen_AVT_3	1.03	(0.99 to 1.06)	0.11
Consult_Spe_AVT_3	1.39	(1.33 to 1.44)	<0.001
Cure_AVT_3	0.89	(0.28 to 1.90)	0.80
SOINS_AUTRES_AVT_3	0.98	(0.96 to 0.99)	<0.001
Dispositifs_med_AVT_3	0.84	(0.81 to 0.86)	<0.001
Dentaire_AVT_3	0.62	(0.57 to 0.67)	<0.001
PHARMACIE_AVT_3	0.97	(0.97 to 0.98)	<0.001
MEDECINE_DOUCE_AVT_3	0.63	(0.55 to 0.72)	<0.001

FIGURE 3.19 – Odds-Ratio base 3 mois

Les croisements de variables n'apportent pas de plus-value, on les omet de l'affichage. En fixant un seuil de 5%, quatre variables sont "à supprimer" : Cures, Forfait jour psy, consultation généraliste et chirurgie optique. Même si nous pensons que cela est une spécificité de notre périmètre et qu'une chirurgie optique ou une consultation chez un généraliste impactent le risque d'incapacité, les intervalles de confiance associés sont larges. Nous décidons de supprimer toutes ces variables sauf celle de consultation généraliste.

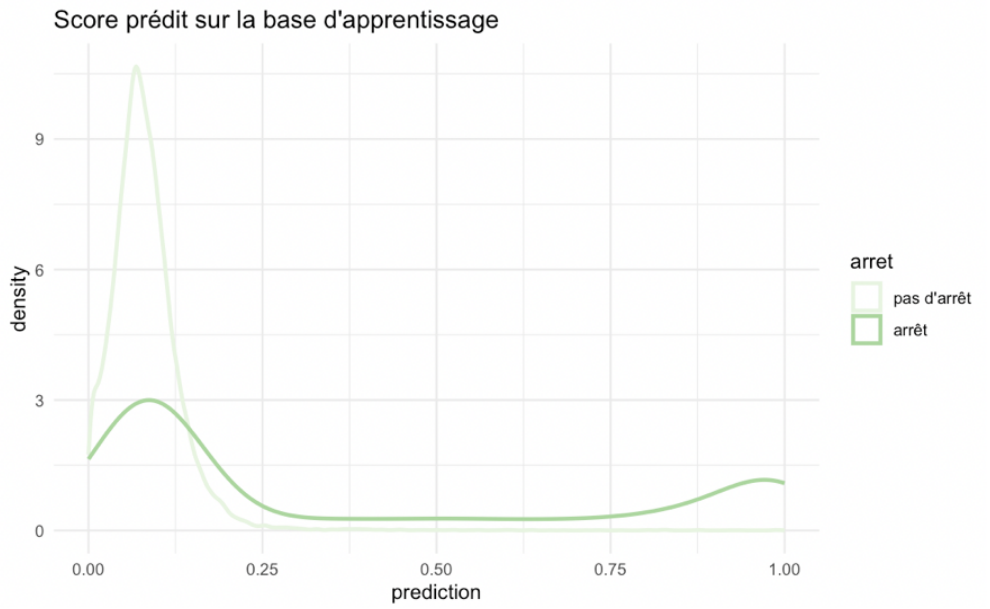


FIGURE 3.20 – Score prédit base 3 mois

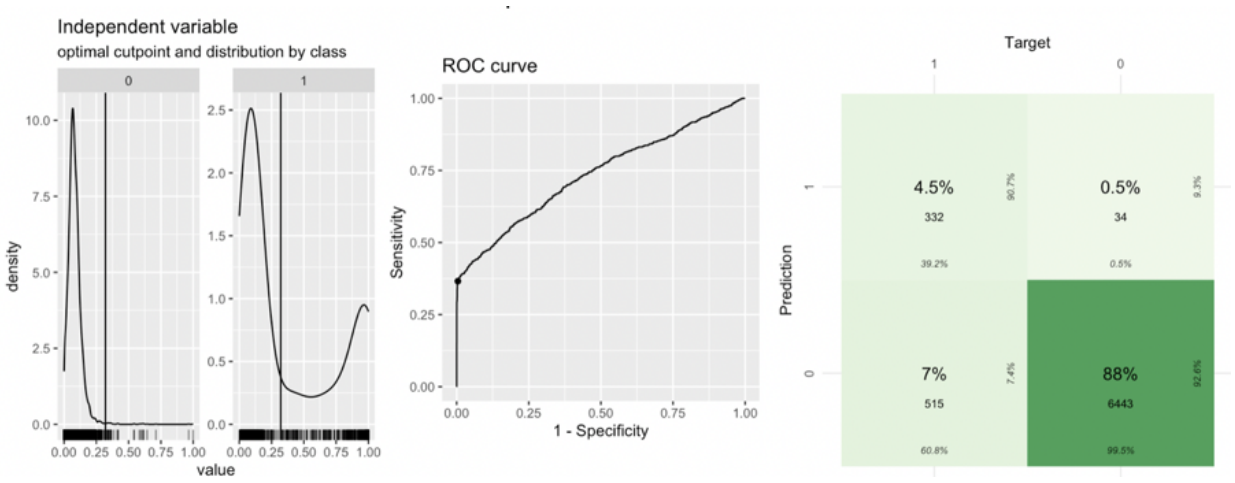


FIGURE 3.21 – Résultats : base 3 mois

En analysant la distribution des modalités de la variable arrêt, nous remarquons une concentration plus importante avant 0.25 de la modalité 1. Nous pensons que cela dégradera la sensibilité de notre modèle.

Le F1-Score est 0.547. Même s'il y a une amélioration de la précision ; plus de 90% des arrêts prédits en sont effectivement ; 99% des non-arrêt prédits sont corrects, nous ne prédisons que 39% des arrêts. Nous pensons ainsi à catégoriser nos variables, en espérant que notre modèle gagnera en sensibilité.

3.5 Affinement de la base

Afin d'affiner notre base et dans le but d'améliorer la sensibilité de nos modèles, nous décidons de catégoriser les variables. En effet, dans le cas où la volumétrie de données n'est pas suffisante

ou lorsqu'on dispose de variables asymétriques, il peut être judicieux de transformer les variables continues en variables catégorielles. Cette transformation permettra de réduire l'impact du bruit statistique. Nous tentons dans un premier temps un "*Binning* simple" en guise de modèle de base. Nous essaierons par la suite la méthode d'*Optimal Binning* et comparerons les résultats.

3.5.1 Catégorisation simple

Nous décidons dans un premier temps de catégoriser chaque variable continue en 3 catégories avec une distribution équitable entre les catégories afin d'éviter que l'une d'entre elle soit sur-représentée ou sous-représentée. Cette approche est relativement simpliste et a pour vocation de comparer les résultats obtenus avec la méthode d'*Optimal Binning*.

Après avoir lancé une première fois le modèle et supprimé les variables non-significatives, nous nous intéressons à la distribution des modalités de notre variable d'intérêt.

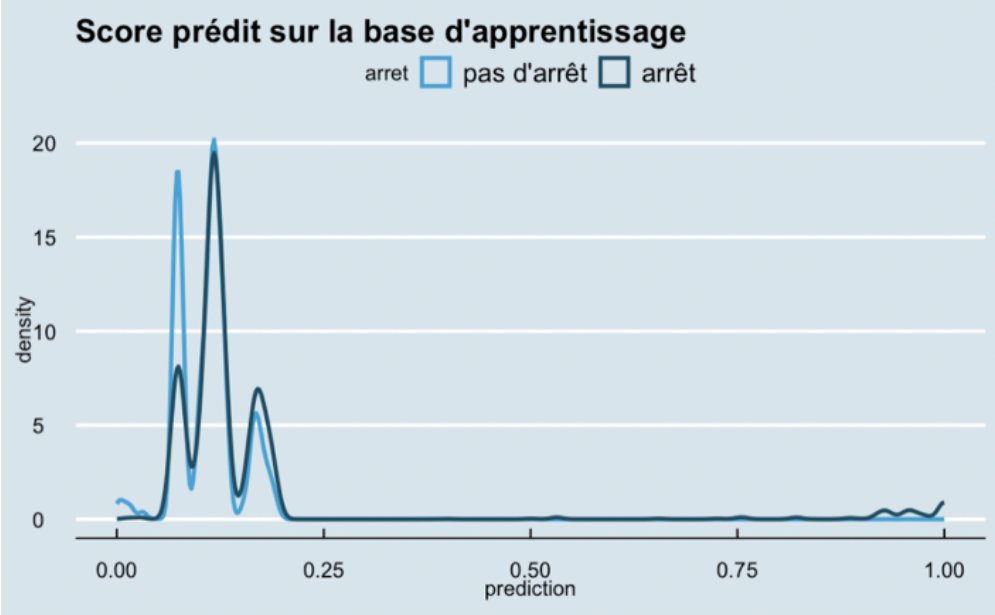


FIGURE 3.22 – Score catégorisation simple

La distribution diffère de celle qu'on a pu observer auparavant. Un bruit supplémentaire semble s'être ajouté. Nous nous attendons ainsi à des résultats bien dégradés.

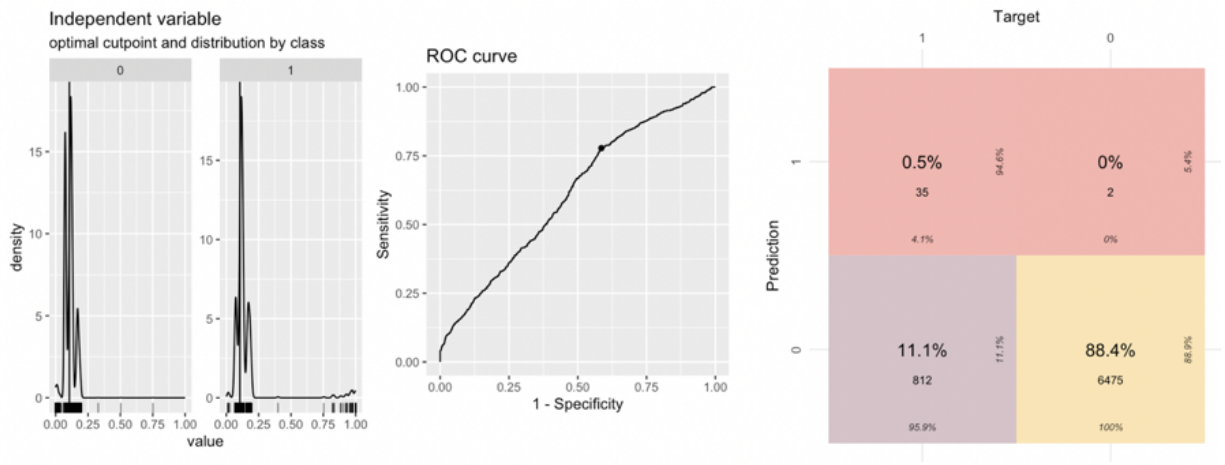


FIGURE 3.23 – Résultats catégorisation simple

Nous obtenons de mauvais résultats : un F1 Score de 0.0792. Le modèle ne capte quasiment pas le signal de la modalité arrêt : seuls 4.1% des arrêts sont prédits par ce modèle.

Nous voyons bien que l'analyse et le découpage des variables doit se faire de manière plus fine.

3.5.2 *Optimal Binning :* Présentation

L'*Optimal Binning* est une méthode de classification supervisée qui permet de discrétiser une variable numérique ou continue à l'aide d'un arbre CART en fonction de son effet sur une variable, dans notre cas, la variable "arret". L'algorithme regroupera des valeurs de la variable en des groupes, ou *bins*, qui ont un impact homogène. Ces points de coupes, ou *cutpoints*, sont sélectionnés de manière à optimiser le lien entre la variable numérique et la variable cible. Plusieurs critères de qualité existent telle que la maximisation de l'information de classe ou le critère de maximisation de la distance entre les groupes. Dans notre cas, nous retenons le critère d'information de classe.

L'algorithme consiste à :

1. Calculer les points de coupes et diviser en 2 segments.
2. Répéter la même procédure sur les segments obtenus tant que chaque segment est supérieur à une valeur donnée.

Cette valeur est définie par défaut dans la fonction "*smbinning*" sur R à 5%. Nous décidons de la conserver. Nous appliquons la fonction à chacune des trois bases. Dans ce qui suit, on se propose de tester la modélisation sur une base discrétisée via le *package* "*smbinning*" de R et comparer avec la base non discrétisée. Nous décidons de transformer toutes les variables numériques en variables catégorielles : l'âge ainsi que les différentes consommations santé.

3.5.3 *Optimal Binning* : avec 12 mois de recul

Nous commençons par catégoriser via "*smbinning*" les variables de la base contenant les données santé avec un recul de 12 mois. On remarque que la distribution de la courbe "pas d'arrêt" est bien

plus concentrée à gauche du graphique. Alors que la courbe "arrêt" parait plus étendue : moins de concentration observée à gauche du graphique. Cela prouve qu'une bonne catégorisation des variables pourrait affiner le modèle. On espère ainsi améliorer la sensibilité du modèle initiale avec les variables non catégorisées qui est de 50%.

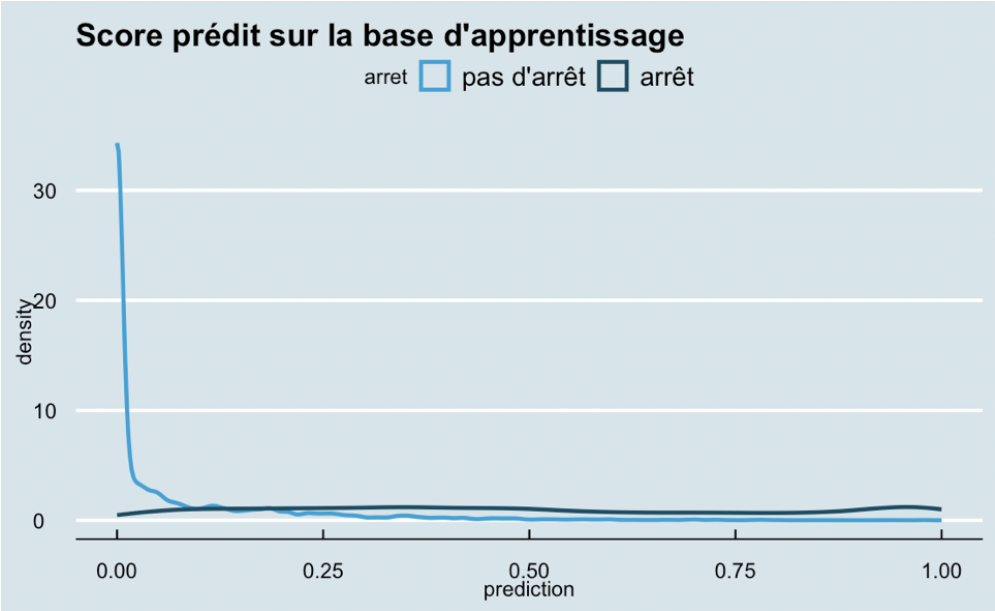


FIGURE 3.24 – Score prédit *Optimal Binning* : avec 12 mois de recul

En effet, les graphiques ci-dessous montrent que la coupure est faite avant le grand "pic" de concentration de valeurs de la classe "arrêt". Nous espérons ainsi mieux capter le signal de la classe positive.

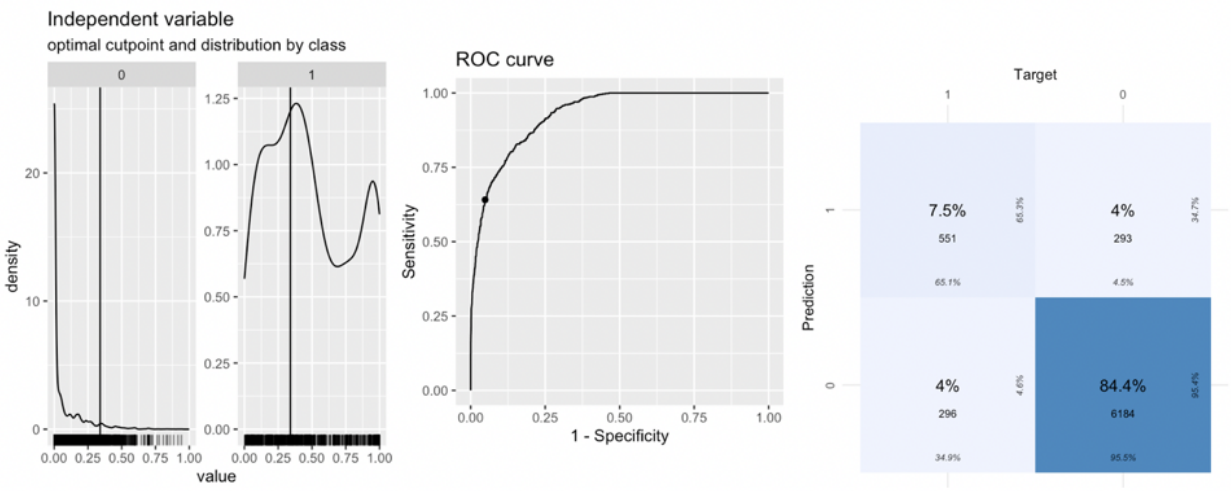


FIGURE 3.25 – Résultats *Optimal Binning* : avec 12 mois de recul

Les résultats obtenus sont satisfaisants : une amélioration du F1 Score qui est désormais de l'ordre de 0.652. Une meilleure sensibilité de 65% est obtenue au détriment d'une dégradation de la précision qui reste tout de même correcte. Ce modèle prédit mieux les arrêts et est plus prudent.

Nous décidons de retenir ce modèle pour la suite.

3.5.4 *Optimal Binning* : avec 3 mois de recul

Nous décidons de suivre la même démarche et de catégoriser les variables de consommation santé observées 3 mois avant l'arrêt. En effet, nous pensons que le lien devrait être plus fort lors de cette période. Si nous prenons l'exemple d'un arrêt suite à une hospitalisation, en réalité, le salarié est arrêté au lendemain de cette occurrence. Même si les temps de déclaration de l'arrêt et de gestion des données nous empêchent d'observer l'événement en temps réel, on pense qu'on devrait être en mesure de l'observer sous 3 mois. En effet, la durée moyenne entre la date de survenance et la date comptable d'observation du sinistre dans les bases est de 26 jours en Santé. En revanche, les arrêts de travail disposent de durée bien plus importante : en moyenne de 125 jours. De ce fait, il est intéressant d'établir un modèle sur les 3 mois afin d'être en mesure d'estimer le nombre d'arrêts futurs : les arrêts survenus mais non comptabilisés et les arrêts non survenus.

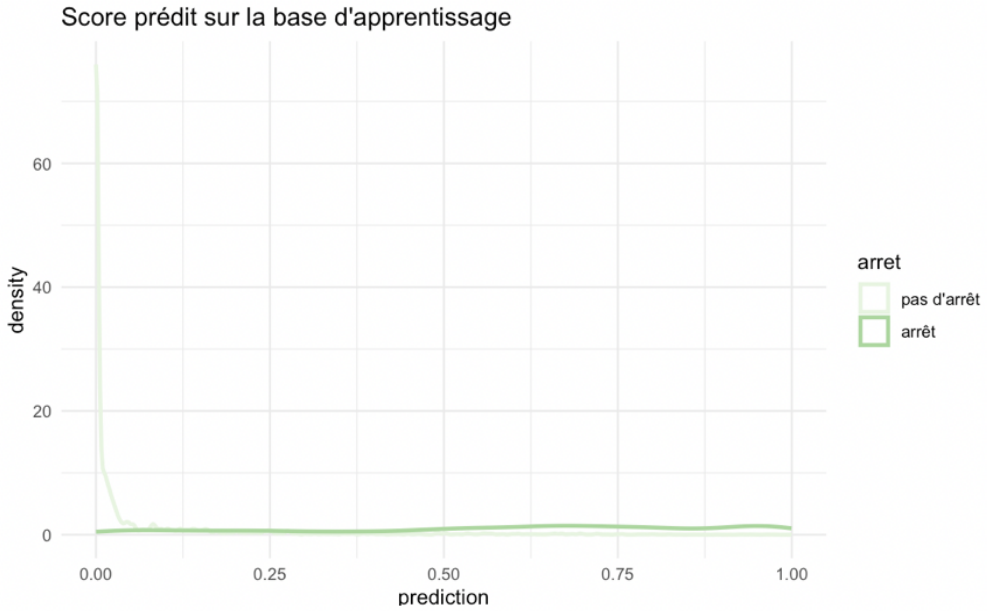


FIGURE 3.26 – Score prédit *Optimal Binning* : avec 3 mois de recul

Nous remarquons une différence flagrante de la distribution des modalités de la variable arrêt. Le grand "pic" de la modalité 1 s'est déplacé vers la droite du graphique, ce qui promet de meilleurs résultats en termes de prédiction.

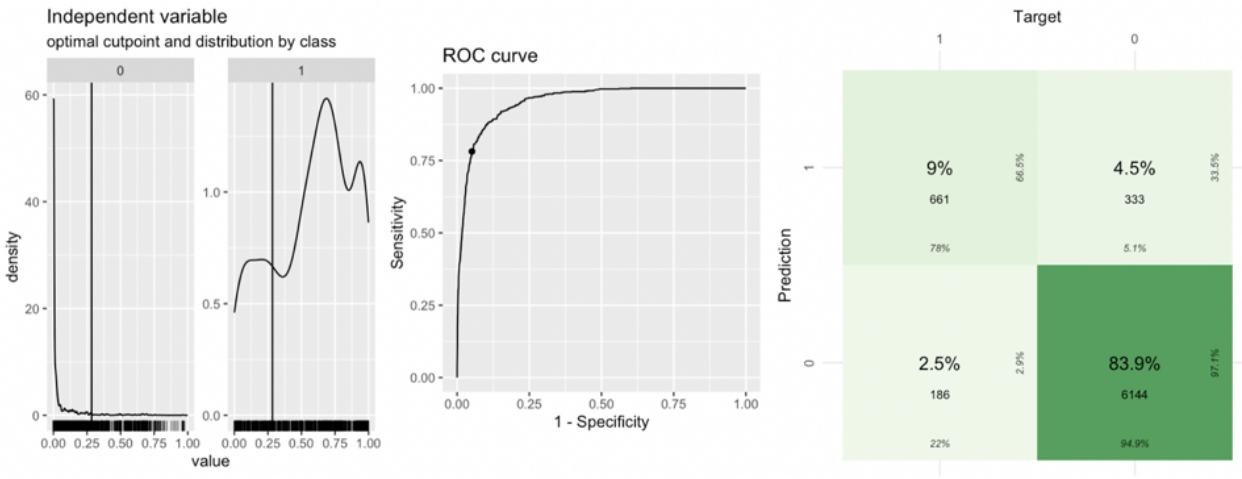


FIGURE 3.27 – Résultats *Optimal Binning* : avec 3 mois de recul

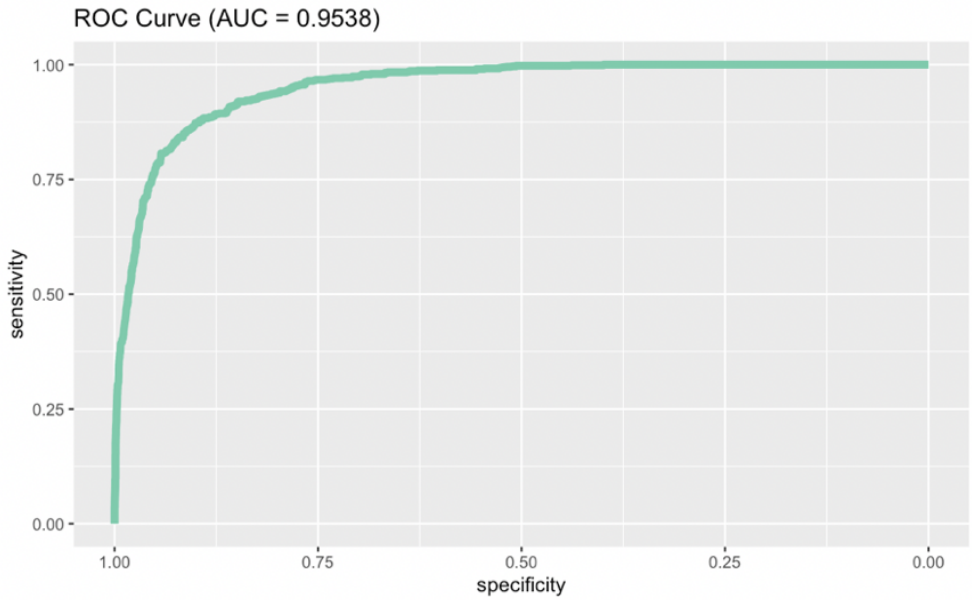


FIGURE 3.28 – Courbe ROC Optimal Binning

En effet, le pouvoir prédictif de ce modèle est le meilleur obtenu jusqu'à maintenant : 78% des arrêt ainsi que 95% des non-arrêts sont prédits par notre modèle. Le F1-Score s'est bien amélioré et est de l'ordre de 0.718. La précision de 66,5% reste correcte. Nous retenons ce modèle pour la suite.

3.6 Modélisation pénalisée

Comme expliqué précédemment, nous testons deux approches pour l'estimation de la probabilité d'entrée en arrêt de travail : un modèle GLM avec les interactions des variables dont la corrélation est supérieure à 0.6, et un modèle de *LASOO* avec l'intégralité des interactions.

3.6.1 Rappels théoriques

Un problème récurrent lors de régressions linéaires est l'instabilité des prédictions. Pour y remédier, la modélisation pénalisée réduit le nombre de variables jugées non discriminantes réduisant ainsi la variance de l'estimation. La régression pénalisée repose sur la fonction de vraisemblance précédemment énoncée avec une pénalisation qui se rajoute à cette dernière.

$$\max_{C(X)} \frac{1}{N} \sum_{i=1}^{n} [y_i \times C(X) - ln(1 + C(X))] - \lambda \left[\alpha ||\alpha||_1 + \frac{1}{2}(1 - \alpha)||\alpha||_2^2 \right] \tag{3.17}$$

Dans le package sur R "*glmnet*", deux paramètres contrôlent la pénalisation :

- λ : le coefficient de pénalité qui ajuste l'impact de la pénalisation. Plus λ est grand, plus les coefficients α tendent vers 0.
- a : Le paramètre de choix de méthode. Pour la régression *Lasso*, le paramètre a est égal à 1.

3.6.2 Utilisation de la validation croisée *K-folds*

Au vu du nombre important de variables corrélées entre elles, on se propose de tester une modélisation qui contient l'ensemble des interactions et une pénalisation de type Lasso. On comparera la performance des modèles par la suite avec les résultats obtenus précédemment.

Dans un premier temps, nous cherchons à déterminer le λ optimal. Pour ce faire, nous avons eu recours à la technique de validation croisée *K-folds*.

L'algorithme découpe le jeu de données en K parties, ou *folds*, sensiblement égales. À chaque itération, une de ces parties est utilisée comme base test, et les (k-1) restantes comme base d'entraînement. *In fine*, chaque partie aura servi 1 fois en validation et (k-1) fois en apprentissage.

Dans notre cas, une large séquence de λ est testée via la validation croisée. Deux paramètres doivent être choisis :

- K : (nombre de parties ou de *folds*) que l'on fixe à 10.
- Le type de métrique : comme expliqué précédemment, le MCO, méthode utilisée par défaut dans la fonction "cv.glmnet" , n'est pas adapté à notre modèle. Nous décidons de choisir le critère de la déviance.

La déviance mesure l'écart entre les valeurs attendues et les observations calculées à partir de la vraisemblance L. $D = -2log(L)$

3.6.3 Application : base avec 12 mois de recul

Nous partons de la base avec les variables catégorielles obtenues avec la méthode d'*Optimal Binning* et rajoutons l'intégralité des interactions possibles.

Tout d'abord, nous utilisons la fonction "cv.glmnet" de R afin de déterminer le meilleur λ.

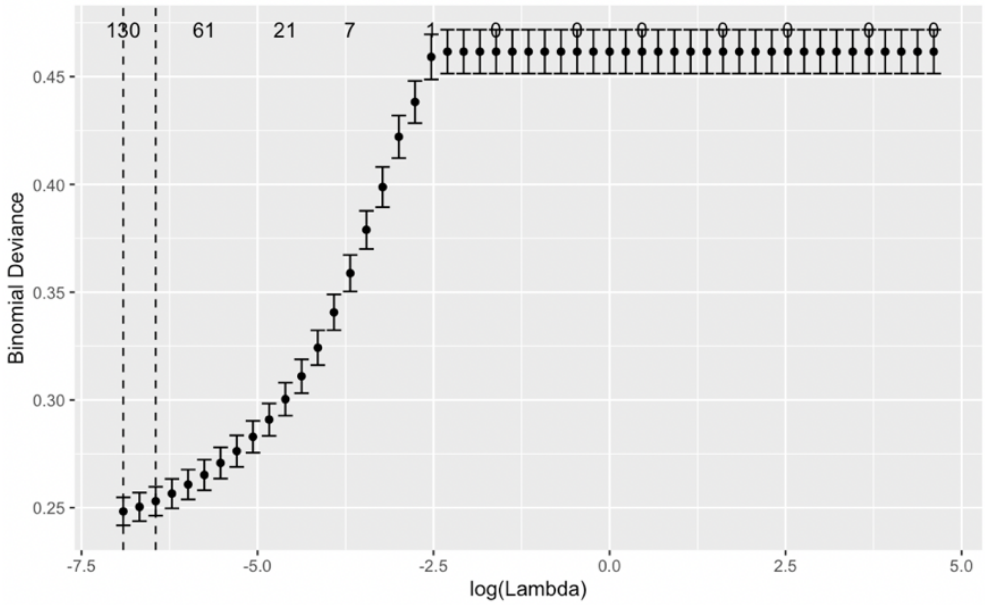


FIGURE 3.29 – Optimisation de lambda avec la fonction "cv.glmnet"

Nous obtenons un λ égal à 0.001, et l'utilisons dans notre régression pénalisée.

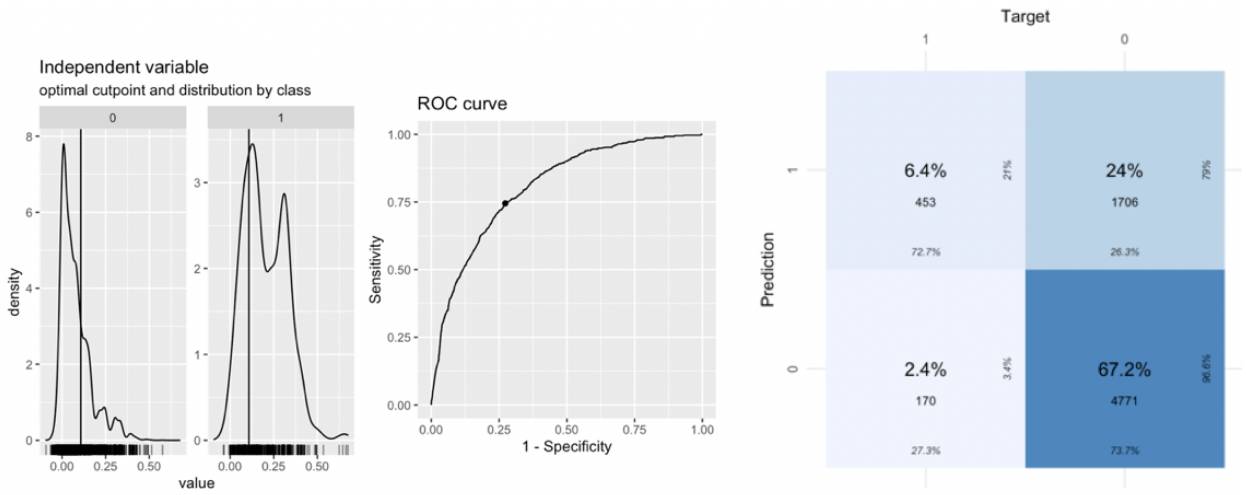


FIGURE 3.30 – Résultats après optimisation de lambda

On voit que la sensibilité s'est améliorée et atteint 72,7%. Toutefois, la précision ainsi que le F1-Score (33%) se sont nettement dégradés. Nous décidons d'opter pour le modèle GLM avec les données catégorielles via l'*Optimal Binning*.

3.7 Analyse

Nous nous intéressons à présent à l'analyse des résultats de notre modèle.

3.7.1 Modèle avec 12 mois de recul

Pour ce faire, nous utilisons la fonction "*summ()*" du package R "*jtools*" afin d'observer les coefficients de la régression logistique appliquée aux données santé observées sur un an.

	Est.	S.E.	z val.	p
(Intercept)	-0.25	0.13	-1.97	0.05
Age_début[25,30)	0.48	0.10	4.82	0.00
Age_début[30,35)	0.46	0.10	4.52	0.00
Age_début[35,40)	0.40	0.10	3.98	0.00
Age_début[40,45)	0.40	0.11	3.74	0.00
Age_début[45,50)	0.53	0.11	4.93	0.00
Age_début[50,60)	0.64	0.10	6.70	0.00
Age_début[60,68)	-0.08	0.18	-0.42	0.67
GENREM	-0.48	0.05	-9.46	0.00
collegeEDP	-0.26	0.09	-2.84	0.00
collegeNONCADRES	0.15	0.10	1.51	0.13
Sejour_Hospi_AVT_12_cat[1,5)	-18.63	236.45	-0.08	0.94
Sejour_Hospi_AVT_12_cat[5,20)	1.86	0.08	24.72	0.00
Sejour_Hospi_AVT_12_cat[20,40)	5.12	0.15	34.96	0.00
Sejour_Hospi_AVT_12_cat[40,100)	7.40	0.31	23.54	0.00
Sejour_Hospi_AVT_12_cat[100 et plus]	20.55	297.25	0.07	0.94
Optique_chirurgie_AVT_12_cat	-1.03	0.53	-1.95	0.05
Chambre_Part_AVT_12_cat	-1.17	0.12	-9.64	0.00
Maternite_AVT_12_cat	-1.38	0.14	-9.56	0.00
Forfait_Jour_Psy_AVT_12_cat	-0.50	0.47	-1.06	0.09
Forfait_Jour_Hospi_AVT_12_cat	-0.16	0.11	-1.51	0.05
SOINS_AUTRES_AVT_12_cat	-0.74	0.07	-11.25	0.00
Dispositifs_med_AVT_12_cat	-1.09	0.08	-13.69	0.00
Dentaire_AVT_12_cat	-1.39	0.06	-23.41	0.00
MEDECINE_DOUCE_AVT_12_cat	-0.91	0.06	-15.15	0.00
Consult_Gen_AVT_12_cat[1,17)	0.50	0.08	6.35	0.00
Consult_Gen_AVT_12_cat[17,21)	-0.93	0.19	-5.02	0.00
Consult_Gen_AVT_12_cat[21 et plus]	-2.71	0.19	-14.21	0.00
Consult_Spe_AVT_12_cat[1,2)	-28.22	421.93	-0.07	0.95
Consult_Spe_AVT_12_cat[2,4)	0.61	0.08	7.86	0.00
Consult_Spe_AVT_12_cat[4 et plus]	0.63	0.08	7.75	0.00
PHARMACIE_AVT_12_cat[6 et plus]	-0.32	0.07	-4.90	0.00

FIGURE 3.31 – Coefficients du modèle avec 12 mois de recul

Hospitalisation

Nous remarquons que les coefficients associés aux tranches supérieures à 5 actes de la variable Séjour_Hospitalier sont croissants et ont des *p-values* inférieures à 5% assurant leurs significativités. Cela se traduit par le fait que plus les actes de séjour hospitalier se multiplient plus la probabilité d'entrée en AT de l'individu augmente. Intuitivement, plus on observe d'actes, plus longue est la durée du séjour, ainsi plus l'assuré a des chances d'être en arrêt. Le coefficient de Chambre_Particulière est négatif et contre-intuitif selon nous. Nous rappelons que les actes sont moins fréquents sur ce poste, ce qui explique la non-interprétabilité du coefficient associé.

Maternité

Le coefficient associé à la variable "Maternité" montre une association négative avec la variable d'intérêt. Nous pensons que cela est principalement lié au déséquilibre de nos données. Des assurés pouvant avoir des actes sur le poste maternité sans être en arrêt, le modèle capte mieux le signal des personnes non arrêtées car ces dernières sont bien plus nombreuses dans notre base.

Consultation

Les coefficients associés aux différents niveaux de la variable Consultation généraliste ne présentent pas de tendance monotone. Comme pour les actes de maternité, nous pensons que cela provient du déséquilibre de nos données. Toutefois, une tendance positive est observée sur les niveaux de la variable Consultation Spécialiste. Plus l'assuré a des actes de consultations chez un spécialiste plus sa probabilité de tomber en arrêt, estimée par notre modèle, augmente.

3.7.2 Modèle avec 3 mois de recul

À présent, nous observons les coefficients de la régression logistique appliquée aux données santé observées sur trois mois.

	Est.	S.E.	z val.	p
(Intercept)	0.50	0.13	3.92	0.00
Age_début[28,40)	0.48	0.08	6.25	0.00
Age_début[40,50)	0.44	0.08	5.22	0.00
Age_début[50,65)	0.60	0.08	7.14	0.00
Age_début[65,68)	-2.82	0.80	-3.52	0.00
GENREM	-0.70	0.06	-12.51	0.00
collegeEDP	-0.27	0.11	-2.58	0.01
collegeNONCADRES	0.15	0.11	1.40	0.16
Sejour_Hospi_AVT_3_cat[1,2)	-17.28	218.01	-0.08	0.94
Sejour_Hospi_AVT_3_cat[2 et plus]	1.97	0.08	23.90	0.00
Consult_Gen_AVT_3_cat[1,2)	2.03	0.09	21.67	0.00
Consult_Gen_AVT_3_cat[2 et plus]	2.82	0.10	28.04	0.00
Consult_Spe_AVT_3_cat[1,2)	-17.78	238.34	-0.07	0.94
Consult_Spe_AVT_3_cat[2 et plus]	1.90	0.08	22.59	0.00
PHARMACIE_AVT_3_cat[2,4)	-0.43	0.10	-4.31	0.00
PHARMACIE_AVT_3_cat[4 et plus]	0.19	0.08	2.40	0.02
Maternite_AVT_3_cat	-2.48	0.24	-10.33	0.00
Forfait_Jour_Hospi_AVT_3_cat	-1.20	0.12	-10.21	0.00
SOINS_AUTRES_AVT_3_cat	-1.86	0.07	-26.45	0.00
Dispositifs_med_AVT_3_cat	-2.40	0.08	-30.21	0.00
Dentaire_AVT_3_cat	-2.50	0.08	-31.89	0.00
MEDECINE_DOUCE_AVT_3_cat	-1.92	0.08	-24.43	0.00

FIGURE 3.32 – Coefficients du modèle avec 3 mois de recul

Dans un premier temps, nous notons que l'*Optimal Binning* a catégorisé une grande partie des variables de ce modèle en 2 catégories. Ce découpage étant plus simple, les catégories regrouperont plus d'individus et pourront être plus faciles à interpréter.

Hospitalisation

Nous remarquons que le coefficient associé à la première tranche de la variable Séjour_Hospitalier n'est pas significatif, donc non-interprétable. Néanmoins, le coefficient de la tranche supérieure à 2 actes de la variable Séjour_Hospitalier est significativement positif. Cela se traduit par le fait que les actes de séjour hospitalier augmentent la probabilité d'entrée en AT. La variable Chambre_part a été supprimée du modèle pour sa non-significativité. La variable Forfait Journalier, n'ayant pas été catégorisée par notre modèle, est difficile à interpréter pour les mêmes raisons que celles évoquées dans la première partie de cette analyse.

Maternité

Le coefficient associé à la variable "Maternité" montre toujours une association négative avec la variable d'intérêt. Comme expliqué précédemment, nous pensons que cela est dû au déséquilibre constaté de nos données.

Consultation

Les coefficients associés aux deux niveaux de la variable Consultation généraliste sont positifs et croissants. La seconde tranche de la variable Consultation Spécialiste est également significativement positive. Un arrêt devant être prescrit par un médecin : généraliste ou spécialiste, l'observation sous 3 mois facilite l'interprétation.

3.8 Conclusion

L'objectif de ce chapitre est l'exploitation des données santé pour prédire un arrêt. Nous avons comparé la consommation médicale des personnes ayant eu au moins un arrêt durant la période d'observation 2017-2019 à ceux qui n'en ont pas eu. Nous nous sommes intéressés aux 12, 6 et 3 mois précédant l'arrêt. Pour ce faire, nous avons opté pour une modélisation de type GLM.

Le GLM estime la probabilité d'entrée en arrêt de chaque individu – donnée intéressante à exploiter. Toutefois, un travail d'analyse et de recherche a été réalisé afin de trouver le seuil optimal à partir duquel l'individu est considéré comme en arrêt. En pratique, ce seuil peut être ajusté en fonction du niveau de prudence souhaité : plus on augmente ce seuil, moindre est la prudence.

Nous avons utilisé plusieurs méthodes afin d'améliorer le pouvoir prédictif de nos modèles : *Optimal Binning*, *LASSO*... et obtenons *in fine* deux modèles satisfaisants : un utilisant les données sur l'année précédant l'arrêt et un deuxième utilisant uniquement les données sur les 3 mois qui précèdent l'arrêt.

Par ailleurs, un programme a été réalisé sur SAS permettant la sélection des individus (double-équipés, assurés uniquement et pas bénéficiaires), la collecte des informations des différentes bases, la sélection des variables ainsi que leurs catégorisations pour la répartition adoptée lors de cette étude. La base obtenue est à rentrer dans le programme R afin d'obtenir les indicateurs souhaités.

Le programme fournit un fichier Excel avec la probabilité d'entrée en arrêt de travail par assuré et surtout une moyenne pour l'ensemble de la base fournie. Il convient à l'utilisateur de sélectionner le modèle souhaité : avec un recul de 12 ou 3 mois.

Ces deux modèles peuvent notamment être utilisés pour le pilotage technique. Cet indicateur est intéressant à utiliser pour affiner la surveillance du portefeuille double-équipé. Le caractère prospectif de l'étude est intéressant : elle sera en mesure de prévenir ou d'alerter sur une hausse de l'absentéisme. Il peut également être pertinent de l'utiliser pour les ajustements tarifaires et en guise de prévention dans le but d'atténuer l'effet d'une éventuelle hausse de la sinistralité en prévoyance.

Une seconde piste envisageable est celle du volet souscription. Le portefeuille étant en majorité composé de contrats santé, le développement du portefeuille prévoyance, et notamment en "multi-équipement", fait partie des objectifs du groupe. L'utilisation de cette étude est notamment pertinente lors de la proposition d'une couverture prévoyance à une entreprise faisant partie du portefeuille santé du Crédit Agricole Assurances. En utilisant ce programme, il est possible d'avoir un indicateur supplémentaire sur le risque arrêt de travail de l'entreprise en question.

En conclusion de cette partie, les différents modèles nous mènent à conclure qu'il y a bien un lien entre les données Santé et la probabilité d'être en arrêt de travail. Une deuxième composante indispensable dans notre étude est la notion de durée. En effet, on peut penser que les données santé peuvent nous aider à prédire la durée de l'arrêt. Par exemple, un arrêt lié à une grippe peut durer moins longtemps qu'un arrêt suite à une hospitalisation pour une chirurgie.

Chapitre 4

Modélisation de la durée d'un arrêt de travail

Sommaire

Nous nous intéressons à présent au lien existant entre la consommation santé et la durée de l'arrêt de travail. Intuitivement, on peut penser qu'une personne forte consommatrice en Santé n'étant donc pas en bonne santé, peut avoir des arrêts plus longs.

4.1 Contexte de l'étude

En assurances collectives, les tables réglementaires du Bureau Commun des Assurances Collectives ou communément appelées tables de BCAC régissent les calculs de provisionnement, et ce, pour le risque d'incapacité comme le risque d'invalidité. L'assureur peut utiliser les tables réglementaires du BCAC, ou peut établir ses tables d'expériences propres à son portefeuille. Dans le cas où il opterait pour la seconde option, ces tables doivent faire l'objet d'une certification par un actuaire agréé indépendant. L'élaboration de tables d'expérience requiert des données de qualité et en quantité suffisante.

Nous précisons que cette partie n'a pas pour vocation d'établir un modèle de provisionnement. En effet, nous avons rencontré plusieurs problèmes quant à la qualité, mais aussi la quantité de données. Cette partie a pour but de mettre en lumière l'impact de la consommation en santé sur l'arrêt de travail - si impact il y a. Idéalement et potentiellement avec une base plus complète, un modèle de durée établit avec des données de consommation santé, lissé ou même ajusté avec les données réglementaires des taux de BCAC constituent une piste intéressante à explorer. En effet,

les tables de BCAC reposent sur deux composantes, l'âge d'entrée en incapacité et l'ancienneté (en mois). L'intégration de données santé à la modélisation peut indiquer une composante, non utilisée jusqu'à présent et particulièrement intéressante à exploiter : le comportement de l'assuré vis-à-vis de sa santé. Prend-il soin de sa santé ? Va-t-il consulter lorsqu'il est malade ? Prend-il correctement son traitement ? Nous pensons que ces interrogations ont un impact direct sur la fréquence, mais aussi la durée des arrêts.

Avant de se lancer dans la modélisation, il est important de présenter des notions auxquelles nous avons affaire dans cette partie.

4.1.1 Manque de données

L'enjeu dans cette partie est la quantité de données dont on dispose afin d'établir une éventuelle modélisation. Comme indiqué précédemment, nous nous intéressons aux assurés ayant souscrit à une assurance santé, mais également à une assurance prévoyance et plus précisément à une couverture incapacité. On rajoute également une condition de présence de minimum 12 mois consécutifs, afin d'avoir une période d'observation suffisamment longue. Ces conditions réduisent considérablement le périmètre de l'étude, d'autant plus que le Crédit Agricole Assurances s'est lancé sur le marché des assurances collectives en 2015, le portefeuille est en constante évolution et le portefeuille santé est plus développé que celui de la prévoyance. Toutefois, en s'intéressant aux années de notre étude, à savoir de 2017 à 2019, et en sélectionnant les individus qui répondent aux différents critères énoncés précédemment, notre base est limitée à 2 117 arrêts.

4.2 Censure et troncature

Par ailleurs, il est commun de rencontrer des problèmes de données incomplètes notamment lorsqu'on parle d'estimation de durée. On distingue deux phénomènes distincts : la censure et la troncature.

4.2.1 Censure

Lorsque l'information sur la date d'occurrence d'un événement est incomplète, par exemple la date de survenance de l'arrêt, on parle de censure.

La censure est dite à gauche lorsque l'événement est survenu avant la période d'observation. De la même manière, la censure est dite à droite lorsque l'événement est survenu après la fin de la période d'observation.

Nos données ne disposent pas de censure à gauche liée à la franchise. Cette dernière est souvent source de censure de données, et est importante à considérer. Dans notre étude, les données ne présentent pas de censure liée à la franchise, car nous disposons des sinistres indemnisés et non indemnisés.

Dans notre cas, et même si elles n'appartiennent pas à l'intervalle d'étude, les dates de début et

de fin de l'arrêt sont renseignées. Dans ce sens, nos données ne sont pas censurées. Toutefois, et afin d'éviter l'introduction d'un biais, nous considérons que tous les sinistres encore en cours au 31/12/2019 sont des sinistres censurés.

4.2.2 Troncature

Lorsque l'information sur la date d'occurrence d'un évènement n'est pas rapportée dans la base de données, bien qu'il ait eu lieu, on parle de troncature.

La troncature est dite à gauche lorsque l'évènement n'est pas rapporté en dessous d'une certaine intensité. De la même manière, la troncature est dite à droite lorsque l'évènement n'est pas rapporté au dessus d'une certaine intensité.

Nous n'avons pas de problème de troncature dans notre étude.

4.3 Environnement et base de données

4.3.1 Construction de la base

Nous partons de la même base présentée précédemment qui regroupe l'ensemble des salariés "double équipés", à savoir les assurés en Santé et en Prévoyance avec la garantie incapacité. Nous gardons uniquement les personnes ayant eu au moins un arrêt durant la période.

Dans le cas où l'assuré présenterait plusieurs arrêts à la suite, nous sélectionnons la date du premier arrêt comme date de début et la date de fin du dernier arrêt comme date de fin dans notre base. Nous rappelons que l'unique censure, dont il est sujet dans cette étude, est celle liée à la période d'observation. En effet, si l'arrêt est toujours en cours à la fin de la période d'observation, la date de fin de l'arrêt dans notre base sera le 31/12/2019. Dans ce cas, nous ne disposons pas de la durée "réelle". Ainsi, nous disposons des données de base :

- Le sexe ;
- La catégorie socioprofessionnelle ;
- La date de naissance ;
- La date de survenance ;
- La date de fin de l'arrêt ;
- Le secteur d'activité ;

À cela, s'ajoutent les données de consommation santé présentées dans la première partie de ce mémoire. Dans cette partie, nous nous limitons aux actes de santé effectués durant les 12 mois précédant l'arrêt.

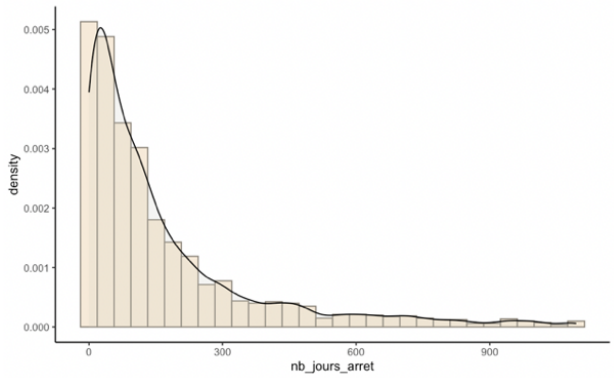
4.3.2 Statistiques descriptives

Dans un premier temps, nous présenterons des statistiques descriptives sur les principales variables.

Durée de l'arrêt de travail

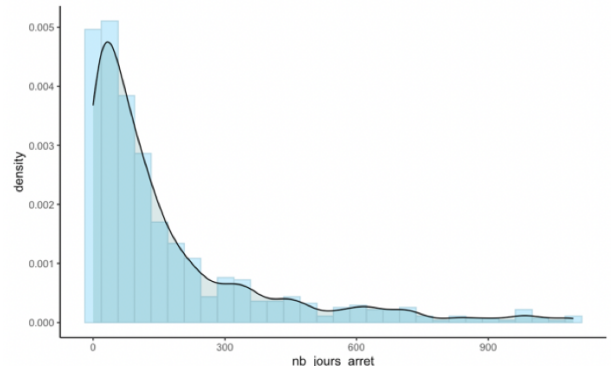
En France, la durée maximale d'un arrêt de travail dépend du motif de l'arrêt et de la catégorie socioprofessionnelle de l'assuré. Il convient de noter que la durée maximale d'un arrêt de travail est de 3 ans ; soit 36 mois. Au-delà de 3 ans, l'individu est considéré en incapacité permanente et passe ainsi en invalidité. Nous vérifions qu'aucun arrêt ne dépasse les 36 mois dans notre base de données. Le cas contraire, cela constituerait des "valeurs aberrantes" à traiter.

En nous intéressant à la répartition de notre variable d'intérêt, nous remarquons une répartition non-uniforme des durées des arrêts. La majorité des arrêts sont de moins d'un an. Toutefois, les arrêts de "courte durée", de moins de 10 jours, sont très peu représentés. Ce phénomène est essentiellement lié au périmètre de l'étude : le portefeuille double équipé ; et ne représente pas les répartitions observées sur l'ensemble du portefeuille, ou encore celles observées sur le marché. Notre base de données est composée de 2 117 arrêts répartis comme suit :



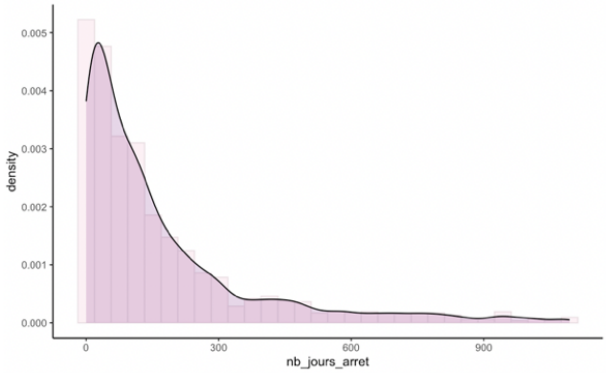
	2017	2018	2019	Global
[0 - 5]	2,2%	2,3%	5,0%	**9,5%**
]5 - 10]	0,9%	0,9%	2,0%	**3,7%**
]10 - 30]	2,5%	3,5%	5,9%	**11,9%**
]30 - 90]	6,9%	6,5%	10,6%	**24,0%**
]90 - 365]	10,2%	11,4%	15,9%	**37,4%**
]365 -]	6,0%	7,4%	0,0%	**13,4%**
Global	**28,6%**	**32,0%**	**39,4%**	**100,0%**

FIGURE 4.1 – Durée de l'arrêt de travail sur l'ensemble de la base



	2017	2018	2019	Global
[0 - 5]	3,0%	2,0%	4,8%	**9,8%**
]5 - 10]	0,5%	1,0%	2,2%	**3,7%**
]10 - 30]	2,2%	4,2%	4,8%	**11,2%**
]30 - 90]	7,9%	6,1%	12,3%	**26,4%**
]90 - 365]	8,5%	11,1%	16,1%	**35,7%**
]365 -]	4,6%	8,6%	0,0%	**13,3%**
Global	**26,8%**	**33,1%**	**40,2%**	**100,0%**

FIGURE 4.2 – Durée de l'arrêt de travail homme



	2017	2018	2019	Global
[0 - 5]	1,7%	2,5%	5,1%	**9,3%**
]5 - 10]	1,0%	0,8%	1,9%	**3,8%**
]10 - 30]	2,6%	3,2%	6,5%	**12,3%**
]30 - 90]	6,4%	6,7%	9,7%	**22,8%**
]90 - 365]	11,0%	11,6%	15,7%	**38,3%**
]365 -]	6,8%	6,7%	0,0%	**13,5%**
Global	**29,5%**	**31,4%**	**39,1%**	**100,0%**

FIGURE 4.3 – Durée de l'arrêt de travail femme

Nous remarquons que l'année 2019 présente beaucoup d'arrêts. Le fait qu'aucun arrêt ne dépasse les 365 jours s'explique par la censure au 31/12/2019. En comparant la durée des arrêts de travail chez les hommes et chez les femmes, on remarque une répartition assez similaire.

Sexe

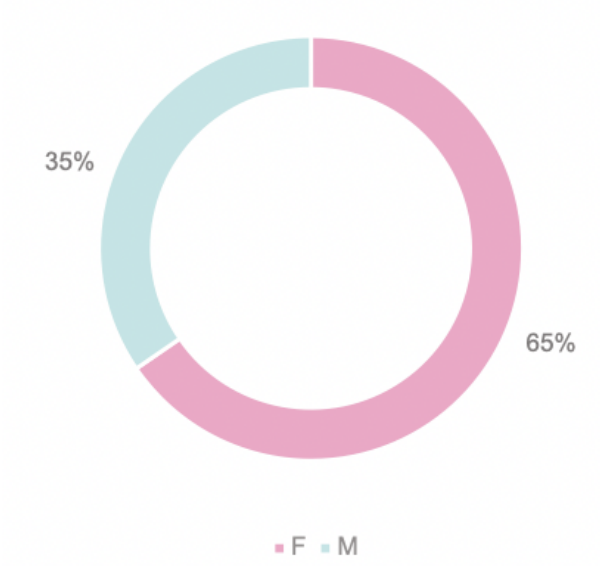


FIGURE 4.4 – Répartition des arrêts en fonction du sexe

Sans surprise, on observe dans notre base de données une majorité de femmes. En effet, on avait noté un taux d'entrée en AT plus élevé chez les femmes dans la première partie de ce mémoire.

Âge à la survenance

En observant l'âge d'entrée en arrêt de travail, on constate que le problème rencontré dans la première partie avec l'âge de début d'observation ne se pose pas dans cette partie. En utilisant l'information sur la date de naissance disponible dans la table de sinistres prévoyance, nous sommes en mesure de calculer l'âge à la survenance sans valeurs aberrantes. En effet, l'âge maximal étant inférieur à 67 ans ; nous n'avons pas le besoin de retraiter cette variable. Comme expliqué dans la

première partie, s'agissant de contrats collectifs couvrant ainsi des salariés, la majorité des assurés ont entre 18 et 60 ans.

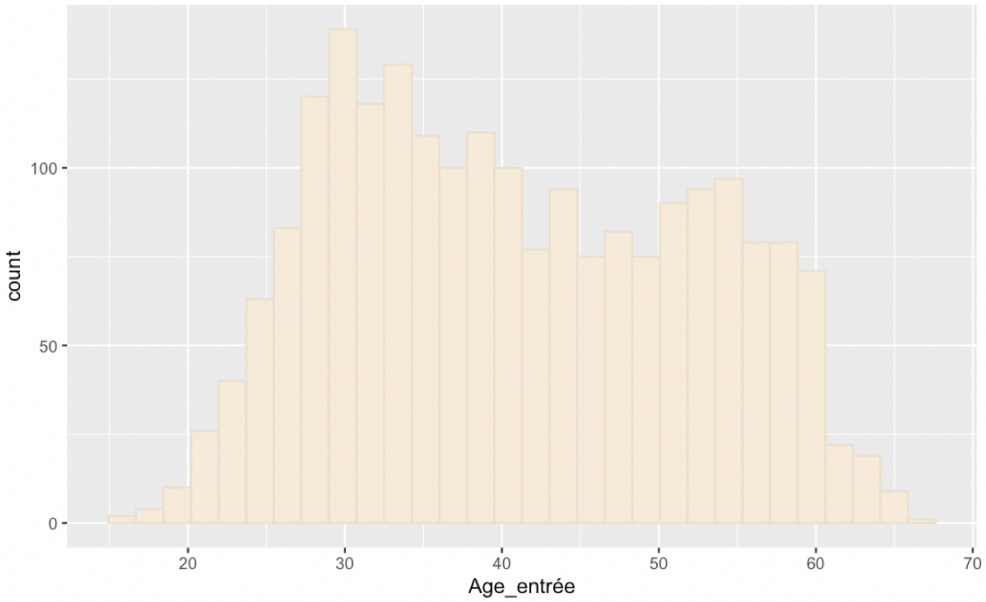


FIGURE 4.5 – Âge à la survenance sur l'ensemble de la base

FIGURE 4.6 – Âge à la survenance en distinguant homme et femme

On note toutefois une répartition différente des âges à la survenance entre les femmes et les hommes. Un pic d'arrêt est constaté entre les âges de 27 et 35 ans chez les femmes ; phénomène non observé chez les hommes. Nous supposons que ce dernier provient des arrêts liés à la maternité, généralement observée autour de ces âges-ci chez les femmes. Nous nous intéressons à présent à la durée moyenne d'un arrêt de travail par tranche d'âge.

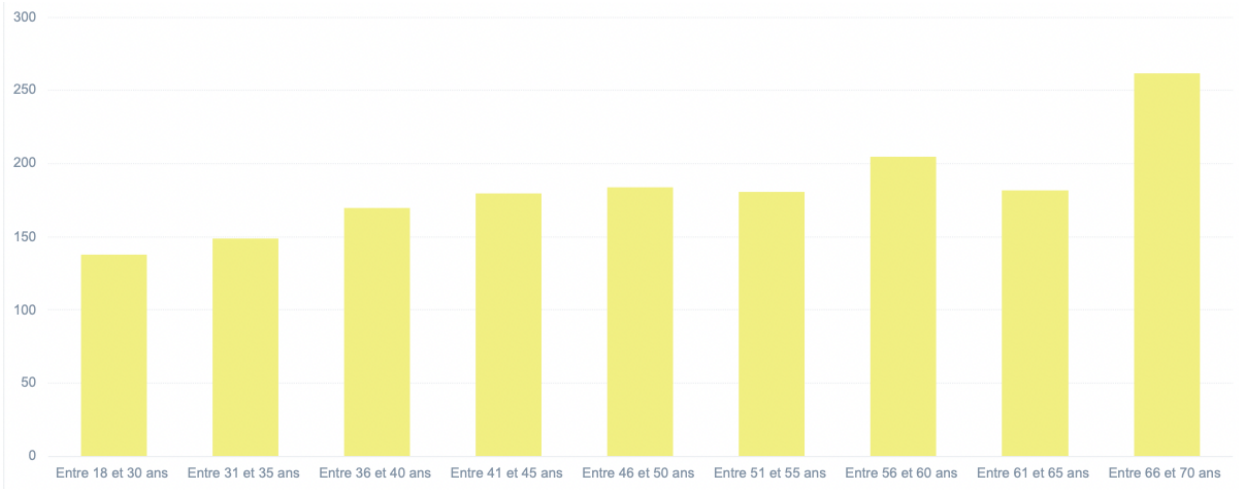


FIGURE 4.7 – Durée moyenne de l'arrêt selon l'âge

Nous remarquons que la durée moyenne d'un arrêt de travail est globalement croissante avec l'âge d'entrée. La légère baisse entre 61 et 65 ans semble être dû au faible nombre d'assurés sur cette tranche. Nous constatons ainsi un lien fort entre la durée d'un arrêt et l'âge d'entrée en arrêt. Plus l'âge de l'assuré est avancé plus la durée de l'arrêt de travail est longue. En effet, cela peut s'expliquer par la dégradation de la santé avec l'âge et donc un temps de traitement nécessaire plus important.

Collège

La majorité des arrêts ne possédant pas de catégorie socio-professionnelle renseignée dans notre base de données. Nous décidons de garder la variable pour le moment mais de la supprimer dans le cas où cette dernière ne s'avère pas significative dans nos modèles.

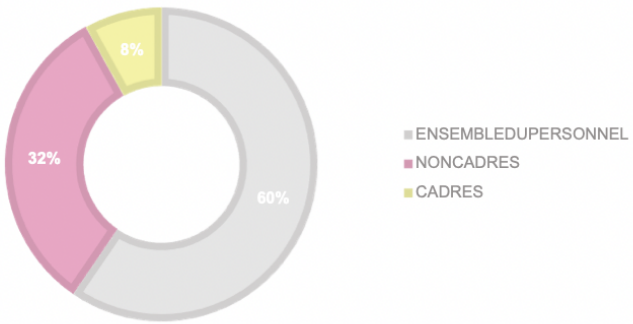


FIGURE 4.8 – Répartition des arrêts selon le collège

Motif de l'arrêt

Nous intégrons dans cette partie l'information sur le motif de l'arrêt. Nous constatons que 4% des arrêts ne disposent pas de motifs dans nos bases. Nous décidons de les assimiler à la catégorie la plus représentée dans notre base : la Maladie (qui, pour information, inclus les actes de maternité).

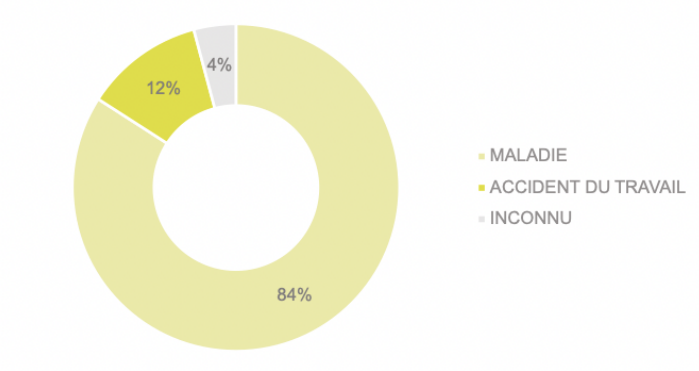


FIGURE 4.9 – Répartition des arrêts selon le motif

En nous intéressant aux motifs d'arrêt par sexe, on constate que les femmes sont moins sujettes aux accidents de travail ; 87% des arrêts étant dus à la maladie.

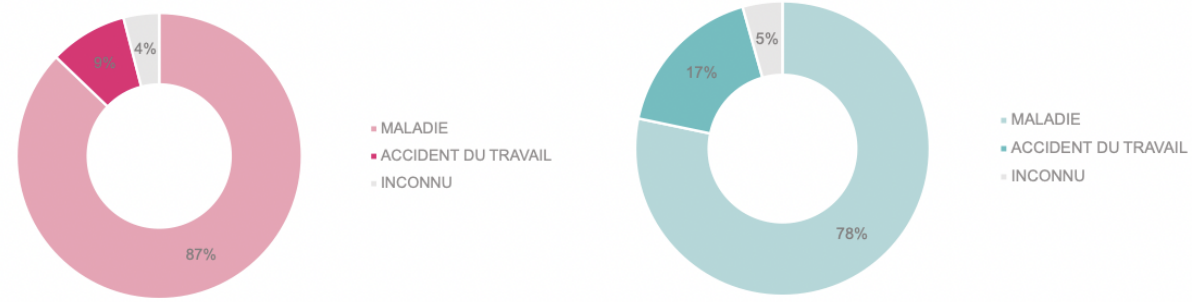


FIGURE 4.10 – Répartition des arrêts selon le motif en distinguant homme et femme

4.4 Modèle de Cox

Le but de cette partie et de modéliser la durée d'un arrêt de travail en utilisant les données de consommation en Santé. Nous optons pour les modèles de durée. Ces derniers modélisent la probabilité de survie d'un individu en fonction du temps. Dans notre étude, la "survie" correspond au temps passé en incapacité. Les modèles de survie peuvent être paramétriques, semi-paramétriques ou non paramétriques. Ce choix doit être effectué selon le type de données disponibles et les hypothèses prises sur la distribution de la durée d'arrêt.

4.4.1 Rappels théoriques

On note T une variable aléatoire à valeur dans $[0, +\infty[$. Sa fonction de distribution sera définie comme suit : $F(t) = P(T \leq \text{t})$ et sa fonction de survie : $S = 1 - F(t) = P(T > t)$

Sa fonction de survie conditionnelle :

$$S_u(t) = P(T > u + t \mid T > u) = \frac{P(T > u + t)}{P(T > t)} = \frac{S(u + t)}{S(u)} \tag{4.1}$$

La fonction de hasard est ainsi définie :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} ln(S(t)) \tag{4.2}$$

La fonction de hasard détermine entièrement la loi de T et on a ainsi la relation suivante :

$$S(t) = exp\left(\int_0^t h(s)ds\right)$$

$$\text{Avec} \begin{cases} S(t) = -exp(H(t)) \\ H(t) = \int_0^t h(s)ds \end{cases}$$

Le modèle de Cox est un modèle de survie semi-paramétrique qui est notamment utilisé pour estimer les effets des covariables, dans notre cas sur la durée d'arrêt. Il suppose que la relation entre la durée d'arrêt et les covariables peut être décrite par une fonction de risque proportionnel.

On se donne une fonction de survie de base, $B(t)$ et on suppose que la fonction de survie, présentée précédemment s'écrit sous la forme $S_\theta(t) = B(t)^\theta$, avec $\theta > 0$ un paramètre inconnu. Dans le modèle de Cox ce dernier s'écrit :

$$\theta = e^{z'\beta}$$

$$\text{Avec} \begin{cases} z = (z_1, z_2, ..., z_p) \text{ un vecteur de p variables} \\ \beta = (\beta_1, \beta_2, ..., \beta_p) \text{ le vecteur de paramètres} \end{cases}$$

La densité s'exprime comme $f_\theta(t) = \theta B(t)^{\theta-1} f(t)$, la fonction de hasard prend donc la forme :

$$h_\theta(t) = \frac{f_\theta(t)}{S_\theta(t)} = \theta \frac{f(t)}{B(t) = \theta h(t)} h_\theta(t) = \frac{f_\theta(t)}{S_\theta(t)} = \theta \frac{f(t)}{B(t) = \theta} \tag{4.3}$$

Ainsi :

$$ln(h(t \mid Z = z)) = ln(h(t)) + \sum_{i=1}^{p} z_i \beta$$

4.4.2 Tests de validation du modèle

Afin de valider le modèle obtenu, différents tests statistiques sont envisageables.

Test de significativité des coefficients

Le test de Wald permet de tester si les coefficients estimés des variables explicatives sont significativement différents de zéro. L'hypothèse nulle est donc que le coefficient est égal à zéro contre

l'hypothèse alternative selon laquelle le coefficient est différent de zéro.

$$\text{Test de Wald : } \begin{cases} \text{H0} : \beta_j = 0 \\ \text{H1} : \beta_j \neq 0 \end{cases}$$

La statistique de Wald est utilisée :

$$W_j = \frac{\sqrt{n}\beta_j}{\sqrt{Var(\beta_j)}}$$

Sous H0, la statistique de Wald suit une loi normale $N(0, 1)$.

Une *p-value* associée inférieure à un certain niveau de signification (fixé à 5%), indique que le coefficient est significativement différent de zéro.

Test de significativité globale des coefficients

Le test de Wald pour la nullité globale des coefficients est utilisé pour évaluer si les covariables ont un effet global significatif sur le risque étudié, dans notre cas la durée de l'arrêt. L'hypothèse nulle est définie par :

$$\text{H0} : \beta_1 = \beta_2 = ... = \beta_j = 0$$

L'hypothèse nulle est que tous les coefficients de régression sont égaux à zéro. L'hypothèse alternative est qu'au moins un des coefficients de régression est différent de zéro.

Test de rapport de vraisemblance

Le *likelihood ratio test* ou test de rapport de vraisemblance compare la vraisemblance du modèle ajusté à la vraisemblance du modèle le plus simple avec uniquement l'intercepte. La statistique du test est fondée sur la différence entre les logarithmes des fonctions de vraisemblance du modèle complet et du modèle réduit.

$$LR = -2log\frac{L_{réduit}}{L_{complet}}$$

$$\text{Avec} \begin{cases} L_{réduit} \text{ la vraisemblance du modèle réduit} \\ L_{complet} \text{ la vraisemblance du modèle réduit} \end{cases}$$

Ainsi, une valeur élevée indique que le modèle complet est significativement meilleur que le modèle réduit. Une faible *p-value* valide que le modèle ajusté est significativement meilleur.

Test de concordance

Le test de concordance évalue le pouvoir de prédiction du modèle. Il mesure la concordance entre les temps de survie prédits par le modèle et les temps de survie réels observés dans les données. Il

mesure la proportion de paires d'individus dans lesquelles le modèle prédit correctement le temps d'événement de la paire.

$$C = \frac{nombredepairesconcordantes - nombredepairesdiscordantes}{nombretotaldepaires}$$

C varie entre 0.5 et 1. Une valeur de concordance élevée, proche de 1 est signe d'une bonne capacité de prédiction. Ainsi, une valeur de 1 indique une concordance parfaite entre les prévisions du modèle et les observations réelles. Un test de concordance significatif indique que le modèle a un pouvoir prédictif supérieur à celui d'un modèle aléatoire.

Test de Score *log-rank*

Le score *log-rank* test permet de comparer la survie entre deux groupes distincts et notamment d'évaluer si les différences observées entre les deux groupes sont significatives. Ce test suppose des hypothèses relativement fortes : les risques doivent être proportionnels et la fonction de survie doit suivre une loi exponentielle.

L'hypothèse nulle est que les deux groupes possèdent des fonctions de survie identiques.

Une *p-value* inférieure à un certain niveau de signification indique que les différences observées entre les deux groupes sont statistiquement significatives ; il y a une différence significative dans la durée des arrêts entre les deux groupes.

Test de proportionnalité

Le test de proportionnalité se base sur la fonction de score de Cox. Son objectif est de tester si le coefficient associé à chaque variable explicative du modèle est constant au fil du temps.

Une des méthodes des plus utilisées est le test de Schoenfeld, qui, comme son nom l'indique, est basé sur la corrélation entre les résidus de Schoenfeld et le temps. Les résidus de Schoenfeld correspondent à la différence entre les valeurs observées et prédites des variables explicatives. Dans le cas où la corrélation est significativement différente de zéro, une violation de l'hypothèse de proportionnalité des risques est donc constatée.

$$\text{Test de proportionnalité :} \begin{cases} \text{H0 : Les risques sont proportionnels pour toutes les covariables} \\ \text{dans le modèle de Cox.} \\ \text{H1 : Les risques ne sont pas proportionnels pour au moins une} \\ \text{covariable dans le modèle de Cox.} \end{cases}$$

Contrairement aux autres tests présentés précédemment, le test de proportionnalité n'est pas directement obtenu à la sortie du modèle de Cox. Il est réalisé en utilisant la fonction *cox.zph()* dans le logiciel R.

4.4.3 Application : Modélisation de base

Une fois le modèle expliqué et les différents tests de validation expliqués, nous passons à présent à l'application. Comme indiqué précédemment, nous disposons de 2 117 individus. En effet, nous décidons de nous restreindre au périmètre "double équipé" ainsi que sur les 3 années d'observation. Au vu du nombre très limité d'arrêts observés et afin d'éviter les éventuels biais d'estimation liés à la taille de notre échantillon, l'enjeu est de réduire la segmentation des individus en gardant les variables et catégories de variables significatives et essentielles pour l'étude.

Nous traitons les variables numériques de consommation Santé en les catégorisant en 2 niveaux avec une répartition quasi-uniforme. Nous tentons une première modélisation en gardant toutes les variables catégorisées en 2 segments. Cette modélisation dite "de base" a pour vocation de mettre en évidence les éventuels besoins de regroupement et de retraitement des variables.

Pour ce faire, nous utilisons la fonction *coxph()* du package *survival*. Cette dernière prend deux variables. Une première, temporelle, indique la durée à laquelle survient l'évènement étudié. Dans notre étude, il s'agit de la durée de l'arrêt de travail : la différence (en jours) entre l'ancienneté à la sortie de l'état d'incapacité et l'ancienneté à l'entrée de ce dernier. Pour ceux n'ayant pas vécu l'évènement (potentielle censure à droite) la variable temporelle n'est autre que la différence entre la date de fin d'observation et la date de survenance de l'arrêt. Par ailleurs, une seconde variable indique si les individus ont vécu l'évènement de censure (0 pour non, 1 pour oui) ; il s'agit de la variable "non_censure" de notre modèle.

La sortie du modèle regroupe l'ensemble des estimateurs des coefficients des covariables ainsi que les intervalles de confiance, les *p-values* correspondantes pour chaque coefficient, mais aussi les résultats de trois tests statistiques : le test de rapport de vraisemblance et la *p-value* associée, un test de Wald, un score *log-rank* et un test de concordance.

Les coefficients indiquent la force, mais aussi la direction de l'association entre chaque covariable et la variable cible, à savoir la durée de l'arrêt. Une valeur de coefficient supérieure à 0 indique un risque accru de prolongation de l'arrêt, tandis qu'une valeur inférieure à 1 indique un risque réduit. Nous observons dans la sortie ci-dessous que très peu de variables sont significatives :

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
GENREM	0.098665	1.103696	0.076170	1.295	0.19521	
collegeENSEMBLEDUPERSONNEL	-0.215112	0.806451	0.147262	-1.461	0.14409	
collegeNONCADRES	-0.447643	0.639133	0.151125	-2.962	0.00306	**
AgeSurvenance	0.002552	1.002556	0.003197	0.798	0.42465	
Chambre_Part_AVT_12	0.009023	1.009063	0.019136	0.472	0.63728	
Sejour_Hospi_AVT_12	-0.003149	0.996856	0.002724	-1.156	0.24770	
Forfait_Jour_Hospi_AVT_12	-0.006619	0.993403	0.011746	-0.564	0.57309	
Secteur_catService	-0.313180	0.731119	0.178193	-1.758	0.07883	.
Mot_ATMALADIE	-0.308912	0.734246	0.110195	-2.803	0.00506	**
Consult_Gen_AVT_12_cat[3,190)	0.037981	1.038711	0.107020	0.355	0.72267	
Consult_Spe_AVT_12_cat[2,182)	0.148885	1.160540	0.085555	1.740	0.08182	.
SOINS_AUTRES_AVT_12_cat[4,231)	0.066991	1.069286	0.091895	0.729	0.46600	
Dispositifs_med_AVT_12_cat[4,113)	0.113612	1.120318	0.114590	0.991	0.32146	
PHARMACIE_AVT_12_cat[16,300)	0.179460	1.196571	0.091875	1.953	0.05078	.
Optique_chirurgie_AVT_12_cat	0.492302	1.636078	1.012151	0.486	0.62669	
Maternite_AVT_12_cat	-0.106281	0.899172	0.252283	-0.421	0.67355	
Forfait_Jour_Psy_AVT_12_cat	0.091165	1.095450	0.374916	0.243	0.80788	
Cure_AVT_12_cat	-0.016215	0.983916	0.458803	-0.035	0.97181	
Dentaire_AVT_12_cat	0.036145	1.036806	0.079083	0.457	0.64763	
MEDECINE_DOUCE_AVT_12_cat	0.160395	1.173975	0.082054	1.955	0.05061	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Concordance= 0.59 (se = 0.013)
Likelihood ratio test= 66.41 on 20 df, p=7e-07
Wald test = 67.51 on 20 df, p=5e-07
Score (logrank) test = 68.14 on 20 df, p=4e-07

FIGURE 4.11 – Coefficients du modèle de base

On rappelle que les *p-values* associées aux coefficients indiquent si la covariable est significativement associée à la durée d'un arrêt. Si la *p-value* est inférieure à un niveau de seuil prédéfini (souvent fixé à 5%), alors on peut rejeter l'hypothèse nulle selon laquelle le coefficient est égal à zéro ; la covariable n'est pas significativement associée à la durée d'un arrêt.

On note qu'un grand nombre des covariables de notre modèle ne sont pas significatives. La non-significativité des variables pose un réel frein à notre modèle, et peut signifier que :

— Les variables non significatives sont des variables non pertinentes dans notre modèle : Il se peut que certaines variables incluses dans le modèle n'aient pas d'impact significatif sur la durée d'un arrêt. En effet, certaines variables peuvent ne pas être corrélées avec la variable d'intérêt ou peuvent ne pas fournir d'informations supplémentaires une fois que d'autres variables ont été incluses dans le modèle.

— L'hypothèse de proportionnalité n'est pas validée. Comme expliqué précédemment, le modèle de Cox est un modèle semi-paramétrique qui suppose que les effets des covariables sont proportionnels. Dans le cas où des variables ne répondent pas à cette hypothèse, le modèle peut rencontrer des difficultés à détecter leurs impacts, même si ces dernières ont en réalité un effet sur la variable cible. Ce phénomène peut être observé dans le cas de censure de données.

— Les variables non significatives sont mal catégorisées. Il se peut que ces dernières soient mal mesurées ou possèdent une faible variabilité. La détection de leur effet sur la durée d'arrêt devient difficile même si elles ont un effet significatif en réalité.

Il est important de faire particulièrement attention aux variables non significatives dans notre modèle. Dans la suite, nous les analyserons pour déterminer si elles sont pertinentes ou non. En effet, l'exclusion de variables non significatives peut améliorer la performance du modèle en réduisant sa complexité et améliorant la précision des estimations des paramètres en réduisant le nombre de ces dernières.

Dans un premier temps, nous regrouperons les variables non significatives qui peuvent être regroupées ou qui sont corrélées entre elles et que l'on considère intéressantes à explorer. Dans un second temps, nous appliquerons des algorithmes de sélection de variables pour améliorer la performance de notre modèle.

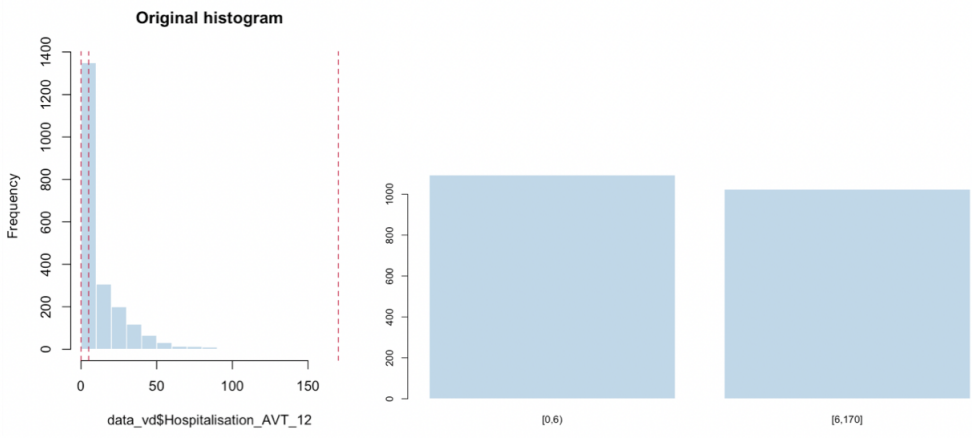
4.4.4 Modélisation après regroupement de variables

Pour les raisons évoquées précédemment, nous décidons de traiter les données comme expliqué dans ce qui suit.

Hospitalisation

Concernant les actes d'hospitalisation, nous décidons de ne garder qu'une seule variable appelée "Hospitalisation_AVT_12 " qui somme les actes des variables "Chambre_Part_AVT_12", "Sejour_Hospi_AVT_12" et "Forfait_Jour_Hospi_AVT_12".

Dans le but de réduire autant que possible la segmentation, nous décidons de diviser cette variable en 2 catégories. Nous souhaitons avoir un nombre équivalent d'individus dans chaque catégorie pour que ces dernières soient bien représentées. En observant l'histogramme de la variable, nous décidons de considérer une catégorie de 0 à 5 actes et une autre avec 6 actes et plus. Nous nous assurons que la répartition est quasi égale : 51.6% pour la première et 48.4% pour la seconde.

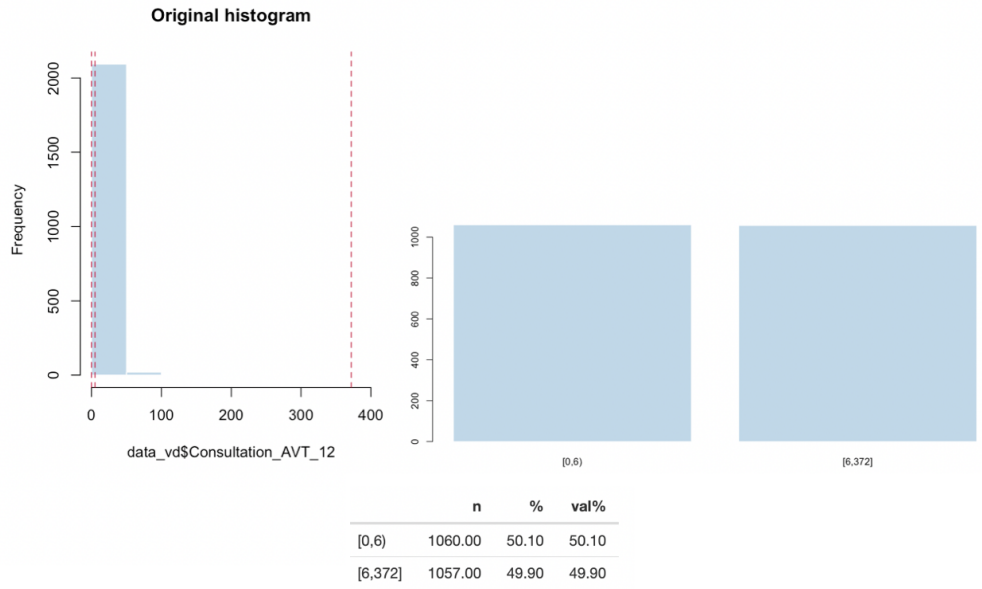


	n	%	val%
[0,6)	1093.00	51.60	51.60
[6,170]	1024.00	48.40	48.40

FIGURE 4.12 – Regroupement : variable Hospitalisation

Consultation

De même, nous décidons de regrouper les consultations de généralistes et de spécialistes. En effet, nous constatons qu'il y a une forte corrélation entre ces variables. Nous créons une variable "Consultation_AVT_12" qui somme ces dernières.



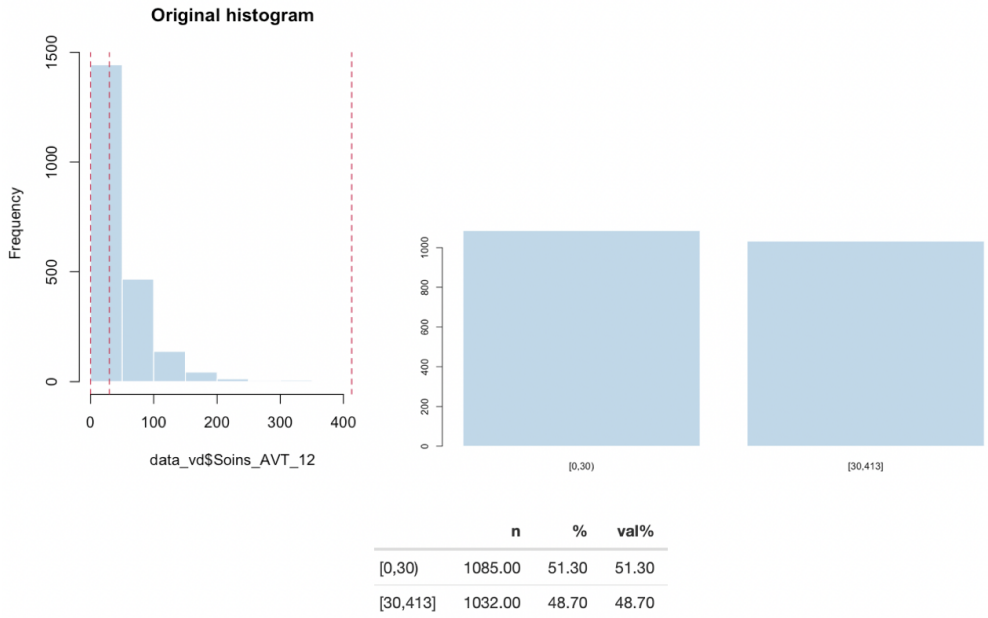
	n	%	val%
[0,6)	1060.00	50.10	50.10
[6,372]	1057.00	49.90	49.90

FIGURE 4.13 – Regroupement : variable Consultation

Nous observons que la majorité des observations ont entre 0 et 50 consultations. En appliquant la même répartition que pour l'hospitalisation, à savoir 0 à 5 actes et 6 actes et plus, nous obtenons que la répartition quasi égale : 50,1% pour la première catégorie et 49,9% pour la seconde catégorie.

Soins

De même, nous décidons de regrouper les consommations de médecine douce, de pharmacie, de dispositifs médicaux et des autres soins. En effet, nous constatons qu'il y a une corrélation importante entre ces variables. Nous créons une variable "Soins_AVT_12" qui somme ces dernières.



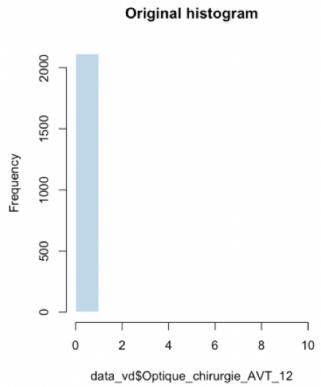
	n	%	val%
[0,30)	1085.00	51.30	51.30
[30,413]	1032.00	48.70	48.70

FIGURE 4.14 – Regroupement : variable Soins

Nous observons que la majorité des observations ont entre 0 et 50 consommations. Nous optons pour la catégorisation suivante : 0 à 29 actes et 30 actes et plus. Nous obtenons une répartition quasi égale : 51,3% pour la première catégorie et 48,7% pour la seconde catégorie.

Chirurgie Optique

En regardant de plus près la répartition de la variable "Optique_chirurgie_AVT_12", nous voyons que 99% des individus n'ont pas eu d'acte de chirurgie optique. Nous jugeons que la segmentation, entre les personnes ayant eu un acte et ceux n'ayant pas eu d'acte, n'est pas pertinente dans ce cas. En effet, les personnes ayant eu un acte sur ce poste seront sous représentées. Cette segmentation risque de complexifier le modèle sans réelle plus-value au modèle. Nous décidons de la supprimer de notre étude



Min	1st quartile	Median	Mean	3rd quartile	Max
0	0	0	0.0094	0	10

FIGURE 4.15 – Regroupement : variable Optique

Maternité

Nous rencontrons le même problème sur le poste de maternité. Nous pensons que ce poste est intéressant à exploiter : la maternité peut provoquer un arrêt, et peut nous aider à déterminer la durée de l'arrêt. Toutefois, par manque de données, nous ne pouvons garder une segmentation avec 50 individus. Nous décidons de nous en séparer.

	n	%	val%
[0,1)	2066.00	97.60	97.60
[1,15]	51.00	2.40	2.40

FIGURE 4.16 – Regroupement : variable Maternité

En suivant la même démarche, nous supprimons les variables Forfait_Jour_Psy_AVT_12, Cure_AVT_12 et Secteur_activite. Par ailleurs, nous estimons que les consommations dentaires n'ont pas d'impact significatif sur la durée d'un arrêt de travail, nous décidons de nous en séparer également.

Application

Nous tentons une seconde modélisation avec les variables réduites ou regroupées et retraitées. Nous espérons qu'en diminuant la complexité du modèle, ce dernier gagnera en performance.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
GENREM	0.0910635	1.0953385	0.0752161	1.211	0.226014	
collegeENSEMBLEDUPERSONNEL	-0.2981950	0.7421566	0.1310479	-2.275	0.022878	*
collegeNONCADRES	-0.5260712	0.5909220	0.1391582	-3.780	0.000157	***
AgeSurvenance	0.0015967	1.0015980	0.0031385	0.509	0.610918	
Mot_ATMALADIE	-0.2706408	0.7628905	0.1093996	-2.474	0.013366	*
Consultation_AVT_12	0.0052849	1.0052989	0.0028216	1.873	0.061066	.
Soins_AVT_12	0.0021124	1.0021146	0.0008072	2.617	0.008875	**
Hospitalisation_AVT_12_cat[6,170]	-0.0129705	0.9871132	0.0809921	-0.160	0.872766	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Concordance= 0.591 (se = 0.013)
Likelihood ratio test= 44.66 on 8 df, p=4e-07
Wald test = 48.87 on 8 df, p=7e-08
Score (logrank) test = 48.83 on 8 df, p=7e-08

FIGURE 4.17 – Coefficients du modèle après retraitements

L'intégralité des tests présentés précédemment ont des *p-values* inférieures à 5% : il existe au moins une variable explicative influente sur le taux de sortie de l'état d'incapacité. Nous nous intéressons aux intervalles de confiance à 95% présentés ci-dessous :

	exp(coef)	exp(-coef)	lower .95	upper .95
GENREM	1.0953	0.9130	0.9452	1.2693
collegeENSEMBLEDUPERSONNEL	0.7422	1.3474	0.5740	0.9595
collegeNONCADRES	0.5909	1.6923	0.4499	0.7762
AgeSurvenance	1.0016	0.9984	0.9955	1.0078
Mot_ATMALADIE	0.7629	1.3108	0.6157	0.9453
Consultation_AVT_12	1.0053	0.9947	0.9998	1.0109
Soins_AVT_12	1.0021	0.9979	1.0005	1.0037
Hospitalisation_AVT_12_cat[6,170]	0.9871	1.0131	0.8422	1.1569

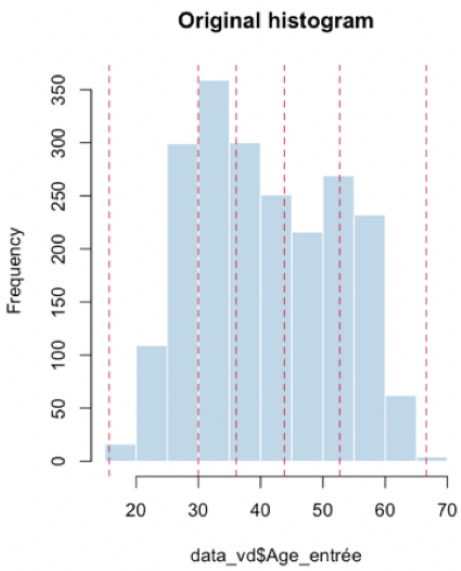
FIGURE 4.18 – Intervalle de confiance après retraitements

On note que selon notre modèle notre variable retravaillée "Soin_AVT_12" a un impact positif significatif sur notre modèle : plus l'assuré a eu de soins l'an précédant son arrêt, plus forte est la probabilité de sortir de l'arrêt. Le motif maladie, lui, a un impact négatif ; ce qui signifie qu'un arrêt pour motif d'accident de travail a tendance à avoir une durée moins importante. Enfin, les non-cadres de notre base de données ont tendance à avoir plus d'arrêts et de plus longue durée.

Néanmoins, plusieurs variables restent non significatives à 5% : le genre, l'âge à la survenance, la consultation ainsi que l'hospitalisation. Le poste de consultation est accepté à 10%, toutefois, l'intervalle de confiance n'assure pas la direction de l'association. L'hospitalisation ne possède pas d'impact significatif. Pourtant, on peut penser qu'une hospitalisation l'année qui précède un arrêt donne une indication sur l'état de santé de l'assuré et peut influer sur la durée de son arrêt. Cette non-significativité peut être liée au manque de données, notamment post segmentation. Le sexe n'est pas discriminant dans notre cas, ce qui correspond à nos premières observations dans la partie statistiques descriptives. Enfin, la non-significativité de la variable d'âge nous pose un réel problème. En effet, nous rappelons que l'âge est une composante principale pour toute modélisation dans ce cadre, notamment celle des tables réglementaires des BCAC. Nous décidons dans un premier temps de la transformer en variable catégorielle et tentons une modélisation pour voir si amélioration il y a.

Application : Âge Survenance

Nous catégorisons la variable de manière à s'assurer qu'il y ait un nombre suffisant d'individus dans chaque segmentation.



	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
GENREM	0.0759847	1.0789461	0.0826539	0.919	0.357932	
collegeENSEMBLEDUPERSONNEL	-0.2847762	0.7521826	0.1387755	-2.052	0.040164	*
collegeNONCADRES	-0.5027613	0.6048581	0.1479919	-3.397	0.000681	***
Mot_ATMALADIE	-0.2927230	0.7462288	0.1201277	-2.437	0.014819	*
Consultation_AVT_12	0.0051716	1.0051850	0.0032184	1.607	0.108082	
Soins_AVT_12	0.0016031	1.0016044	0.0008935	1.794	0.072775	.
Hospitalisation_AVT_12_cat[6,170]	-0.0226683	0.9775867	0.0882973	-0.257	0.797390	
AgeSurvenance_cat[35,45)	0.0525106	1.0539137	0.1153368	0.455	0.648908	
AgeSurvenance_cat[45,55)	0.1520320	1.1641975	0.1166266	1.304	0.192377	
AgeSurvenance_cat[55,68]	-0.0294621	0.9709677	0.1286663	-0.229	0.818884	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Concordance= 0.582 (se = 0.014)
Likelihood ratio test= 32.73 on 10 df, p=3e-04
Wald test = 34.94 on 10 df, p=1e-04
Score (logrank) test = 34.9 on 10 df, p=1e-04

	exp(coef)	exp(-coef)	lower .95	upper .95
GENREM	1.0789	0.9268	0.9176	1.2687
collegeENSEMBLEDUPERSONNEL	0.7522	1.3295	0.5731	0.9873
collegeNONCADRES	0.6049	1.6533	0.4526	0.8084
Mot_ATMALADIE	0.7462	1.3401	0.5897	0.9443
Consultation_AVT_12	1.0052	0.9948	0.9989	1.0115
Soins_AVT_12	1.0016	0.9984	0.9999	1.0034
Hospitalisation_AVT_12_cat[6,170]	0.9776	1.0229	0.8222	1.1623
AgeSurvenance_cat[35,45)	1.0539	0.9488	0.8407	1.3212
AgeSurvenance_cat[45,55)	1.1642	0.8590	0.9263	1.4632
AgeSurvenance_cat[55,68]	0.9710	1.0299	0.7545	1.2495

FIGURE 4.19 – Regroupement : variable Âge survenance

Nous constatons qu'aucune des catégories de la variable AgeSurvenance_cat n'est significative dans notre modèle. En revanche, l'interprétation nous semble plus cohérente que celle d'un coefficient unique positif. Nous pensons que cela est lié à une quantité insuffisante de données. Nous décidons de garder la variable et de tenter une autre approche pour améliorer la performance de notre modèle.

4.4.5 Modélisation après regroupement et sélection de variables

Nous envisageons d'utiliser des algorithmes de sélection de variables. Ces derniers aident à identifier les variables les plus importantes du modèle. Nous choisissons de tester la régression de Cox basée sur le critère d'information d'Akaike (AIC).

L'algorithme AIC utilise une mesure de l'ajustement du modèle et pénalise la complexité du modèle. Il s'agit d'une mesure qui prend en compte la log-vraisemblance et le nombre de paramètres dans le modèle. L'AIC pénalise les modèles via un certain nombre de paramètres, de sorte que les modèles plus simples avec des scores AIC les plus faibles sont favorisés.

Nous utilisons la fonction *stepAIC()* pour ajuster plusieurs modèles en éliminant et en ajoutant des variables au modèle initial jusqu'à trouver celui qui minimise l'AIC. Nous utilisons l'option *scope* pour forcer le modèle à prendre au moins l'âge comme covariable du modèle.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
collegeENSEMBLEDUPERSONNEL	-0.3051354	0.7370236	0.1308446	-2.332	0.019698	*
collegeNONCADRES	-0.5311881	0.5879061	0.1389701	-3.822	0.000132	***
AgeSurvenance	0.0021765	1.0021788	0.0031024	0.702	0.482961	
Mot_ATMALADIE	-0.2819352	0.7543226	0.1085520	-2.597	0.009398	**
Consultation_AVT_12	0.0050993	1.0051123	0.0028544	1.786	0.074021	.
Soins_AVT_12	0.0019362	1.0019381	0.0007594	2.550	0.010787	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance= 0.59 (se = 0.013)

Likelihood ratio test= 43.16 on 6 df, p=1e-07

Wald test = 46.86 on 6 df, p=2e-08

Score (logrank) test = 46.92 on 6 df, p=2e-08

	exp(coef)	exp(-coef)	lower .95	upper .95
collegeENSEMBLEDUPERSONNEL	0.7370	1.3568	0.5703	0.9525
collegeNONCADRES	0.5879	1.7010	0.4477	0.7720
AgeSurvenance	1.0022	0.9978	0.9961	1.0083
Mot_ATMALADIE	0.7543	1.3257	0.6098	0.9332
Consultation_AVT_12	1.0051	0.9949	0.9995	1.0108
Soins_AVT_12	1.0019	0.9981	1.0004	1.0034

FIGURE 4.20 – Coefficients du modèle final

La *p-value* des trois tests globaux est significative. Toutefois, l'âge ne semble pas être une variable discriminante dans notre modèle : variable continue ou catégorisée. Les résultats de ce modèle sont à considérer avec précaution. Nous décidons tout de même d'analyser les différents coefficients obtenus.

Dans la première partie de ce mémoire, nous avions observé un taux d'entrée en arrêt plus important chez les non-cadres ainsi que sur la catégorie ensemble du personnel. Dans cette partie, nous constatons que leurs arrêts ont tendance à durer plus longtemps sur ce périmètre.

On note que selon notre modèle notre variable retravaillée de Soins a un impact positif significatif sur notre modèle : plus l'assuré a eu de soins l'an précédant son arrêt, plus forte est la probabilité de sortir de l'arrêt. En effet, on peut penser qu'un individu qui prend soin de sa santé aura plus de facilité à "guérir" en cas de maladie ou se remettre en forme. Toutefois, cette interprétation est à considérer avec précaution, d'autant plus que la volumétrie des données ne nous permet pas à ce stade de pouvoir établir des analyses fiables.

Le motif maladie, lui, a un potentiel impact négatif ; ce qui signifie qu'un arrêt pour motif d'accident de travail a tendance à avoir une durée moins importante. Nous pensons que cela dépend fortement de notre périmètre et qu'on ne peut tirer conclusion de cette information.

4.5 Conclusion

Ce chapitre a pour unique objectif d'analyser l'éventuel lien entre la durée d'un arrêt et la consommation médicale. Il n'a en aucun cas pour vocation d'établir un modèle ajusté pour le provisionnement du risque arrêt de travail. L'unique conclusion que l'on peut en tirer est le potentiel impact positif de la consommation en pharmacie, en médecine douce et en autres soins, sur la probabilité de sortir de l'arrêt. Au vu du peu de données à disposition sur le périmètre double-équipé, il n'est pas possible d'envisager à ce stade un ajustement. Cette modélisation est à enrichir avec de nouvelles données (par exemple, des années supplémentaires d'observation) et reste à analyser et à ajuster avant toute utilisation pratique.

Chapitre 5

Impact d'un arrêt sur la consommation santé

Sommaire

L'arrêt de travail peut causer une dégradation de l'état de santé d'une personne. Cette dégradation s'accompagne souvent par une consommation médicale plus importante. D'un point de vue métier, il peut être intéressant d'exploiter cette piste dans l'optique d'éviter une dérive du portefeuille. Le but de cette partie est de s'intéresser à la consommation santé post-arrêt et de la comparer à celle des personnes n'ayant pas eu d'arrêt lors de la période d'observation. L'objectif est d'observer si les assurés ont des comportements différents, si les personnes ayant été arrêtées consomment plus sur certains postes et d'identifier et d'analyser les éventuels postes à risque.

5.1 Environnement et base de données

Pour ce faire, nous utiliserons les quatre bases de données SAS utilisées dans le premier chapitre :

- La base EFFECTIFS Santé comportant l'ensemble des effectifs bénéficiant d'une couverture Santé, leurs numéros de contrat, leurs identifiants assurés, leurs dates de naissance, leurs dates et durées d'affiliation ;
- La base EFFECTIFS Prévoyance comportant l'ensemble des effectifs bénéficiant d'une couverture Prévoyance, ainsi que les différentes garanties souscrites ;
- La base SINISTRES Prévoyance comportant l'ensemble des sinistres incapacité, invalidité et Décès (Y compris les sinistres non indemnisés) ;
- La base SINISTRES Santé comportant toutes les consommations en santé et les informations telles que le nombre d'actes, le montant remboursé...

Nous croisons les tables Effectifs Santé et Effectifs Prévoyance et sélectionnons les individus "double équipés" munis d'une garantie incapacité.

Concernant la période d'observation et pour les mêmes raisons que celles évoquées dans le premier chapitre de ce mémoire, nous nous limitons aux exercices 2017, 2018 et 2019.

Dans le but d'avoir un recul conséquent sur leurs prestations, nous exigeons une condition de 12 mois consécutifs d'observation des assurés.

Pour la sélection des prestations santé, nous optons pour le nombre d'actes par poste et décidons de conserver la répartition présentée précédemment :

— Chirurgie Yeux ;
— Dentaire. On regroupe l'ensemble des consommations dentaires en une seule famille ;
— Forfait Journalier ;
— Chambre Particulière ;
— Séjour Hospitalier. On regroupe l'ensemble des frais liés à un séjour hospitalier ;
— Maternité ;
— Psychiatrie et Psychologie ;
— Consultation Généraliste ;
— Consultation Spécialiste ;
— Pharmacie ;
— Médecine Douce ;
— Soins Autres ;
— Dispositifs médicaux ;
— Cures ;

Nous nous intéressons aux consommations santé après la survenance d'un arrêt et souhaitons les comparer aux consommations des assurés n'ayant pas eu d'arrêt. Nous sélectionnons les sinistres de santé survenus sur l'année qui suit la date de survenance. Dans le cas où il y aurait plusieurs arrêts, nous retenons la première date de survenance.

La période d'observation regroupe ainsi l'intégralité des consommations santé sur l'année qui suit la date de survenance, incluant la durée de l'arrêt. L'observation peut concerner uniquement la durée de l'arrêt dans le cas où l'arrêt dure plus d'un an. Nous comparons ces données à la moyenne des consommations sur 12 mois pour les individus n'ayant pas eu d'arrêt lors de la période d'observation.

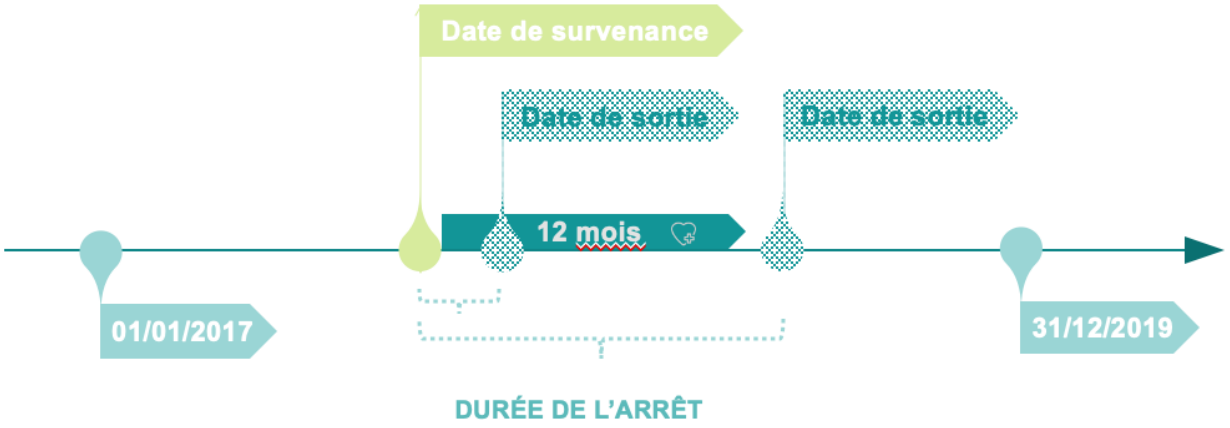


FIGURE 5.1 – Sélection temporelle des données santé post-arrêt

5.2 Analyses statistiques

Avant toute modélisation, nous effectuons des analyses statistiques. Nous nous intéresserons aux taux d'individu en arrêt de travail en fonction des comportements de la population en termes de consommation médicale.

Ainsi, nous nous intéressons à la segmentation du portefeuille selon le comportement des assurés en termes de consommation médicale. Nous évaluons le nombre d'individus ayant eu un arrêt dans la population sur le nombre total d'individus de la population.

Maternité : Nous nous intéressons aux taux de personnes ayant eu arrêt des individus ayant eu au moins un acte sur le poste de maternité et ceux ayant eu plus d'un acte sur l'année qui suit l'arrêt. Le taux est de 5,4% chez la population peu consommatrice alors que celle de la population plus consommatrice est de 22,8%. Intuitivement, les femmes consommant en acte de maternité sont plus susceptibles d'être en arrêt.

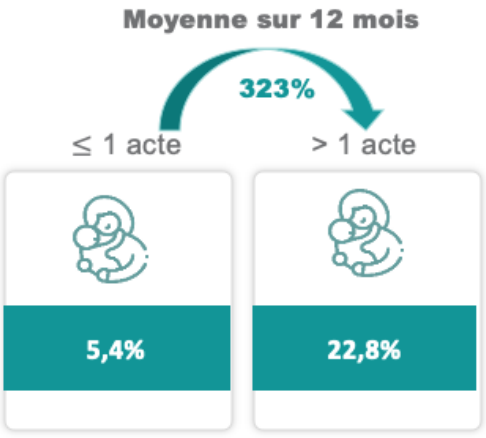


FIGURE 5.2 – Taux d'arrêt de travail : Maternité

Pharmacie et Médecine Douce : En nous intéressant aux consommations de pharmacie et de médecine douce, nous constatons que l'écart est fort sur ces actes. En effet, nous pensons que les variables liées à la consommation de médicaments ou de médecine douce peuvent être discriminantes dans nos modèles ; une personne ayant eu plusieurs consommant "beaucoup" en pharmacie ou en médecine douce peut être plus susceptible d'être en arrêt de travail.

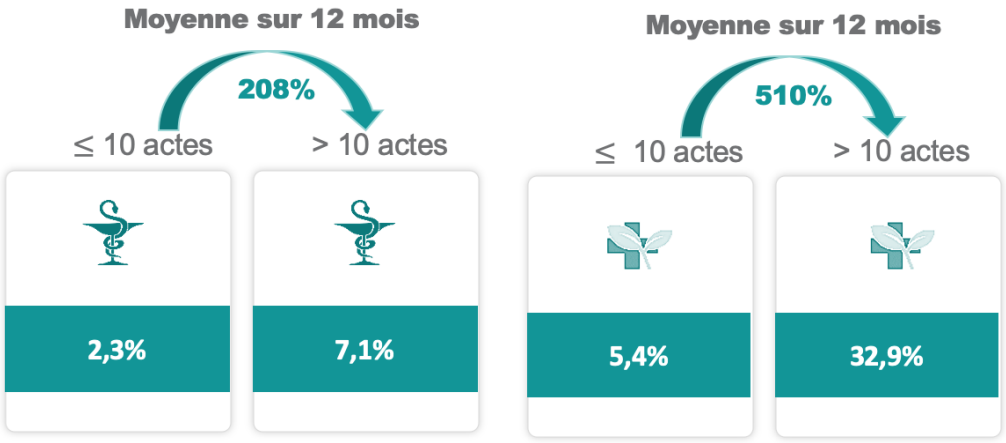


FIGURE 5.3 – Taux d'arrêt de travail : Pharmacie et Médecine douce

5.3 Modélisation

5.3.1 Application

Nous conservons par ailleurs la même démarche : nous optons pour une modélisation de type GLM et tentons d'améliorer les prédictions du modèle avec les mêmes méthodes utilisées au premier chapitre. Nous utilisons en particulier, le sur-échantillonnage ainsi que l'*optimal binning* afin d'améliorer la performance du modèle.

Dans ce qui suit, nous n'expliciterons pas les étapes réalisées et présenterons uniquement les résultats du modèle optimisé. Nous appliquons le modèle sur les variables traitées et catégorisées et obtenons les résultats ci-dessous :

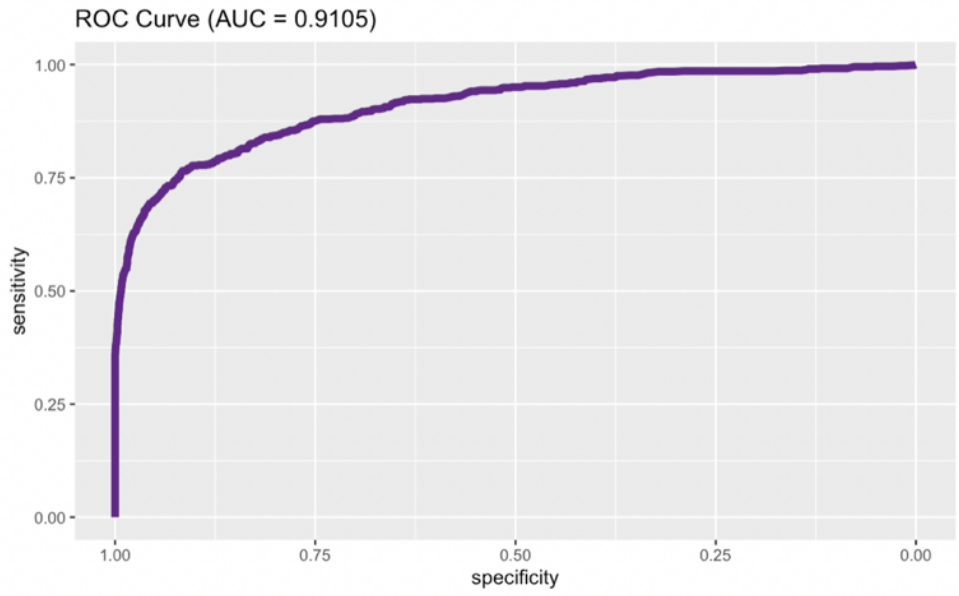


FIGURE 5.4 – Courbe ROC GLM Arrêt de Travail

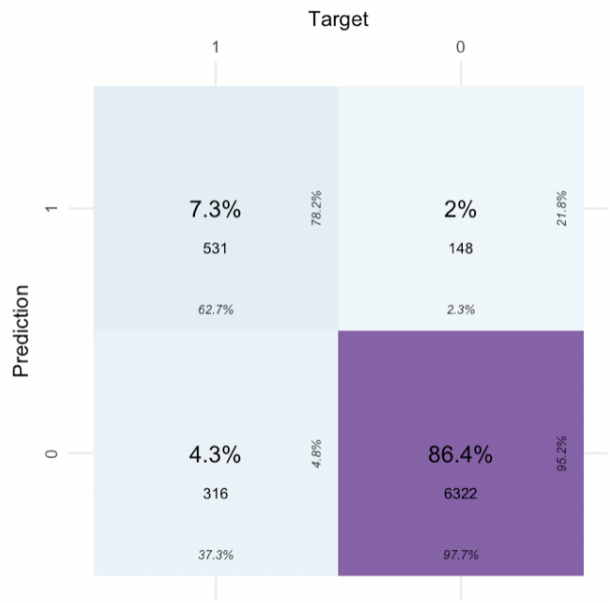


FIGURE 5.5 – Matrice de confusion GLM Arrêt de Travail

Les résultats du modèle final sont très bons : l'AUC du modèle est de 0.91, le F1-Score de 0.69. En nous intéressant à la matrice de confusion, nous constatons que dû au manque de données avec arrêt, le modèle prédit mieux les non-arrêts que les arrêts : 98% des non-arrêts prédits sont corrects, alors que 63% des arrêts prédits sont des vrais. Toutefois, la sensibilité du modèle demeure satisfaisante.

5.3.2 Analyse

Nous nous intéressons à présent à l'interprétation de la sortie du modèle. Nous rappelons que les coefficients d'un modèle GLM mesurent l'effet de chaque variable indépendante sur la variable

dépendante, après avoir tenu compte de l'effet des autres variables indépendantes incluses dans le modèle. Le coefficient associé à chaque variable correspond ainsi à une mesure de la contribution de cette variable à la prédiction de la variable cible. Plus grand est le coefficient, plus l'effet de la variable sur la variable cible est important.

	Est.	S.E.	z val.	p
(Intercept)	-3.62	0.15	-23.95	0.00
GENREM	-0.38	0.06	-6.72	0.00
collegeENSEMBLEDUPERSONNEL	-0.86	0.10	-8.29	0.01
collegeNONCADRES	-0.16	0.11	-1.47	0.14
Age_entrée[25,30)	0.31	0.12	2.54	0.01
Age_entrée[30,35)	0.35	0.12	2.94	0.00
Age_entrée[35,40)	0.30	0.12	2.45	0.01
Age_entrée[40,45)	0.37	0.13	2.98	0.00
Age_entrée[45,50)	0.42	0.13	3.37	0.00
Age_entrée[50,60)	0.34	0.12	2.94	0.00
Age_entrée[60 et plus)	0.15	0.18	0.79	0.43
Sejour_Hospi_APR_cat[20,50)	-0.04	0.12	-0.30	0.76
Sejour_Hospi_APR_cat[50 et plus)	1.96	0.47	4.16	0.00
Chambre_Part_APR_cat	1.01	0.10	10.12	0.00
Maternite_APR_cat	2.90	0.17	17.07	0.00
Forfait_Jour_Psy_APR_cat[5,10)	-2.08	0.74	-2.80	0.01
Forfait_Jour_Psy_APR_cat[10 et plus)	5.71	1.07	5.32	0.00
MEDECINE_DOUCE_APR_cat	-0.77	0.06	-13.39	0.00
Consult_Gen_APR_cat[5,17.5)	0.29	0.07	3.93	0.00
Consult_Gen_APR_cat[17.5,40)	0.64	0.10	6.45	0.00
Consult_Gen_APR_cat[40,50)	0.91	0.18	5.16	0.00
Consult_Gen_APR_cat[50 et plus)	1.25	0.16	7.84	0.00
Consult_Spe_APR_cat[8,30)	0.66	0.07	8.84	0.00
Consult_Spe_APR_cat[30,60)	1.40	0.17	8.13	0.00
Consult_Spe_APR_cat[60 et plus)	2.95	0.47	6.34	0.00
PHARMACIE_APR_cat[22,50)	1.30	0.09	14.39	0.00
PHARMACIE_APR_cat[50,76)	1.73	0.10	16.81	0.00
PHARMACIE_APR_cat[76,200)	1.59	0.10	15.22	0.00
PHARMACIE_APR_cat[200,300)	1.22	0.16	7.76	0.00
PHARMACIE_APR_cat[300 et plus)	2.40	0.18	13.31	0.00
Forfait_Jour_Hospi_APR_cat[5,10)	1.08	0.19	5.74	0.00
Forfait_Jour_Hospi_APR_cat[10 et plus)	2.26	0.30	7.44	0.00
SOINS_AUTRES_APR_cat[90,164)	1.98	0.09	21.18	0.00
SOINS_AUTRES_APR_cat[164,300)	1.77	0.16	11.00	0.00
SOINS_AUTRES_APR_cat[300et plus)	2.53	0.30	8.48	0.00

MODEL FIT:
$\chi^2(34) = 10458.73$, $p = 0.00$
Pseudo-R² (Cragg-Uhler) = 0.59
Pseudo-R² (McFadden) = 0.50
AIC = 10585.48, *BIC* = 10875.42

FIGURE 5.6 – Coefficients GLM Arrêt de Travail

Les R^2 de Cragg-Uhler et de McFadden sont respectivement de 0.59 et 0.50 ; le modèle explique une proportion importante de la variance dans les données.

Ainsi, en analysant les coefficients de notre modèle, nous sommes en mesure d'évaluer l'importance

relative de chaque variable dans la prédiction du risque d'arrêt de travail.

Hospitalisation

Nous remarquons que le coefficient associé à la tranche "50 et plus" de la variable Séjour_Hospitalier, celui associé au poste Chambre Particulière ainsi que les coefficients des niveaux de la variable Forfait Journalier sont tous positifs. Le fait qu'une personne en arrêt consomme plus sur ce poste est cohérent. Intuitivement, un salarié en hospitalisation a plus de chances d'être en arrêt.

Maternité

Le coefficient associé à la variable "Maternité" montre une association positive avec la variable cible "arret". Une personne ayant eu un arrêt de travail consomme plus sur le poste de maternité, sur l'année après la date de survenance de l'arrêt. En effet, on peut penser que les femmes en arrêt de travail pour cause de maternité peuvent avoir des problèmes de santé liés à la grossesse ou même à l'accouchement, nécessitant ainsi une surveillance accrue et des actes de soins supplémentaires. Elles peuvent ainsi se rendre plus souvent chez leur gynécologue ou l'obstétricien, ce qui entraînerait une augmentation de la fréquence d'actes sur le poste de maternité.

Consultation, Pharmacie et Soins autres

Les coefficients associés aux différents niveaux des variables Consultation généraliste et Consultation Spécialiste, Pharmacie et Soins autres sont tous également positifs. En effet, les personnes qui sont en arrêt de travail sont plus sujettes à des problèmes de santé plus importants que les personnes en activité professionnelle. Cela peut nécessiter un traitement plus lourd ou plus régulier qui se traduit en une augmentation sur les postes de pharmacie, consultations, hospitalisation ainsi que d'autres soins.

5.4 Conclusion

Ce chapitre a pour unique objectif d'analyser le lien, sujet de ce mémoire, post-arrêt. Nous nous sommes intéressés aux consommations durant l'année suivante la date de survenance, que l'individu soit en arrêt ou qu'il en soit sorti. Nous constatons un réel impact de l'arrêt sur la consommation en Santé : les prestations de consultations, pharmacie, maternité et hospitalisation sont bien plus importantes. Ainsi, un arrêt engendre une augmentation des dépenses santé. Il serait intéressant de regarder les consommations en dehors de la période en arrêt pour évaluer s'il existe un éventuel impact "long terme" sur la santé des assurés.

Conclusion

L'objectif de ce mémoire est de mettre en avant le lien entre l'assurance Santé et l'assurance Prévoyance, en particulier la garantie Incapacité. En effet, ces risques sont généralement étudiés de manière indépendante. Or, les statistiques descriptives révèlent des écarts de taux d'absentéisme entre différentes populations du portefeuille ayant des comportements différents en termes de consommation Santé. Pour mener une analyse plus poussée, nous avons opté pour différentes modélisations afin d'observer les effets de chaque variable indépendamment des autres. Nous avons analysé ce lien sous plusieurs angles.

Dans un premier temps, nous avons étudié l'impact de la consommation en santé sur l'exposition au risque d'arrêt de travail. Nous utilisons un GLM que nous avons optimisé de diverses manières. Dans un second temps, nous avons étudié l'impact de cette consommation sur la durée de l'arrêt. Pour ce faire, nous optons pour un modèle de Cox que nous cherchons à améliorer via une segmentation. Enfin, nous nous sommes intéressés à l'influence d'un arrêt sur les consommations en santé des assurés.

En conclusion, nous pouvons affirmer qu'un lien existe entre ces deux risques. D'une part, les données Santé permettent d'améliorer la prédiction du risque arrêt de travail, notamment en termes de probabilité d'entrée en arrêt. Certaines variables relatives aux consommations Santé telles que les consultations ou encore l'hospitalisation sont significatives, parfois plus que les variables usuelles. D'autre part, et même si les résultats du modèle de durée sont à considérer avec précaution, nous pouvons envisager une piste selon laquelle il y aurait un impact positif de la consommation de soins sur la durée de l'arrêt : plus un assuré est consommateur en soins de santé tels que la pharmacie ou encore la médecine douce, plus il prend soin de sa santé, et plus le risque de rester en arrêt est faible et la probabilité de reprendre son activité professionnelle est importante. Enfin, nous remarquons un réel changement de consommations santé chez les individus ayant eu un arrêt. Une augmentation de la consommation est observée sur plusieurs postes tels que la pharmacie, les consultations, mais aussi les actes liés à la maternité.

Dans le cadre du pilotage du risque d'arrêt de travail, les modèles de prédiction de la probabilité d'entrée sont intéressants à utiliser pour affiner la surveillance du portefeuille double-équipé. Leurs utilisations pourront permettre de prévenir et d'anticiper une éventuelle dérive de l'absentéisme. Ils peuvent également constituer un indicateur complémentaire pour les ajustements tarifaires et compléter les programmes de prévention dans le but d'atténuer l'effet d'une éventuelle hausse de

la sinistralité en prévoyance. De plus, l'utilisation de ce lien dans le cadre de tarification constitue une piste envisageable. Dans ce cas, des études complémentaires avec une volumétrie de données plus importante doivent être menées.

Par ailleurs, il faudra également veiller à mener des études complémentaires et à enrichir l'étude menée par des années d'observations supplémentaires pour envisager l'utilisation des modèles de prédiction de la durée d'un arrêt. Concernant les consommations santé post-arrêt, cette étude montre qu'un arrêt engendre souvent une augmentation des prestations santé et permet d'identifier les éventuels postes "à risque". En outre, il peut être intéressant de mener une étude complémentaire sur les consommations après la fin d'un arrêt pour évaluer le risque d'un éventuel impact "long terme" sur la santé des assurés.

Enfin, en excluant les années 2020 et 2021, notre étude ne permet pas la mise en avant des risques psychosociaux. En effet, à la suite de la pandémie de Covid-19, les arrêts liés à ces risques ont considérablement augmenté. Un nombre plus important d'assurés souffrent de troubles psychologiques les empêchant de maintenir leurs activités professionnelles : la dépression, l'anxiété liée au contexte sanitaire, économique et politique, mais aussi à l'appréhension du retour en entreprise. Les données santé pourraient être particulièrement utiles pour anticiper ce type d'arrêt ; l'observation d'actes de psychologie pourrait notamment alerter sur le risque.

Bibliographie

Documents internes (2021-2022). Crédit Agricole.

PLANCHET, Frédéric (2022). *Modèles de durée, applications actuarielles.*

GUIZOUARN Jean Charles, MARESCAUX Nicolas (2004). *Assurance Santé, Segmentation et compétitivité.* Economica.

SAADAOUI Ali, DUMONT Emmanuel (2019). *Impact de la consommation Santé sur la probabilité d'entrée en arrêt de travail et sur sa durée.* Institut des Actuaires.

PLANCHET, Frédéric (2016). *Tarification et modèles linéaires généralisés.*

Karine, NGUYEN (2018). *Méthodes de provisionnement du maintien en incapacité des contrats dits franchises courtes.* Institut des Actuaires.

NAKACHE Jean-Pierre, CONFAIS Josiane (2003). *Statistique explicative appliquée.* Editions TECHNIP.

FFA - Quel rôle demain pour les complémentaires santé? (2022). URL : `https://www.franceassureurs.fr/nos-positions/lassurance-qui-protege/quel-role-demain-pour-les-complementaires-sante/`.

FFA - Coronavirus COVID-19 et assurance (2021). URL : `https://www.franceassureurs.fr/lassurance-protege-finance-et-emploie/lassurance-protege/lassurance-en-pratique-pour-les-particuliers/coronavirus-covid-19-et-assurance`.

Argus de l'assurance (2020-2022). URL : `https://www.argusdelassurance.com/`.

GitHub (2008-2022). URL : `https://github.com/`.

Freakonometrics (2007-2022). URL : `https://freakonometrics.hypotheses.org/`.

Pratique de la Régression Logistique (2017). URL : http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf.

Regularized regression (2017). URL : https://eric.univ-lyon2.fr/~ricco/cours/slides/regularized_regression.pdf.

Ridge and Lasso Regression (2018). URL : https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b.

Table des figures

Annexe A

Complément sur les statistiques descriptives

A.1 Données Santé précédant l'AT

Forfait Journalier Hospitalier : En comparant le taux d'entrée en arrêt des individus ayant eu au moins un acte de forfait journalier hospitalier et ceux ayant eu plus d'un acte sur l'année. Nous constatons que le taux est de 5,9% chez la population peu consommatrice et de 12,2% chez la population plus consommatrice.

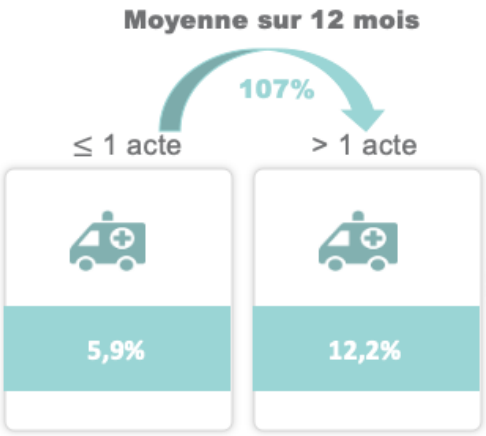


FIGURE A.1 – Taux d'entrée en AT : Forfait Journalier Hospitalier

Chambre particulière : Nous notons un écart d'autant plus fort sur des actes de chambre particulière. Le taux est de 5,9% chez la population peu consommatrice et de 16,9% chez la population plus consommatrice. Intuitivement, une personne ayant eu plus d'une consommation de chambre particulière est plus susceptible de devoir arrêter son activité professionnelle.

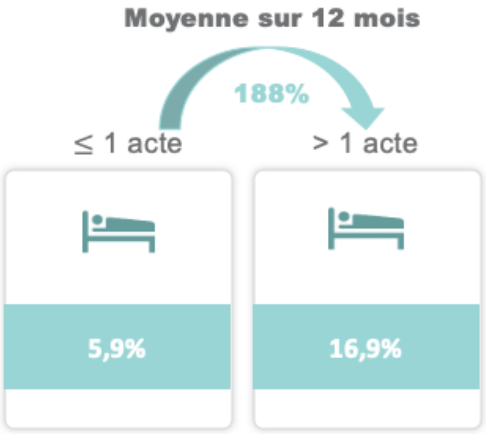


FIGURE A.2 – Taux d'entrée en AT : Chambre particulière

Séjour Hospitalier : De la même manière, nous constatons un taux d'entrée en arrêt de travail plus important chez la population ayant eu plus de deux actes sur le poste "Séjour Hospitalier". Ce dernier passe de 4,4% à 8,3%.

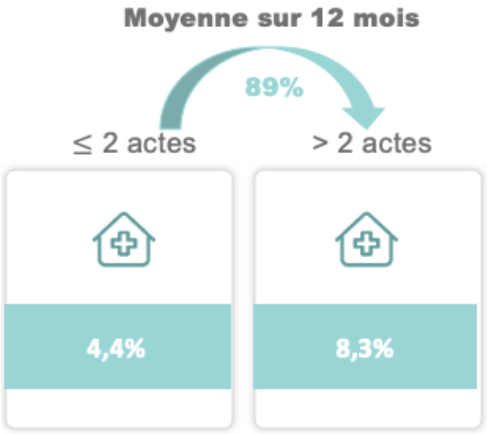


FIGURE A.3 – Taux d'entrée en AT : Séjour Hospitalier

A.2 Données Santé suivant l'AT

Consultation spécialiste : Nous décidons de comparer le taux de personnes des individus ayant eu entre 0 et 10 actes de consultation chez un spécialiste et ceux ayant eu plus de 10 actes sur l'année qui suit l'arrêt. Nous constatons que le taux est de 4,4% chez la population peu consommatrice en consultation d'un spécialiste et de 21,3% chez la population plus consommatrice. Les individus plus grands consommateurs sur ce poste ont plus de chances d'avoir eu un arrêt.

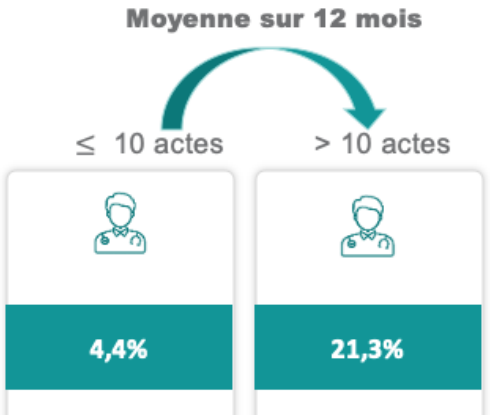


FIGURE A.4 – Taux d'arrêt de travail : Consultation spécialiste

Hospitalisation : En nous intéressant aux actes d'hospitalisation, nous constatons également des écarts élevés entre les populations peu consommatrices et celle plus consommatrices. Les taux de personnes eu arrêt des individus ayant eu entre 0 et 5 actes d'actes d'hospitalisation sont bien plus faibles que ceux ayant eu plus de 5 actes sur l'année qui suit l'arrêt. Intuitivement, un individu qui a des actes d'hospitalisation est plus susceptible d'avoir arrêter son activité professionnelle.

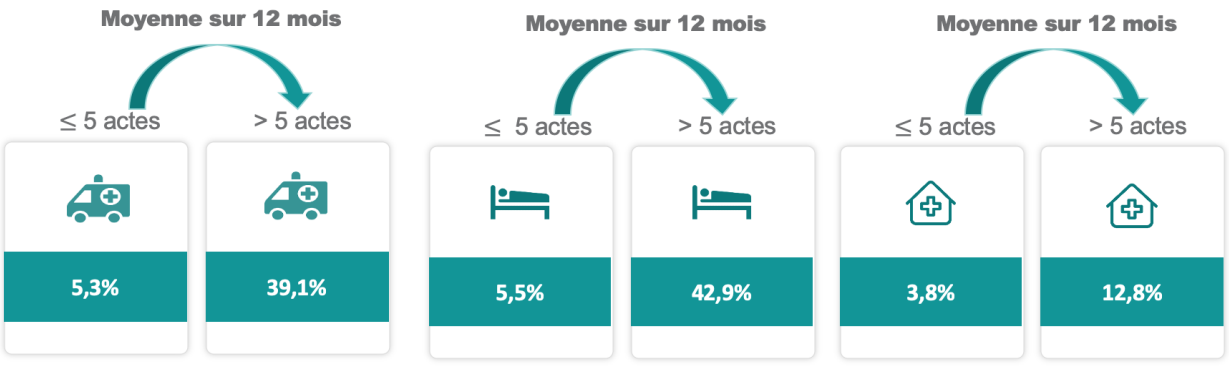


FIGURE A.5 – Taux d'arrêt de travail : Hospitalisation

Annexe B

Graphique de score théorique

Un modèle "parfait" aurait des densités de "score" réparties comme ci-dessous. La concentration des non-arrêts se situerait à gauche et la concentration des arrêts se situerait à droite du graphique. Ainsi, la détermination du seuil serait facilitée et les performances du modèle excellentes.

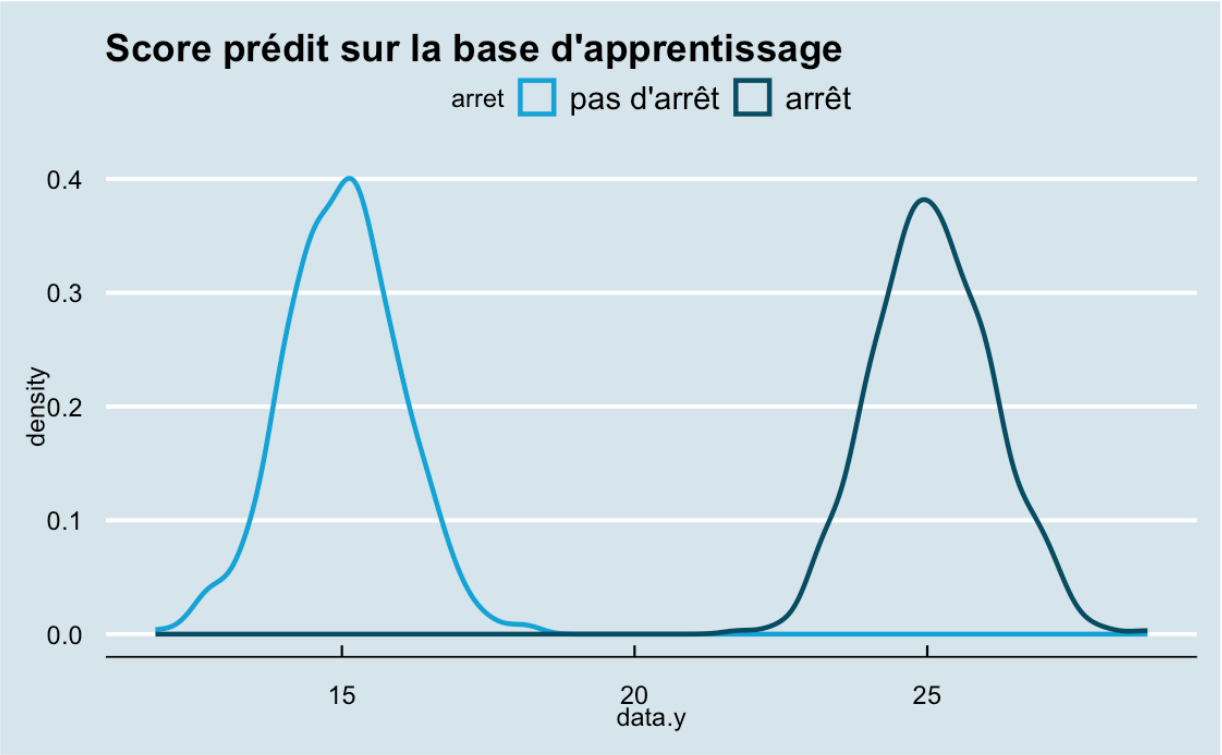


FIGURE B.1 – Score théorique