Mémoire présenté devant le Cnam

pour l'obtention du diplôme du Master Droit Economie Gestion mention Actuariat et l'admission à l'Institut des Actuaires

le 20/01/2022

Par : Ellen OLYMPIO

Titre: Provisionnement des sinistres d'arrêts de travail avec des méthodes d'apprentissage automatique – Impacts en termes de gestion des risques

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

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

Membres présents du jury de l'Institut des Actuaires
David DUBOIS
Jean-Marie NESSI

Membres présents du jury du Cnam
Présidente du jury : Sandrine LEMERY
Nathanaël ABECERA
François WEISS

Entreprise :

Nom : AG2R La Mondiale

Directeur de mémoire en entreprise :
Nom : Isabelle CHEROUVRIER
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

Secrétariat :

Bibliothèque :

Signature du candidat

Remerciements

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je souhaite témoigner toute ma gratitude.

Je voudrais, en premier lieu, exprimer toute ma reconnaissance à Isabelle Cherouvrier, ma directrice de mémoire chez AG2R La Mondiale, pour sa disponibilité et la constance de son soutien depuis de nombreuses années.

Je remercie tout particulièrement Slim Belghith, pour sa précieuse collaboration et ses conseils judicieux autour des données indispensables à la réalisation de ce mémoire.

Je tiens à témoigner toute ma gratitude à Ginette Agopomé, Actuaire, pour son investissement, sa patience et son accompagnement à toutes les étapes du mémoire.

Je voudrais adresser ma reconnaissance à Héloïse Yvroud, Actuaire, pour sa disponibilité et sa pédagogie lors d'échanges constructifs autour des méthodes d'apprentissage automatique.

Mes remerciements vont également à l'ensemble de l'équipe pédagogique de la Chaire d'Actuariat du Cnam.

Enfin, je remercie toutes celles et tous ceux qui, à des degrés divers, m'ont permis d'avancer.

Résumé

Les organismes assureurs sont tenus réglementairement de provisionner les prestations futures d'un sinistre d'arrêt de travail en cours d'indemnisation jusqu'à la fin de l'arrêt. L'approche classique utilisée est basée sur les triangles de *Chain Ladder* pour les provisions pour sinistres à payer (PSAP) et le calcul paramétrique pour les provisions mathématiques. Toutefois, face à la sinistralité importante d'un contrat (augmentation continue de la sinistralité, population sous risque avec des caractéristiques particulières, etc.), il est opportun d'envisager des méthodes alternatives de calcul des provisions.

L'objectif du mémoire est d'évaluer la pertinence des méthodes d'apprentissage automatique pour challenger les méthodes de provisionnement usuelles. La modélisation mise en avant s'inspire de l'article de recherche « *Tree-based censored regression with applications to insurance* » de Lopez et al. (2015) [2] qui propose une méthode de provisionnement en présence d'observations censurées. Elle s'appuie sur des données historiques réelles d'arrêts de travail pour lesquels la charge ultime est disponible (*backtesting*). La mise en œuvre réalisée avec des données nouvelles a permis de poser un regard critique tant sur le provisionnement que sur la tentative d'explication de la sinistralité élevée du portefeuille.

Une application opérationnelle illustre l'apport du modèle prédictif et explicatif en termes de gestion des risques pour un pilotage ciblé du contrat. D'une part, la fonction Gestion des Risques et plus particulièrement la Fonction Actuarielle induites par la directive de Solvabilité 2 [1] pourront se faire leur propre opinion sur les méthodes actuelles de provisionnement et proposer des améliorations pour le suivi prospectif des arrêts de travail. D'autre part, face à un premier état des lieux des absences liées aux arrêts de travail, de nouveaux outils de gestion des arrêts de travail pourraient être envisagés et ouvrir la voie à un pilotage approprié de l'absentéisme.

Mots clés : provisionnement, arrêt de travail, maintien de salaire, incapacité, censure, apprentissage automatique, arbres de régression, CART, *Random Forest*, *Gradient Boosting*, IPCW, Kaplan-Meier, gestion des risques

Abstract

Insurers are required, by regulation, to fund future benefits payments for a work interruption during compensation until its end. The traditional approach used is based on the Chain Ladder method for claims in course of settlement reserving and the parametric calculation of mathematical actuarial reserves. However, faced with a high loss experience of a contract (continuous increase in claims, population at risk with specificities, etc.), it is advisable to consider alternative methods of reserving calculation.

The purpose of the master thesis is to define the implementation of machine learning algorithms to challenge current reserving methods. The modeling is based on the research article "Tree-based censored regression with applications to insurance" by Lopez and al. (2015) [2] which presents a reserving method including censored data. Built from real historical data of work interruption for which the ultimate cost is available, the implementation carried out with new data made it possible to take a critical look at both the reserving and the attempt of explanation of the high loss experience.

An operational application illustrates the contribution of the predictive and explanatory model in terms of risk management for targeted monitoring of the contract. On the one hand, the Risk Management function and more particularly the Actuarial Function induced by the Solvency 2 directive [1] will be able to form their own opinion on current reserving methods and suggest improvements for the prospective monitoring of work interruptions. On the other hand, faced with a first overview of absences related to work interruption, new monitoring tools could be considered and pave the way for an efficient absenteeism management policy.

Keywords: reserving, work interruption, salary continuance, incapacity, censored observations, machine learning, regression trees, CART, Random Forest, Gradient Boosting, IPCW, Kaplan-Meier, risk management

Sommaire

Introduction

Les provisions techniques sont les réserves de prestations que doivent constituer les compagnies d'assurance pour être en mesure d'honorer leurs engagements futurs : indemnités journalières et rentes en arrêt de travail, capitaux en cas de vie ou de décès, frais médicaux, etc. La thématique du provisionnement est d'autant plus importante qu'elle représente une exigence de la directive de Solvabilité 2 [1], qui impose que les provisions calculées soient *best estimate*. L'approche classique utilisée pour déterminer le montant des provisions, dans le cas des sinistres en frais médicaux par exemple, est celle des triangles de *Chain Ladder*. Cette méthode agrégée doit sa popularité à sa facilité d'implémentation sur des données variées. Ces données incluent les paiements cumulés, les charges de sinistres déclarés ou tardifs, et le nombre de sinistres. Selon le risque, le calcul des provisions peut également être réalisé sinistre par sinistre. C'est le cas des provisions mathématiques constituées pour les sinistres en incapacité et invalidité, ou pour les garanties décès, mais également le cas des provisions en assurance automobile. Ces provisions prennent en compte les charges de sinistres en cours de paiement et non clos. Elles ne permettent pas de prendre en compte les sinistres tardifs. Les provisions pour sinistres tardifs peuvent ensuite être estimées par le biais des méthodes agrégées telles que les triangles de *Chain Ladder*.

Dans le cadre des arrêts de travail, les assureurs ont plus particulièrement l'obligation de provisionner chaque sinistre d'arrêt de travail en incapacité et invalidité en cours d'indemnisation au sein de leur portefeuille. Cette provision matérialise les prestations futures que devra verser l'assureur jusqu'à la fin de l'arrêt. Les provisions représentent généralement le poste le plus important d'un compte de résultats sur ces risques.

A l'ère du big data, des méthodes d'apprentissage automatique font leur preuve dans le domaine assuranciel. Les algorithmes de *machine learning* se révèlent être des outils très utiles et adaptés à l'analyse des données du monde de l'assurance et permettent d'individualiser les tarifs ainsi que le provisionnement. Parmi ces méthodes, l'approche développée par Lopez et al. (2015) [2] dans leur article « *Tree-based censored regression with applications to insurance* » propose la prédiction de la charge ultime d'un sinistre, clos ou toujours ouvert (pour lequel on ne connaît pas la charge ultime), en utilisant des arbres de régression. L'originalité de cette méthode réside dans le fait d'être appliquée entre autres à des observations censurées, en introduisant une pondération comme levier de correction du biais induit par la censure.

L'objectif de ce mémoire est de juger de l'efficacité des méthodes d'apprentissage de provisionnement et de montrer leur apport sur l'estimation des provisions des sinistres d'arrêts de travail au sein des compagnies d'assurance. Ces méthodes pourront d'une part, fournir un nouvel éclairage pour challenger le provisionnement réalisé dans le cadre de l'environnement social, et d'autre part permettre de stabiliser les boni-mali de provisionnement.

En partie 1 de ce mémoire, est présenté le contexte des arrêts de travail. Le cadre général du provisionnement est ensuite abordé, afin de préciser les enjeux auxquels sont confrontées les entreprises d'assurance. Le provisionnement associé aux arrêts de travail y est également décrit. Dans la partie 2, les choix méthodologiques retenus sont exposés. La partie 3 présente les statistiques exploratoires du portefeuille de l'étude. La partie 4 est consacrée à l'application des modèles choisis aux données ainsi qu'à l'analyse des résultats obtenus. Une application opérationnelle en termes de gestion des risques pour le pilotage ciblé du portefeuille étudié sera mise en œuvre en partie 5.

Partie 1 – Arrêts de travail et provisionnement

1. Arrêts de travail

1.1. Contexte

Depuis plusieurs années, les dépenses liées aux arrêts de travail sont en constante hausse. La Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (DREES), dans son analyse des dépenses de santé en 2019 [3], estime le montant des indemnités journalières prises en charge par la Sécurité sociale à 15,7 milliards d'euros. La DREES constate une augmentation continue des indemnités journalières :

- En 2014, la hausse est de +4,2% ;
- Entre 2015 et 2018, elle évolue entre +2% et +3,8% ;
- En 2019, elle atteint +4,5%.

En particulier, la constante augmentation des arrêts maladie s'explique par différents facteurs :

- La population active vieillit, avec une hausse de +3,2 points de la population des travailleurs de plus de 60 ans entre 2010 et 2019 ;
- Les arrêts de travail pour les actifs les plus âgés et plus fragiles sont plus longs, avec l'état de santé qui se dégrade. Le travail précaire participe également à cette augmentation, du fait de la pénibilité des missions réalisées ;
- L'évolution des salaires qui servent de référence au calcul des indemnités journalières explique en partie cette hausse.

Les provisions mathématiques pour les arrêts de travail constituent un enjeu particulièrement important pour les organismes assureurs. L'enjeu est d'autant plus important qu'il est fonction des taux non-vie en vigueur[1] au sein de l'environnement social. Plus les taux sont élevés, plus les provisions techniques à constituer sont faibles. Mais face à la problématique des taux bas voire négatifs, les provisions techniques augmentent considérablement.

Les acteurs les plus impactés par les enjeux du provisionnement des arrêts de travail restent les institutions de prévoyance, spécialisées dans l'assurance santé et prévoyance des salariés pour les branches professionnelles.

1.2. Risques d'arrêts de travail

L'arrêt de travail est une garantie non-vie et on distingue 3 différents risques.

1.2.1. Maintien de salaire

En complément des versements de la Sécurité sociale, l'employeur est tenu de maintenir pendant une durée déterminée un certain niveau de la rémunération du salarié en arrêt de travail. Ce complément de salaire intervient après 10 jours d'arrêt de travail. Cette obligation a été mise en place par la loi de mensualisation de 1978 et revue par l'Accord National Interprofessionnel (ANI) en 2008. L'ancienneté requise au sein de l'entreprise pour le salarié est réduite de 3 ans à 1 an, et la franchise passe de 10 jours à 7 jours.

[1] Le taux technique non-vie est le taux d'actualisation maximal que les organismes assureurs ont le droit d'utiliser pour l'évaluation des engagements arrêt de travail et dépendance. Il représente 75% de la moyenne sur 24 mois du Taux Moyen d'emprunt d'Etat (TME) pour des échéances supérieures à 7 ans.

L'employeur peut financer ce complément d'indemnisation sur sa propre trésorerie ou confier à un organisme assureur la gestion de ces engagements en souscrivant un contrat de mensualisation. Les cotisations sont à la charge exclusive de l'employeur. La durée de la garantie maintien de salaire augmente de 10 jours par 5 ans d'ancienneté. Au terme du maintien de salaire (durée variable), l'assuré bascule en incapacité.

1.2.2. Incapacité

L'incapacité vient en relais du maintien de salaire, lorsque celui-ci est une garantie du contrat. Si l'assuré se retrouve dans un état d'incapacité, à cause d'une maladie ou d'un accident, le régime obligatoire de son assurance maladie assure le versement d'une partie de ses revenus d'activité. En général, les indemnités journalières correspondant au délai de carence sont déduites de cette prestation en espèces. La durée maximale de l'arrêt initial en incapacité est de 1 095 jours. Au terme de cette durée, l'assuré bascule en invalidité.

Dans le cas d'une couverture complémentaire souscrite auprès d'un organisme assureur, l'assuré perçoit des indemnités journalières indexées sur son salaire annuel de référence. Les prestations cumulées de la Sécurité sociale et de la Complémentaire ne doivent pas permettre l'enrichissement des assurés.

1.2.3. Invalidité

Lorsqu'une maladie ou un accident d'origine non professionnelle entraîne une réduction de la capacité de travail, l'assuré perçoit, grâce à son assurance maladie, une rente d'invalidité pour compenser la perte de salaire. Le montant est défini en fonction de la gravité de son invalidité et de son salaire de référence.

Dans le cas d'une couverture complémentaire souscrite auprès d'un organisme assureur, l'assuré perçoit une rente mensuelle ou trimestrielle indexée sur son salaire annuel de référence. Tout comme pour l'incapacité, les prestations cumulées de la Sécurité sociale et de la Complémentaire ne doivent pas permettre l'enrichissement des assurés.

2. Provisionnement

2.1. Cadre général du provisionnement en assurance non-vie

L'Article R. 331-6 du Code des Assurances [4] précise les différents types de provisions techniques à constituer par les organismes d'assurance.

- **La provision mathématique (PM) des rentes** est définie comme la « valeur actuelle des engagements de l'entreprise en ce qui concerne les rentes et accessoires de rentes mis à sa charge. » ;

- **La provision pour sinistres à payer (PSAP)** représente « la valeur estimative des dépenses en principal et en frais, tant internes qu'externes, nécessaires au règlement de tous les sinistres survenus et non payés, y compris les capitaux constitutifs des rentes non encore mises à la charge de l'entreprise. (…) ». Les PSAP se décomposent comme suit :
    - Les provisions RBNS (*Reported But Not Settled*) : il s'agit de l'estimation à dire d'expert pour les sinistres survenus et connus de l'assureur. Lorsqu'un sinistre est en cours de paiement et non clos, les provisions constituées sont les provisions RBNS ;

  - Les provisions IBNR (*Incurred But Not Reported*) : elles prennent en compte les sinistres dits tardifs. Ce sont les sinistres survenus mais non encore déclarés. Ils sont composés des :
    - IBNeR (*Incurred But Not Enough Reported*) : il s'agit des provisions complémentaires pour les dossiers ouverts pour lesquels l'estimation du coût total doit être revu à la hausse ;
    - IBNyR (*Incurred But Not Yet Reported*) : il s'agit des sinistres survenus mais dont l'assureur n'a pas encore connaissance ;

- **La provision pour risques croissants (PRC)** peut être exigée pour couvrir les risques de maladie et d'invalidité. Il s'agit de « la différence entre les valeurs actuelles des engagements respectivement pris par l'assureur et par les assurés » ;

- **La provision pour primes non acquises (PPNA)** est définie comme la « provision destinée à constater, pour l'ensemble des contrats en cours, la part des primes émises et des primes restant à émettre se rapportant à la période comprise entre la date de l'inventaire et la date de la prochaine échéance de prime ou, à défaut, du terme du contrat » ;

- **La provision pour risques en cours (PREC)** vient compléter la PPNA : elle permet de couvrir, « pour l'ensemble des contrats en cours, la charge des sinistres et des frais afférents aux contrats, pour la période s'écoulant entre la date de l'inventaire et la date de la première échéance de prime pouvant donner lieu à révision de la prime par l'assureur ou, à défaut, entre la date de l'inventaire et le terme du contrat, pour la part de ce coût qui n'est pas couverte par la provision pour primes non acquises » ;

- **La provision pour égalisation (PE)** permet de se prémunir face à d'éventuels pics de sinistralité ;

- **La provision pour risque d'exigibilité (PRE)** permet à l'organisme assureur d'honorer ses engagements en cas de moins-value de certains actifs [2]. On parle de moins-value latente de placements lorsque la valeur nette comptable de ces placements est supérieure à leur valeur globale.

2.2. Provisionnement des arrêts de travail

En arrêt de travail, un assuré peut faire l'objet de plusieurs risques (cf. section 1.2) ou états. Il est nécessaire d'établir les liens entre ces risques et le calcul des provisions associées à chacun d'entre eux. Pour les sections suivantes, le maintien de salaire est considéré comme une garantie présente au sein du contrat de prévoyance.

[2] Les placements concernés sont décrits dans l'article R. 332-20 du Code des Assurances.

2.2.1. *Différents états d'un assuré en arrêt de travail*

Les états possibles pour un assuré en arrêt de travail sont présentés dans le graphique ci-dessous :

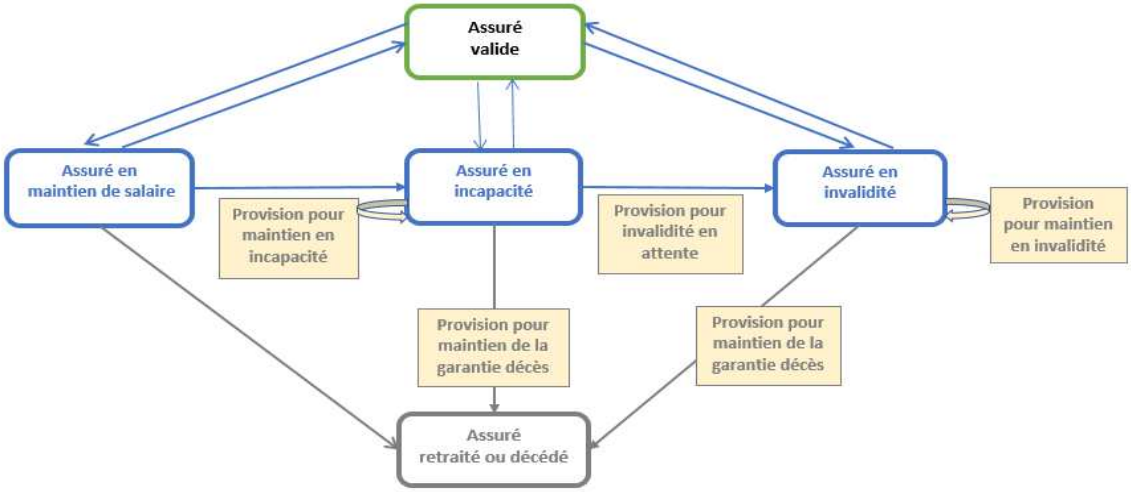


Figure 1 : Différents états possibles pour un assuré en arrêt de travail et provisions associées

Lorsqu'un arrêt de travail survient pour un assuré, il passe d'un état valide à un état d'assuré en maintien de salaire. Lorsque ces droits en maintien de salaire sont épuisés, c'est la garantie incapacité qui est activée. L'assuré passe alors en état d'incapacité. Au bout de 1 095 jours (ou 36 mois) depuis l'arrêt initial, son statut d'arrêt de travail évolue et l'assuré passe en état d'invalidité. L'assuré peut passer d'un état en arrêt de travail à un état valide et reprendre son activité professionnelle. Mais il peut également atteindre l'âge de départ en retraite en étant en arrêt de travail et arriver au terme de son contrat en prévoyance, ou décéder. Dans ces 2 derniers cas, les garanties d'arrêts de travail prennent fin.

En maintien de salaire, l'assuré peut :

- Passer en incapacité ;
- Retourner à un état valide ;
- Atteindre l'âge de départ en retraite ;
- Décéder.

En incapacité, l'assuré peut :

- Passer en invalidité ;
- Retourner à un état valide ;
- Atteindre l'âge de départ en retraite ;
- Décéder.

En invalidité, l'assuré peut :

- Retourner à un état valide (cas rare) ;
- Atteindre l'âge de départ en retraite ;
- Décéder.

Le cadre sur le fonctionnement en arrêt de travail étant posé, les différentes provisions calculées seront par la suite abordées en fonction de l'état d'arrêt de travail dans lequel se trouve l'assuré.

2.2.2. *Provisionnement pour le maintien de salaire*

Aucune provision dossier par dossier n'est calculée pour la garantie maintien de salaire, s'agissant d'un risque court. Dans ce cadre, les organismes assureurs calculent les provisions pour sinistres à payer

(PSAP) en utilisant les méthodes agrégées usuelles, comme la méthode de *Chain-Ladder* qui s'appuie sur les triangles de cadences de règlements.

2.2.3. Provisionnement pour l'incapacité

Comme requis par la réglementation, les organismes assureurs doivent calculer les provisions mathématiques pour chacun des sinistres connus en incapacité. La provision totale d'un sinistre se décompose en 3 provisions différentes.

Premièrement, le maintien en incapacité est provisionné. Pour cela, les organismes ont recours soit à la table réglementaire de maintien en incapacité du BCAC [3] [5], soit à leur table de maintien d'expérience certifiée en incapacité pour calculer la provision basée sur l'âge de l'assuré, son ancienneté dans le sinistre en incapacité et la durée restante maximale pour ce sinistre. Un assureur peut toutefois utiliser sa propre formule de calcul de provision mathématique, tout en conservant la comparabilité du résultat obtenu avec celui qui le serait avec la table réglementaire du BCAC.

L'assuré peut basculer en invalidité au terme de l'état en incapacité. A ce titre, l'invalidité en attente doit être prise en compte par les organismes assureurs. La provision associée se calcule en s'appuyant sur la table réglementaire d'invalidité du BCAC ou sur une table de maintien d'expérience en invalidité, l'âge de l'assuré et l'âge de départ à la retraite.

Le taux technique non-vie en vigueur est utilisé pour calculer les provisions de maintien en incapacité et d'invalidité en attente.

Enfin, la provision mathématique pour le maintien des garanties décès (MGDC) est déterminée. Parmi les garanties décès figurent :

- Le capital décès : le décès de l'assuré peut donner lieu au versement d'un capital. Le montant à verser est généralement déterminé en fonction de la rémunération de l'assuré décédé. Il varie selon l'âge de l'assuré au moment du décès et la situation familiale (majoration pour un enfant à charge par exemple) ;

- La rente conjoint : le décès de l'assuré peut conduire au versement d'une rente au conjoint survivant. Déterminée sur la base du dernier salaire de l'assuré décédé et en fonction des droits acquis auprès des organismes assureurs, cette rente peut être viagère ou temporaire, selon les termes du contrat ;

- La rente éducation : les enfants à charge de l'assuré peuvent bénéficier d'une rente d'éducation, calculée généralement sur la base du dernier salaire de l'assuré décédé. Cette rente peut être fixe, ou bien évoluer avec l'âge des enfants.

Le taux technique vie [4] en vigueur est utilisé pour le calcul de la provision de maintien des garanties décès.

En résumé, la provision mathématique au titre de l'incapacité se décline comme suit :

$$PM_{incapacité} = PM_{maintien\ en\ incapacité} + PM_{invalidité\ en\ attente} + PM_{MGDC}$$

[3] Bureau Commun d'Assurances Collectives

[4] Le taux technique vie est utilisé dans le cadre d'opérations d'assurance vie (rentes viagères, assurance décès, etc.). Il est au maximum de 60% de la moyenne des 6 derniers mois du Taux Moyen d'Emprunt (TME).

2.2.4. Provisionnement pour l'invalidité

La table de maintien en invalidité du BCAC, ou les tables de maintien d'expérience en invalidité, sont utilisées par les organismes assureurs pour calculer la provision en invalidité en cours, à laquelle s'ajoute celle du maintien des garanties décès.

La provision mathématique en invalidité permet le versement d'une rente jusqu'au départ à la retraite de l'assuré, ou jusqu'à son décès, si celui-ci intervient avant le départ à la retraite.

Les garanties décès sont identiques à celles décrites dans le cadre du calcul de la provision mathématique en incapacité (cf. section 2.2.3).

En résumé, la provision mathématique au titre de l'invalidité se décline comme suit :

$$PM_{invalidité} = PM_{maintien\ en\ invalidité} + PM_{MGDC}$$

3. Problématique du mémoire

Le calcul des provisions en arrêts de travail, tel que présenté dans le cadre des dispositions réglementaires pour le maintien de salaire, l'incapacité et l'invalidité, permet de répondre généralement à la problématique du provisionnement pour les arrêts de travail dans le cas où il n'y a pas une sinistralité élevée. Toutefois, lorsque le portefeuille de l'organisme assureur présente des spécificités (augmentation continue de la sinistralité, population sous risque avec des caractéristiques particulières, etc.), il est pertinent d'explorer des méthodes alternatives pour le calcul des provisions.

Dans ce mémoire, est proposée la mise en œuvre de méthodes de provisionnement en apprentissage automatique pour les dossiers d'arrêts de travail connus, y compris pour les sinistres en maintien de salaire. Ces méthodes permettront d'estimer la charge ultime des sinistres en prenant en compte les caractéristiques propres à chacun d'eux. Lorsqu'un sinistre est clos, la charge ultime correspond à la somme des règlements effectués au titre du dossier. Lorsque le sinistre est toujours ouvert, ces méthodes permettront d'estimer la charge de sinistre restante. Il s'agit des provisions RBNS pour les sinistres en maintien de salaire, incapacité et invalidité, et des provisions mathématiques requises pour les sinistres en incapacité et invalidité.

Les méthodes proposées ne permettent pas de prendre en compte les sinistres survenus mais non encore connus de l'assureur, c'est-à-dire les sinistres tardifs ou IBNR qui pourront faire l'objet d'une future étude.

L'objectif dans ce mémoire est de se focaliser sur le provisionnement pour les sinistres d'arrêts de travail connus et en cours de paiement. Ce mémoire traitant exclusivement des garanties d'arrêts de travail, le provisionnement du maintien des garanties décès n'y sera pas présenté.

Partie 2 – Cadre méthodologique

1. Méthodes d'apprentissage automatique pour le provisionnement

L'approche classique pour déterminer le provisionnement des arrêts de travail est basée sur les triangles *Chain Ladder* pour les PSAP et le calcul paramétrique pour les provisions mathématiques.

Ces dernières décennies, de nouvelles approches fondées sur l'apprentissage automatique ont émergé. Ces approches sont basées sur une stratégie *micro level* consistant en l'utilisation des données individuelles et de la richesse de l'information.

Les méthodes d'apprentissage ont apporté précision et robustesse aux modèles prédictifs, à travers les algorithmes de *machine learning* tels que les arbres de décision CART, ainsi que des méthodes plus élaborées telles que le *Random Forest* et le *Gradient Boosting*. C'est l'apport d'une meilleure appréhension du risque que nous souhaitons mesurer à travers leur implémentation dans le cadre du provisionnement, et plus spécifiquement dans celui des arrêts de travail.

2. Généralités sur l'apprentissage automatique

Le domaine de l'apprentissage automatique ou *machine learning* est une catégorie d'algorithmes qui permet de détecter les informations pertinentes pour prédire de manière plus précise des résultats, sans avoir été explicitement programmés. Les algorithmes tirent des enseignements du traitement des données par itération.

2.1. Types d'apprentissage

Il existe plusieurs types d'apprentissage dans le domaine du *machine learning* :

- **L'apprentissage supervisé**
  Pour prédire des événements futurs, les algorithmes supervisés utilisent ce qui s'est produit dans le passé, en utilisant des données labellisées. Avec les données X en entrée, ils peuvent prédire la variable réponse $Y = f(X)$.
  A partir d'algorithmes supervisés, il est possible de résoudre des problèmes de classification (lorsque la variable réponse est qualitative) ou de régression (lorsque la variable réponse est quantitative et continue) ;

- **L'apprentissage non supervisé**
  Les algorithmes disposent de données non labellisées en entrée, et il n'existe pas de variable réponse. La machine crée alors ses propres réponses, en réalisant des déductions à partir des données disponibles. Ces algorithmes permettent de résoudre des problématiques de regroupement ou *clustering* (création de groupes ayant des caractéristiques similaires en les assignant en grappes) ou d'association (détection de relations intéressantes, voire cachées, entre des variables d'une base de données importante) ;

- **L'apprentissage semi-supervisé**
  Les algorithmes de ce type d'apprentissage utilisent à la fois des données labellisées et non labellisées ;

- **L'apprentissage par renforcement**
  Cette méthode d'apprentissage interagit avec son environnement, pour déterminer le comportement idéal, à travers des actions qui permettent de découvrir des erreurs ou des avantages.

A partir des différentes descriptions présentées, l'apprentissage supervisé peut être considéré comme étant adapté à la mise en œuvre de la prédiction du provisionnement des arrêts de travail (prédiction d'une variable réponse correspondant à la charge ultime des sinistres).

Dans la section suivante, les arbres de décision, qui font partie des méthodes d'apprentissage supervisé, seront exposés.

2.2. Arbres de décision

Un arbre de décision est un outil d'aide à la décision, basé sur une méthode de partitionnement récursif sur un échantillon de données pour obtenir un modèle à la fois explicatif et prédictif. Les principaux avantages d'un arbre de décision résident dans la simplicité d'interprétation de leurs résultats visualisés sous la forme d'un arbre. La structure de la donnée est représentée hiérarchiquement, sous la forme d'une suite de décisions ou de tests pour prédire un résultat. Les décisions sont prises à chaque nœud de l'arbre. Plusieurs types de nœuds existent pour un arbre de décision :

- Le nœud « racine » : il s'agit du premier nœud. Il se situe au sommet de l'arbre. C'est le nœud à partir duquel se trouve l'échantillon entier, avant les subdivisions au moment de la première question posée ;
- Les nœuds « internes » : ce sont les nœuds qui sont intermédiaires, et qui sont à la fois fils d'un nœud prédécesseur et père d'autres nœuds successeurs ;
- Les nœuds « terminaux » : encore appelés « feuilles », ils ne possèdent pas de nœuds fils.

Pour un échantillon de données, chaque observation est décrite par un ensemble de variables, pour lesquelles des décisions sont prises au niveau des nœuds de l'arbre, depuis le nœud racine et en se poursuivant au niveau des nœuds internes. Les résultats de la prédiction se trouvent au niveau des feuilles.

Voici ci-dessous, un exemple d'arbre de décision binaire :

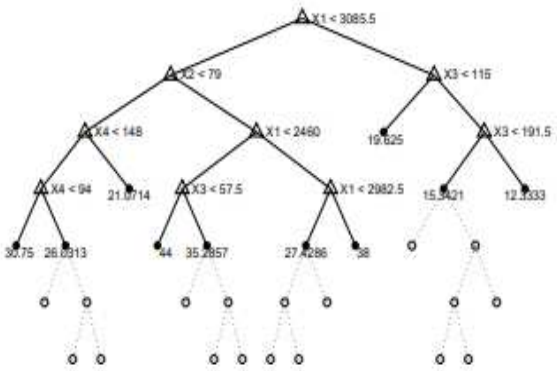


Figure 2 : Exemple d'arbre de décision binaire

Il existe 2 cas d'utilisation des arbres de décisions :

- **Les arbres de classification**
  Ils sont utilisés pour prédire une variable qualitative, et répartissent les individus d'un échantillon au sein de classes homogènes ;

- **Les arbres de régression**
  Ils permettent la prédiction d'une variable à expliquer quantitative et continue. Considérons Y, la variable quantitative à expliquer, et $\{X_1, \dots, X_j, \dots, X_p\}, 1 \leq j \leq$p, les p variables explicatives. Au sein d'un échantillon de n observations du $p+1$-uplet $\{Y, X_1, \dots, X_j, \dots, X_p\}$, l'exploration de l'ensemble de ces variables permettra d'une part de trouver une relation entre celles-ci pour chacune des observations, et d'autre part de distinguer les variables explicatives les plus pertinentes.

Pour prédire le provisionnement des arrêts de travail, l'utilisation d'arbres de régression paraît la plus appropriée (prédiction d'une variable réponse correspondant à la charge ultime des sinistres).

2.3. Sur-apprentissage et sous-apprentissage

Dans le cadre d'apprentissage supervisé, l'algorithme apprend des données passées pour pouvoir prédire les données nouvelles. L'objectif est donc de créer un modèle généralisable, pouvant être appliqué aux nouvelles données, n'ayant pas servi à l'apprentissage. En effet, la qualité du modèle se mesure à l'erreur de prédiction sur les données qui n'ont pas été utilisées lors de la phase d'apprentissage. Toutefois, ces algorithmes peuvent être confrontés à des problématiques de sur-apprentissage et de sous-apprentissage.

Il y a sur-apprentissage lorsqu'un algorithme, tel que construit, s'adapte strictement aux données de l'échantillon de la base d'apprentissage. Dans ce cas, l'algorithme n'a pas la capacité de prédire la variable réponse d'une observation différente de celle des données de l'apprentissage. Pour prévenir le sur-apprentissage, la qualité de prédiction de l'algorithme s'effectue sur un échantillon différent de celui de l'apprentissage, l'échantillon de test.

Il y a sous-apprentissage lorsque l'algorithme s'adapte mal aux données d'apprentissage. L'algorithme a des difficultés à détecter les interactions entre les différentes variables explicatives, contrairement au cas du sur-apprentissage qui capte « trop » toutes les interactions, y compris les bruits, des observations.

La conséquence du sur-apprentissage et du sous-apprentissage est un modèle peu performant. Il convient donc de pallier ces problématiques en créant un modèle équilibré. C'est l'objet de la section suivante.

2.4. Compromis biais-variance

Pour tester un modèle d'arbre de régression, l'erreur quadratique ou *Mean Squared Error* (MSE) est l'indicateur statistique principalement utilisé. Généralement, l'espérance de l'erreur quadratique d'un modèle est estimée de façon empirique sur l'échantillon de test.

Soit V, cet échantillon, de taille n. L'estimation de l'espérance de l'erreur quadratique s'exprime de la manière suivante :

$$\hat{E}(\hat{f}, V) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

avec y_i, la réalisation et $\hat{f}(x_i)$ la valeur prédite par le modèle.

Supposons $y = f(x) + e$, la fonction à modéliser. e est le bruit d'espérance nulle et de variance σ. On crée un estimateur $\hat{f}(\,)$ avec l'échantillon d'apprentissage, de sorte qu'une estimation $\hat{f}(x_i)$ soit disponible pour chaque point y_i de l'échantillon de test.

Ainsi, l'erreur quadratique peut s'écrire :

$$\begin{aligned} MSE &= E[(y - \hat{f}(x))^2] \\ &= E[y^2 - 2y\hat{f}(x) + \hat{f}^2(x)] \\ &= Var[y] + Var[\hat{f}(x)] + E[f(x) - \hat{f}(x)]^2 \end{aligned}$$

avec :

- $Var[\hat{f}(x)] = E[\hat{f}^2(x)] - E[\hat{f}(x)]^2$;
- $Biais = E[\hat{f}(x) - f(x)]$.

La décomposition du MSE est obtenue sous la forme :

$$MSE = Biais^2 + Variance + Erreur\ irréductible$$

La variance représente la sensibilité du modèle face aux fluctuations de l'échantillon d'apprentissage. Plus un modèle aura une variance importante, plus il changera si les données sont nouvelles. Les modèles à grande variance sont complexes. En effet, ils modélisent le bruit de l'échantillon d'apprentissage. De fait, ils ne sont pas applicables à d'autres échantillons. Il s'agit du sur-apprentissage.

Le biais représente l'erreur moyenne de $\hat{f}(x)$. Au sein d'un modèle, plus le biais est important, plus les modèles sont simples. Ils présentent un inconvénient majeur : ils ne détectent pas correctement les relations et interactions entre les variables explicatives et la variable de sortie. Il s'agit du sous-apprentissage.

Il est à préciser cependant que le MSE est utilisé dans le cas d'observations complètes.

2.5. Bases d'apprentissage, de test et de validation

Un échantillon est généralement découpé en 3 parties dans le cadre d'un apprentissage automatique :

- **L'échantillon d'apprentissage** : il permet de construire le modèle ;
- **L'échantillon de test** : la performance du modèle est mesurée grâce à cet échantillon. Il sert à calibrer le modèle ;
- **L'échantillon de validation** : il permet de mesurer la qualité prédictive du modèle, avec des données nouvelles.

Pour un modèle performant, l'objectif est de découper l'échantillon de manière à avoir le maximum de données possibles pour l'apprentissage, tout en conservant un nombre suffisant pour les échantillons de test et de validation.

La problématique de cette pratique est la perte de données en apprentissage. Ainsi, d'autres méthodes de rééchantillonnage peuvent être appliquées, telle que la validation croisée par exemple. Elle est présentée en annexe A.

2.6. Calibrage des modèles

Les algorithmes d'apprentissage automatique requièrent un étalonnage, qui se matérialise par la fixation de valeurs pour des paramètres importants, encore appelés hyperparamètres. C'est avec l'échantillon de test que s'effectue la mesure de la performance des algorithmes. Voici les étapes qui permettent ce calibrage :

- Etape 1 : déterminer la liste des valeurs possibles pour les paramètres du modèle ;
- Etape 2 : découper les bases de données, construire le modèle avec la base d'apprentissage et l'appliquer à la base de test ;
- Etape 3 : analyser les scores obtenus avec les différentes valeurs des paramètres ;
- Etape 4 : choisir les paramètres optimaux du modèle ;
- Etape 5 : mettre en œuvre le modèle final optimal sur la base de validation ;

Sur la base de l'article « *Tree-based censored regression with applications to insurance* », Lopez et al. (2015) [2] ont adapté l'algorithme CART *(Classification And Regression Tree)* pour tenir compte, non seulement des observations complètes, mais également des observations censurées.

Dans le chapitre suivant, l'algorithme CART est présenté ainsi que son adaptation en présence de données censurées.

3. Algorithme CART

3.1. Arbre de régression CART

L'algorithme CART proposé par Breiman et al. (1984) [6] est un arbre de décision exclusivement binaire. L'arbre binaire représente un ensemble de questions binaires qui vont segmenter les données en fonction de leurs particularités. L'arbre CART permet la prédiction de variables tant qualitatives (arbre de classification) que quantitatives (arbre de régression). C'est dans le cadre d'un arbre de régression qu'il sera utilisé pour nos travaux de provisionnement.

La mise en œuvre de l'algorithme CART s'effectue en 3 étapes :

- Construction de l'arbre maximal ;
- Élagage de l'arbre maximal ;
- Sélection de l'arbre optimal parmi les différents arbres élagués.

3.1.1. Construction de l'arbre maximal

La variable à expliquer dans le cadre d'une régression est une variable quantitative continue. Les démonstrations suivantes sont extraites de l'article « *Tree-based censored regression with applications to insurance* » de Lopez et al. (2015) [2].

La création de l'arbre maximal s'articule comme suit :

Soient :

- π_0, la quantité à prédire ;

- $\{Y, X_1, \dots, X_j, \dots, X_p\}$, le $p+1$-uplet contenant la variable réponse Y et les variables explicatives X_j, $1 \leq j \leq$p.

Tout au long de ce mémoire, le contexte général retenu est celui au sein duquel la quantité à prédire est une espérance :

$$\boldsymbol{\pi_0 = E[Y|X = x]}$$

Dans ce cas, la fonction de perte est l'erreur quadratique moyenne, ou *Mean Squared Error* (MSE), et est représentée par : $E[(\pi(x) - Y)^2]$.

L'espérance permet de minimiser l'erreur quadratique. Ainsi, la quantité à prédire sous la forme d'espérance, pour un échantillon constitué de n observations, est solution de l'équation :

$$\pi_n(x) = argmin\, E[(Y - \pi(x))^2 | X = x]$$

En pratique, l'espérance et la variance appliquées sont empiriques, avec :

- $E_n(Y) = \frac{1}{n}\sum_{i=1}^{n} y_i$, l'espérance empirique ;
- $V_n(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - E_n(Y))^2$, la variance empirique.

Or $Var(Y) = E[(Y - E(Y))^2]$. La variance empirique est alors choisie comme étant le critère de sélection de l'arbre.

Visualisons maintenant un arbre de décision. Pour expliquer Y, l'algorithme sélectionne la première variable explicative $X_j, 1 \leq j \leq p$, qui donne le meilleur découpage de l'échantillon de données en deux nœuds, en réduisant l'hétérogénéité des nœuds. En effet, chaque nœud N de l'arbre doit contenir des observations de Y les plus homogènes possibles. La mesure de l'hétérogénéité des nœuds croit avec la dispersion des observations de Y. Au sein d'un arbre de régression, cette mesure est la variance :

$$V_N(Y) = \frac{1}{|N|}\sum_{j=1}^{N}(Y_j - \overline{Y_N})^2,$$

avec $\overline{Y_N}$, la moyenne des $|N|$observations Y_j dans le nœud. Au nœud N, la variable sélectionnée sera celle qui permettra de minimiser l'hétérogénéité au sein des nœuds fils. C'est ainsi que l'algorithme CART permet également de distinguer les variables discriminantes dans la prédiction de la variable réponse. A chaque construction de nœud, le nouvel estimateur de $E[Y]$ est l'espérance empirique des observations du nœud concerné. L'algorithme recommence cette opération jusqu'à ce qu'il n'y ait plus qu'un seul individu ou que le critère de fin de génération de l'arbre soit atteint. Ce critère peut être la prise en compte d'un nombre minimal d'individus au sein d'un nœud. Les nœuds finaux sont les feuilles de l'arbre. L'arbre maximal est ainsi construit.

3.1.2. Elagage avec la fonction coût-complexité

L'élagage consiste à créer de nouveaux sous-arbres, permettant la prédiction de la variable d'intérêt, tout en retirant les branches (lien entre 2 nœuds père et fils) qui entrainent le sur-apprentissage ou le sous-apprentissage. Autrement dit, les branches sont retirées lorsque leur suppression n'impacte pas significativement la mesure de l'hétérogénéité, en veillant à conserver un nombre suffisant de données.

Pour construire les sous-arbres, il faut se référer à la fonction coût-complexité. Cette fonction estime le nombre d'opérations élémentaires effectuées par un algorithme (coût) et mesure sa performance en termes d'utilisation de la mémoire et de la vitesse d'exécution (complexité).

Soit T_{max}, l'arbre CART maximal. Considérons un sous-arbre T, comprenant $|T|$ feuilles.

Pour chaque valeur de α, avec $\alpha \in [0; +\infty[$, il existe un arbre $T \subset T_{max}$ qui minimise la fonction coût-complexité décrite par la relation suivante :

$$R_\alpha(T) = E\left[\left(Y - \pi(x)\right)^2 \middle| X = x\right] + \alpha . |T|$$

Soit $T(\alpha)$, le sous-arbre qui minimise la fonction coût-complexité en α. C'est ainsi que se produit la construction de plusieurs sous-arbres qui minimisent la fonction coût-complexité pour une valeur α donnée.

La dernière étape de l'algorithme CART est la sélection du meilleur arbre optimal.

3.1.3. Sélection du meilleur arbre optimal

A la suite de l'élagage, plusieurs sous-arbres optimaux ont été construits.

La méthode privilégiée dans ce mémoire, et très majoritairement utilisée en pratique, consiste à utiliser un échantillon de test différent de celui d'apprentissage de l'algorithme pour vérifier la qualité de la prédiction, avec des indicateurs de performance présentés ci-après.

3.1.4. Mesure de la qualité de prédiction

La robustesse de l'algorithme se mesure avec des indicateurs statistiques sur une base de données de test. Après avoir construit le modèle avec les données d'apprentissage, la base de test servira à mesurer la performance du modèle construit, à travers des indicateurs statistiques de performance. Dans cette section, les différents indicateurs adaptés aux arbres de régression, et auxquels il convient de recourir, seront exposés.

Le ***Mean Squared Error*** **(MSE)**, ou erreur quadratique moyenne, est la moyenne arithmétique des carrés des écarts entre les prévisions du modèle et les observations. C'est la valeur à minimiser afin de s'assurer de la robustesse et de la précision d'un algorithme.

On note :

- n, le nombre d'observations de l'échantillon ;
- y_i, la charge ultime réelle du sinistre i ;
- $\hat{y}_i$, la charge ultime estimée du sinistre i.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Le ***Root Mean Squared Error*** **(RMSE)** est un indicateur d'écarts qui est la racine carrée du MSE. Plus il est élevé, moins la prédiction est fiable. Il est également adapté à l'arbre de régression CART.

$$RMSE = \sqrt{MSE}$$

Le ***Mean Absolute Error*** **(MAE)** ou erreur absolue moyenne peut aussi être utilisé pour l'arbre de régression CART. Plus il est élevé, moins la prédiction est fiable.

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

Lorsque l'amplitude de la variable à prédire est importante, comme la charge ultime d'un sinistre par exemple, le **coefficient de détermination linéaire de Pearson**, noté **R²**, est adapté pour mesurer la qualité d'une prédiction. Il est défini par la relation suivante :

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

Plus il est élevé, plus la prédiction est robuste. Il n'est autre que le complémentaire de l'erreur carrée relative, ***Relative Squared Error*** ou **RSE.** Un RSE élevé est synonyme d'une prédiction peu fiable.

$$RSE = 1 - R^2$$

Dans le cadre de cette étude, une importance sera attachée à ces 2 derniers indicateurs de performance, R^2 et RSE, jugés plus adaptés à la prédiction de la charge ultime d'un sinistre.

3.2. Méthode d'apprentissage en présence de données censurées

Lorsqu'un sinistre est clos, les algorithmes disposent de la valeur exacte de la charge ultime à prédire. Dans ce cas, aucune provision n'est à constituer au titre du sinistre.

En revanche, pour un sinistre toujours ouvert à la date d'arrêté comptable (cas d'un sinistre censuré), la charge de sinistres disponible ne représente pas la charge ultime, mais une charge de sinistres moindre par rapport à l'ensemble des prestations attendues au titre de ce sinistre.

Pour corriger le biais induit sur la charge ultime par la censure des données, une approche est proposée par Vock et al. (2016) [7] pour prendre en compte les données censurées en utilisant la probabilité inverse de la pondération de la censure ou *Inverse Probability of Censoring Weighting* (IPCW). Cette méthode permet de calibrer le modèle en attribuant un poids à chacune des observations. Le principal avantage de cette méthode est de pouvoir être intégrée à de nombreux algorithmes d'apprentissage lorsque les données observées sont incomplètes.

Lopez et al. (2015) [2] proposent également une pondération des observations pour une meilleure qualité de prédiction dans le cadre du provisionnement des sinistres en assurance non-vie, en adaptant l'algorithme CART aux observations censurées.

3.2.1. Censure et formalisme

Dans cette section, l'approche de la censure proposée par Lopez et al. (2015) [2] est reprise.

Considérons le vecteur aléatoire (M, T, X). $M \in \mathbb{R}$ est la variable aléatoire d'intérêt représentant la charge de sinistre. $T \in \mathbb{R}^+$est la variable aléatoire représentant la durée, depuis la date de survenance jusqu'à la date de clôture du sinistre. X est le vecteur des variables explicatives de la variable M, et/ou de la variable T.

La censure est introduite par la variable $C \in \mathbb{R}^+$. En présence de données censurées, les données M et T ne sont pas connues. Mais on observe $Y = \min(T, C)$, et $N = \delta$M, avec $\delta = 1_{T \leq C}$.

L'individu i $(1 \leq \mathrm{i} \leq \mathrm{n})$ présent dans l'échantillon est ainsi représenté par $(N_i, Y_i, \delta_i, X_i)$. L'échantillon est composé de réplications indépendantes et identiquement distribuées de $(N_i, Y_i, \delta_i, X_i)_{1 \leq \mathrm{i} \leq \mathrm{n}}$.

L'objectif, en mettant en œuvre une méthode d'apprentissage automatique dans ces conditions, est de détecter l'importance des variables X et de la durée non censurée T sur la variable d'intérêt M. Pour ce faire, la méthode IPCW est présentée dans la section suivante pour faire le lien entre ces différentes variables.

3.2.2. Méthode IPCW - Inverse Probability of Censoring Weighting

Pour chaque observation i dans un échantillon de n individus, le poids $w_{i,n}$ est associé et correspond à l'inverse de l'estimation de la probabilité d'être non censuré $G(t) = P(C \leq t)$, à condition que l'observation i ait une durée de vie de t.

Etant donné que M et T ne peuvent être observés entièrement dans le cas de données censurées, un estimateur alternatif prenant en compte la censure doit être considéré pour ne pas introduire un biais à la prédiction.

Lopez et al. (2015) [2] retiennent alors les hypothèses suivantes :

- C est indépendante de (M, T) ;
- $P[T \leq C|M, T, X] = P[T \leq C|T]$.

A partir de ces hypothèses, pour toute fonction $\psi \in L^1$, on obtient :

$$E\left[\frac{\delta\psi(N,Y,X)}{1-G(Y-)}\right] = E[\psi(M,T,X)],$$

avec $G(t) = P(C \leq t)$. Il s'agit bien d'appliquer ici la méthode IPCW.

La fonction $G(\,)$ est cependant inconnue. Compte tenu des hypothèses retenues, Lopez et al. (2015) [2] proposent d'estimer cette fonction avec l'estimateur de survie de Kaplan-Meier. Il permet de déduire la fonction de survie à partir des durées de vie. Il sera utilisé pour estimer la fonction de répartition de la durée de vie des sinistres, depuis la date de survenance jusqu'à la date de clôture (date censurée ou non) du sinistre.

La section suivante expose l'estimateur de Kaplan-Meier ainsi que ses avantages.

3.2.3. Estimateur de Kaplan Meier et modèle de survie

L'estimateur de Kaplan-Meier (Kaplan et Meier (1958) [8]), connu également sous le nom de l'estimateur produit-limite, permet de simuler la fonction de survie à partir de données de durée de vie.

Il présente 3 avantages dans le cadre de ce mémoire :

- Il s'agit d'un estimateur non paramétrique. Il n'y a pas d'hypothèse forte à formuler autour de la loi de distribution, comme c'est le cas pour un estimateur paramétrique[5] ou semi-paramétrique [6] ;
- Il permet de mettre en évidence les spécificités du portefeuille étudié ;
- Il tient compte des censures et troncatures.

La section suivante s'inspire des travaux sur l'introduction à l'analyse des durées de survie de Saint Pierre (2015) [18].

Considérons les éléments suivants :

- α, la fonction de survie déterminée juste avant la date x_1 ;
- n, le nombre d'individus dans le cadre d'une étude ;
- x_1, la première date à laquelle des décès sont constatés ;
- r_1, le nombre d'individus présents à la date x_1 ;

[5] Parmi les estimateurs paramétriques, on distingue entre autres les modèles de risque instantané constant et monotone.

[6] Parmi les estimateurs semi-paramétriques, on peut distinguer le modèle de Cox et les modèles à hasards proportionnels.

- s_1, le nombre d'individus décédés à la date x_1.

$\hat{p}_{x_1} = \frac{s_1}{r_1}$ correspond à la proportion de décès à la date x_1. La fonction de survie conditionnelle au temps de survie x_1 peut être déduite, et exprimée par $1 - \frac{s_1}{r_1}$.

Plus explicitement, la fonction de survie à la date x_1 est égale à $\alpha \times \left(1 - \frac{s_1}{r_1}\right)$.

Par analogie, la fonction de survie à la date x_2 est égale à $\alpha \times \left(1 - \frac{s_1}{r_1}\right) \times \left(1 - \frac{s_2}{r_2}\right)$, avec $x_2 > x_1$.

Plus généralement, pour les décès qui surviennent aux dates x_i, $i \geq 1$, on observe s_i décès pour une population de r_i individus.

A la date $t < x_1$, la fonction de survie $S_n(t) = 1$.

Si on se place à la date du $i^{\text{ème}}$ événement, $S_n(x_i)$ est la fonction de survie et son expression est la suivante :

$$S_n(x_i) = S_n(x_{i-1}) \times \left(1 - \frac{s_i}{r_i}\right)$$

A la date t comprise entre 2 dates, c'est-à-dire $t \in [x_{i-1}; x_i]$, la fonction de survie peut être définie comme suit :

$$S_n(t) = \prod_{k=1}^{i-1} \left(1 - \frac{s_k}{r_k}\right)$$

Dans le cadre du mémoire et de manière pratique, l'estimateur de Kaplan-Meier servira à estimer la fonction de répartition de la durée de vie des sinistres. Pour ce faire, on considère que les individus sont les sinistres d'arrêts de travail. Les dates de naissance correspondent aux dates de survenance de l'arrêt de travail et le décès est matérialisé par la date de fin d'indemnisation du sinistre.

3.3. Adaptation de l'algorithme CART en présence de données censurées

Considérons à nouveau la relation suivante, abordée dans la section 3.2.2. :

$$E\left[\frac{\delta\psi(N,Y,X)}{1-G(Y^-)}\right] = E[\psi(M,T,X)]$$

En partant d'un échantillon de taille n, on estime $E[\psi(M,T,X)]$ empiriquement avec l'estimateur $\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \psi(N_i,Y_i,X_i)}{1-G(Y_i^-)}$.

Soit $\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \psi(N_i,Y_i,X_i)}{1-G(Y_i^-)} = \sum_{i=1}^{n} w_{i,n}\psi(N_i,Y_i,X_i)$, avec $w_{i,n} = \frac{1}{n} \times \frac{\delta_i}{1-G(Y_i^-)}$ qui n'est autre que le poids issu de la méthode IPCW.

Dans la pratique, les poids IPCW sont attribués aux sinistres clos. Ils sont déterminés en calculant la fonction de survie pour chaque sinistre observé, avec l'estimateur de Kaplan-Meier qui tient compte des sinistres non clos. Le poids correspondant aux sinistres censurés est égal à 0 et celui des sinistres clos est égal à $\frac{1}{n} \times \frac{1}{(1-G(Y_i^-)}$. Enfin, le poids est d'autant plus important que y_i est grand. On en déduit également que la somme des poids est égale à 1.

3.4. Qualité de prédiction en présence de données censurées

En section 3.1.4., les indicateurs R^2 et RSE ont été retenus pour la mesure de la qualité de prédiction de charges de sinistres. Nous nous attardons sur ces 2 indicateurs dans le cadre de données censurées. Le ***Weighted Relative Squared Error*** **(WRSE)** est adapté à ce cas de figure :

$$WRSE = \frac{\sum_{i=1}^{n} \omega_i \times (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} \omega_i \times (y_i - \bar{y})^2}, \text{ avec } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Son complémentaire, **le coefficient de détermination pondéré** R_w^2 s'obtient facilement à partir du $WRSE$:

$$R_w^2 = 1 - WRSE$$

Ce dernier indicateur est facilement interprétable, car il indique la réelle corrélation entre les valeurs prédites et les valeurs réelles. La prédiction est de meilleure qualité lorsque l'indicateur R_w^2 est élevé.

3.5. Limites de l'algorithme CART

L'algorithme CART présente quelques inconvénients :

- Chaque segmentation de l'arbre s'opère sur une seule variable, et non pas une combinaison de variables explicatives ;
- L'ordre des variables impacte le pouvoir de prédiction des modèles ;
- De légères variations de données peuvent conduire à des arbres différents ;
- La sélection de l'arbre optimal n'est pas garantie.

Après avoir exposé l'adaptation de l'algorithme CART aux observations non complètes, nous souhaitons prolonger l'exploration en tenant compte des méthodes d'agrégation telles que le *bagging* et le *boosting*. L'objectif est de mesurer leur apport dans le cadre du provisionnement des arrêts de travail en présence de données censurées. Le chapitre suivant présente ces méthodes d'agrégation.

4. Considération des méthodes d'agrégation

Les méthodes d'agrégation ou méthodes d'ensemble présentent l'avantage d'assembler les résultats d'algorithmes dans le but d'obtenir une amélioration de l'ajustement par la combinaison d'un grand nombre de modèles, tout en évitant le sur-apprentissage. Il existe plusieurs stratégies d'agrégation :

- Les stratégies aléatoires, comme le *bootstrap* et le *bagging*, et les algorithmes *Random Forest* (ou forêt aléatoire) ;
- La stratégie adaptative, comme le *boosting*, une construction adaptative d'une famille de modèles.

Lorsqu'on se trouve dans des cas de modèles potentiellement instables, tels que les arbres CART, les stratégies de *bagging* et de *boosting* prennent tout leur sens, en apportant de la précision aux prédictions.

Les stratégies de *bootstrap*, *bagging* et *boosting* sont décrites ci-après, ainsi que les algorithmes retenus *Random Forest* et *Gradient Boosting*, adaptés ensuite aux données censurées.

4.1. *Bootstrap*, *bagging* et algorithme *Random Forest*

4.1.1. Principe du bootstrap

Le *bootstrap* est une méthode statistique de rééchantillonnage d'une base de données initiale composée de n individus. Il consiste à créer de nouveaux échantillons à partir de la base de données initiale, en procédant à des tirages aléatoires avec remise de n individus afin d'évaluer la précision d'une méthode : estimer une erreur de prédiction, une variance, etc.

Soit $X = (X_1, X_2, \dots, X_n)$, un échantillon initial dont la fonction de répartition $F(x) = P[X \leq x]$ est inconnue.

On distingue :

- Le *bootstrap* paramétrique : la loi de F est inconnue ;
- Le *boostrap* non paramétrique : la loi de F est connue, mais le paramètre est inconnu.

Pour toute fonction de répartition de la forme $\theta(F) = \int h(x)dF(x)$, on utilise l'approximation $\theta(\hat{F}) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$.

La loi de $\hat{F}$ est déterminée avec des simulations de Monte Carlo [7], en réalisant des tirages avec remises dans l'échantillon X.

4.1.2. Principe du bagging

Le boostrap aggregating, encore appelé *bagging,* est une technique d'agrégation de modèles utilisée pour améliorer la prédiction des arbres de décision.

De nombreux modèles peuvent être agrégés dans le but de réduire la variance de l'estimateur pour corriger l'instabilité des arbres de décision (Breiman (2001) [9]). En effet, une petite modification au sein des données peut complètement changer les critères de construction de l'arbre. Le principe du *bootstrap* est de créer de nouveaux échantillons par tirage aléatoire dans l'ancien échantillon, avec remise. L'algorithme est testé sur ces sous-échantillons de données.

Considérons la charge ultime d'un sinistre comme la variable quantitative à prédire. Une explication du fonctionnement du *bagging* dans le cadre d'une régression est décrite ci-dessous. Soient :

- Y, la variable quantitative à expliquer ;
- $(X_1, X_2, \dots, X_k)$, les covariables, avec $x = (x_1, x_2, \dots, x_k) \in \mathbb{R}^k \; et \; k < n$;
- $\emptyset(x)$ représente le modèle et $\widehat{\emptyset}(x)$ en est son estimateur ;
- $z = \{(x_1, y_1), \dots, (x_n, y_n)\}$, l'échantillon observé.

On considère m échantillons indépendants, notés $z_m, m \in [1, M]$. Un modèle agrégé est ensuite construit :

$$\widehat{\emptyset}_M(\,) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\emptyset}_m(\,)$$

Il s'agit du principe de moyennisation utilisé dans le *bagging*. Une moyenne est calculée pour les prévisions afin de réduire la variance, et donc l'erreur des prédictions sur plusieurs modèles indépendants.

[7] Les simulations de Monte Carlo représentent des algorithmes utilisés dans l'objectif d'estimer la probabilité d'occurrence d'un scénario au sein duquel les paramètres sont aléatoires. Cette technique statistique permet de comprendre l'influence de l'incertitude dans des modèles de prédiction, notamment dans le domaine de la finance.

En comparaison avec CART, le *bagging* présente l'avantage d'être moins impacté par le sur-apprentissage. En effet, les arbres individuels ne sont pas élagués. Et dans le cas où il y aurait sur-apprentissage, la variance spécifique de l'arbre sera élevée, mais présentera un biais faible.

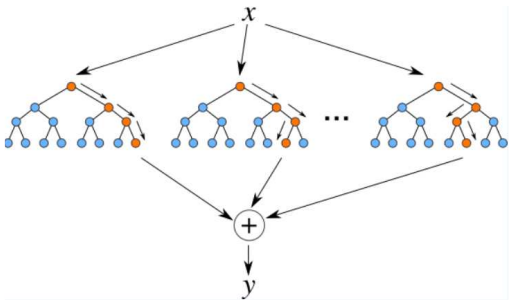


Figure 3 : *Bagging* – Principe de fonctionnement de la prédiction

4.1.3. Algorithme Random Forest

4.1.3.1. Principe

Les forêts aléatoires ou *Random Forest* sont des techniques de *machine learning* proposées par Breiman (2001) [9]. Ces techniques permettent de pallier le problème généré par l'ordre des variables explicatives pour les arbres de décisions. Leur but est de construire un certain nombre d'arbres partiellement indépendants pour lesquels l'apprentissage s'effectue sur des sous-ensembles de données légèrement différents. Toujours dans le cas des arbres de régression, chaque arbre va réaliser ses prédictions, puis une moyennisation de l'ensemble des prédictions sera effectuée. Le résultat est donc la combinaison des sous-espaces aléatoires et de la technique du *bagging*.

Notons toutefois que pour les arbres de classification, l'attribution d'une classe se fera par « vote » à partir des prédictions de chaque arbre. En effet, un arbre donné va désigner (ou voter pour) la classe qui définit le mieux un objet à classer. La forêt aléatoire agrège ensuite l'ensemble des prédictions pour donner la classification de l'objet.

Dans l'objectif de prédire une charge ultime de sinistres, on se focalise sur le fonctionnement des arbres de régression. Pour créer chaque arbre, on spécifie un nombre m de variables parmi les M variables d'entrée ($m << M$) pour qu'elles soient sélectionnées à chaque nœud, afin d'obtenir la meilleure séparation sur ces variables. On ne réalise pas d'élagage sur chaque arbre construit. Il est d'usage de fixer $m = \frac{M}{3}$ dans le cadre d'arbres de régression. Une autre approche invite à tester la moitié puis le double de cette valeur pour tenter d'améliorer les prédictions.

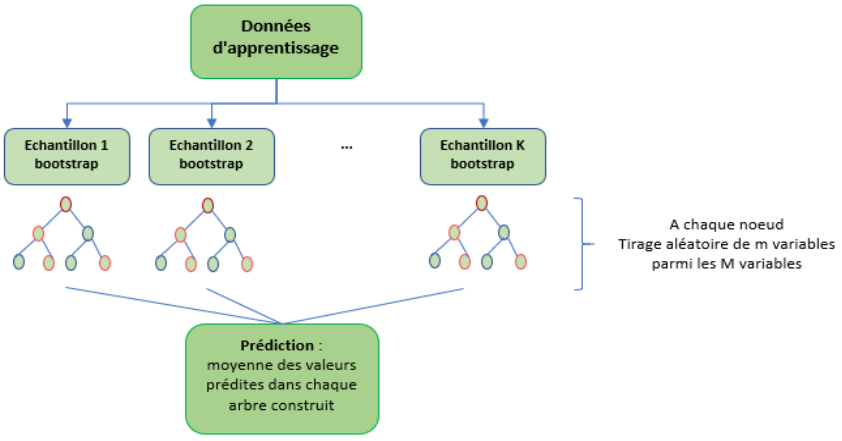


Figure 4 : *Random Forest* – Principe de construction de l'algorithme

Les algorithmes *Random Forest* présentent de nombreux avantages. La précision est meilleure que les algorithmes CART. Leur mise en œuvre peut s'effectuer sur une base de données volumineuse, en prenant également en compte un nombre important de variables d'entrée. Les forêts aléatoires donnent accès aux variables discriminantes dans la régression. Au fur et à mesure qu'elle se construit, la forêt aléatoire génère un estimateur non biaisé de l'erreur de généralisation. De plus, elle prend en charge les données manquantes, et peut se généraliser sur d'autres échantillons de données.

4.1.3.2. Erreur *Out Of Bag*

Chaque arbre de la forêt aléatoire est construit sur une fraction « *in bag* » de données, selon le principe de *bootstrap*. Cette fraction sert à l'apprentissage de l'algorithme. Les données non utilisées sont appelées les données *out of bag* ou OOB.

Considérons une observation (i, X_i, Y_i) de l'échantillon d'apprentissage. Pour estimer la prédiction $\widehat{Y}_i$, on calcule l'erreur obtenue par l'agrégation de tous les arbres n'ayant pas été appris avec cette observation. La somme des erreurs calculée ainsi sur l'ensemble de l'échantillon d'apprentissage représente l'erreur *out of bag*. Elle se définit par $\frac{1}{n}\sum_{i=1}^{n} 1_{\hat{Y}_i \neq Y_i}$. L'erreur *out of bag* n'utilise jamais les prédictions de la forêt elle-même, mais plutôt celles des prédictions qui sont des agrégations d'arbres de cette forêt. Elle sert à déterminer la pertinence des variables explicatives.

4.1.3.3. Importance des variables

Les forêts aléatoires permettent de mesurer l'importance des variables explicatives dans la prédiction.

Après avoir construit la forêt aléatoire, on calcule l'erreur *out of bag* associée. Pour calculer le score d'une variable explicative, la procédure est précisée ci-dessous :

- Lors de l'apprentissage, les valeurs de la variable explicative sont permutées aléatoirement parmi les observations ;
- A ce stade, on calcule une nouvelle fois l'erreur *out of bag*, puis on calcule la différence avec l'erreur *out of bag* associée à la forêt ;
- Les scores sont ensuite normalisés[8].

4.1.3.4. Limites des algorithmes *Random Forest*

Malgré les avantages pluriels des forêts aléatoires, elles présentent un inconvénient majeur, lié à leur sensibilité face au nombre de variables pour la construction des arbres individuels. Genuer et al. (2012) [10] expliquent que la variation de ce paramètre induit des indices d'importance différents au sein des variables retenues.

Gregorutti et al. (2014) [11] montrent également que la corrélation entre les variables explicatives peut impacter sensiblement la mesure de l'importance des variables.

[8] La normalisation consiste à soustraire la moyenne et à la diviser par l'écart-type. Dans ce cas, chaque valeur refléterait la distance par rapport à la moyenne en unités d'écart-type.

4.1.4. *Adaptation de l'algorithme Random Forest en présence de données censurées*

La méthode décrite dans cette section est proposée par Olympio (2019) [25] dans sa thèse « Contributions au provisionnement en assurance de personnes et à la gestion des risques ».

En présence de données censurées, la méthode IPCW est également appliquée aux forêts aléatoires. Les individus sont pondérés à l'aide de l'estimateur de Kaplan-Meier, en calculant l'estimateur $\hat{G}$. Dans ce cas, la sélection des individus dans la construction des arbres individuels se fera en appliquant le principe *bootstrap*.

Soit M, le nombre de modèles (ou arbres) à construire au sein de la forêt aléatoire.

- Pour $m = 1$ à M, un tirage aléatoire est effectué dans la base z d'un échantillon *bootstrap*, avec remise. L'échantillon issu de ce tirage est noté $z_m^{(b)}$;
- On estime l'arbre de régression optimal $\hat{f}_{z_m^{(b)}}$ en fonction des critères de l'algorithme « *Tree-based censored* » afin d'obtenir les M modèles ;
- Ensuite, on procède à l'agrégation des M modèles ;
- Enfin, on calcule l'estimateur qui est égal à la moyenne des prédictions de chaque modèle, en appliquant simplement la méthode du *bagging*.

La qualité de prédiction intègre également des indicateurs statistiques pondérés, tout comme pour l'arbre CART pondéré.

4.2. *Boosting* et algorithme *Gradient Boosting*

4.2.1. *Principe du boosting*

Une idée originale de Schapire (1990) [12] a conduit au *boosting*, une modélisation en apprentissage automatique. L'objectif était d'améliorer les performances d'un modèle de discrimination dont la probabilité de succès sur la prédiction d'une variable qualitative est supérieure à celle d'une variable aléatoire. Ce modèle s'appelle un « faible classifieur ». Quelques années plus tard, Freund et Schapire (1996) [13] ont affiné cet algorithme pour la prédiction d'une variable binaire. Viendront ensuite de nombreuses adaptations pour la prise en compte de *k* classes, de régression, etc. sur des jeux de données divers et variés. Les résultats ont démontré les performances de l'algorithme pour réduire sensiblement, non seulement la variance, mais aussi le biais de prévision, en comparaison avec d'autres algorithmes.

Le *boosting* est une méthode d'agrégation de modèles. Plutôt que de créer plusieurs arbres et de finaliser la prédiction uniquement lorsque tous les estimateurs auront été déterminés en calculant la moyenne ou en réalisant un vote (à l'image du *bagging*), le *boosting* travaille de manière séquentielle :

- Tout d'abord, l'algorithme construit un premier modèle qui va être évalué ;
- A ce stade, l'algorithme affecte à chaque individu du modèle un poids en fonction de la performance de la prédiction. Moins la prédiction est performante, plus le poids de l'individu est important ;
- Ces individus pour lesquels la valeur prédite est peu performante seront introduits dans la construction du modèle suivant ;
- Les poids sont donc corrigés au fur et à mesure, avec pour objectif de mieux prédire les valeurs les plus singulières ;

4.2.2. *Algorithme Gradient Boosting*

Parmi les modèles agrégés adaptatifs, Friedman (2002) [14] propose un algorithme, *Gradient Boosting Machine*, basé sur une fonction de perte supposée convexe et différentiable, notée l. Cette approche propose de chercher la meilleure combinaison linéaire d'arbres binaires en minimisant les résidus ou écarts entre la valeur prédite et la moyenne calculée.

La figure suivante résume ce fonctionnement :

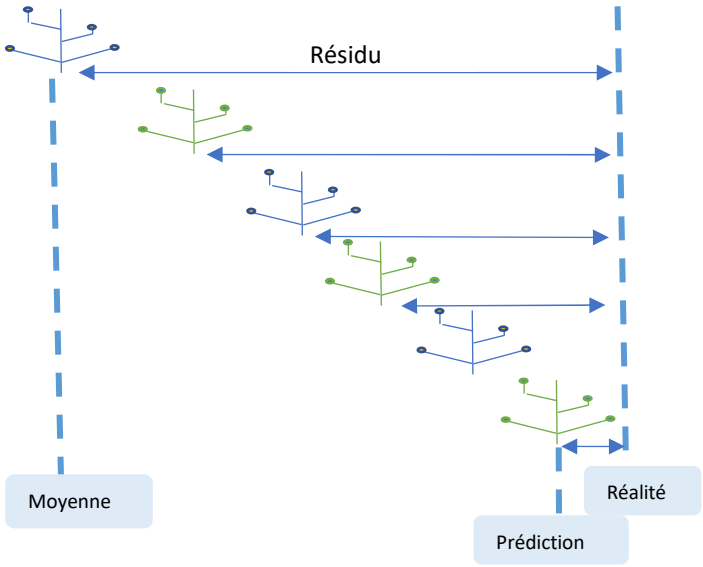


Figure 5 : *Gradient Boosting* – Principe de fonctionnement de l'algorithme

2 idées innovantes sont proposées, en complément du *boosting* :

- Pour améliorer les propriétés de convergence de l'algorithme, on considère le gradient de la fonction de perte. Le calcul du poids des individus s'effectue avec le gradient de la fonction de perte, dans l'optique d'optimiser les propriétés de convergence ;
- La descente du gradient est une méthode itérative qui permet d'approcher la solution optimale d'un problème. Pour se prémunir contre le sur-apprentissage, Le gradient est simulé par un arbre de régression.

La suite de cette section s'inspire de l'illustration de Rakotomalala (2016) [19] au sujet du *Gradient Boosting* pour un modèle de régression.

Cas d'un problème de régression

Considérons Y, une variable quantitative cible et X, un ensemble de variables quantitatives. Soit le problème de régression suivant :

$$Y_i = M_1(X_i) + \varepsilon_{1_i}$$

avec :

- (i, X_i, Y_i), une observation ;
- ε_1, les insuffisances du modèle ;
- M_1, un arbre de régression. Notons qu'il peut s'agir de tout type de méthode.

On note :

$$e_{i1} = Y_i - M_1(X_i)$$

où e représente les résidus. L'objectif est ensuite de modéliser ce résidu avec un second arbre de régression M_2 et de l'associer au modèle précédent, afin de réaliser une meilleure prédiction. Dans le cadre d'un modèle additif, on obtient :

$$e_{i1} = M_2(X_i) + \varepsilon_{2_i}$$

$$\hat{Y}_i = M_1(X_i) + M_2(X_i)$$

Cette combinaison d'arbres de régression a pour rôle de compenser les insuffisances de M_1. Ce processus peut être prolongé par l'ajout d'arbres de régression supplémentaires M_3, etc. Le modèle additif permet une illustration claire de la méthode. Notons qu'il peut s'agir d'un modèle différent.

Fonction de coût global et relation avec la descente du gradient

La somme des carrés des erreurs est un indicateur global de qualité des modèles de régression. On note :

- $l(Y_i - f(X_i)) = \frac{1}{2}(Y_i - f(X_i))^2$, la fonction de coût ;
- $L(Y, f) = \sum_{i=1}^{n} l\big(Y_i - f(X_i)\big)$, l'indicateur global de qualité.

Le calcul du gradient se décompose comme suit :

$$\nabla l\big(Y_i, f(X_i)\big) = \frac{\partial l(Y_i - f(X_i))}{\partial f(X_i)} = \frac{\partial[\frac{1}{2}(Y_i - f(X_i))^2]}{\partial f(X_i)} = f(X_i) - Y_i$$

On remarque ici que la valeur négative du gradient représente le résidu.

De fait, en revenant au problème de régression, la modélisation des résidus à l'étape $m, 1 \leq m \leq M$ de l'algorithme peut s'écrire plus généralement et de manière itérative :

$$\hat{f}_m = \hat{f}_{m-1} - \gamma_m \sum_{i=1}^{n} \nabla_{f_{m-1}} l\,(Y_i, f_{m-1}(X_i))$$

Le problème à résoudre revient à trouver un meilleur pas de descente γ avec :

$$min_\gamma \sum_{i=1}^{n} [l(Y_i, f_{m-1}(X_i) - \gamma \frac{\partial l\big(Y_i, f_{m-1}(X_i)\big)}{\partial\, f_{m-1}(X_i)})]$$

Généralement, *Gradient Boosting* utilise des arbres de régression CART pour sa mise en œuvre. Pour se prémunir contre le sur-apprentissage, une limite de la taille des arbres peut être utilisée lors du paramétrage de l'algorithme.

4.2.3. Limites des algorithmes Gradient Boosting

Malgré ses nombreux avantages, notamment sa flexibilité dans le choix des fonctions de coût et son adaptabilité à des jeux de données variés, l'algorithme *Gradient Boosting* présente toutefois des inconvénients :

- Une complexité non négligeable reste le nombre important de paramètres à calibrer (nombre d'itérations, taille des arbres, etc.) pour obtenir une réelle performance ;

- Il est également très couteux en termes d'occupation mémoire, car tous les arbres en déploiement en sont consommateurs. Plus le nombre d'arbres est important, plus les temps de calculs sont longs ;
- Enfin, le *Gradient Boosting* est menacé par le sur-apprentissage. Le choix d'arbres équilibrés est ainsi requis.

4.2.4. Adaptation de l'algorithme Gradient Boosting en présence de données censurées

La méthode décrite dans cette section est proposée par Olympio (2019) [25]. On note :

- M, le nombre d'arbres optimaux élagués prédéfinis ;
- n, le nombre d'individus composant un échantillon.

L'échantillon est représenté par $z = \{(N_i, Y_i, \delta_i, X_i)\}$, avec $i = 1, \dots, n$.

Plusieurs étapes sont à considérer pour la mise en œuvre de l'algorithme en présence de données censurées :

- Premièrement, les calculs de l'estimateur $\hat{G}$ et des poids IPCW avec l'estimateur de Kaplan-Meier pour les n individus sont réalisés ;
- Ensuite, la fonction de perte $\hat{f}_0 = argmin_\gamma \sum_{i=1}^{n} l(Y_i, \gamma)$ est estimée ;
- On calcule récursivement $r_i^m = -\left[\frac{\partial l(Y_i, f(X_i))}{\partial f(X_i)}\right]_{f=f_{m-1}}$, pour $i = 1, \dots, n \; et \; pour \; m = 1, \dots, M$;
- Puis, l'arbre de régression optimal élagué δ_m est ajusté au couple $(x_i, r_i^m)_{i=1,\dots,n}$ avec l'algorithme « *Tree-based censored* » ;
- γ_m est calculé par la suite, en résolvant l'équation $min_\gamma \sum_{i=1}^{n} l(y_i, f_{m-1}(x_i) - \gamma \delta_m(x_i))$;
- Puis, la mise à jour est effectuée pour $\hat{f}_m = \hat{f}_{m-1} - \gamma_m \delta_m(x)$;
- Enfin, la valeur finale $\hat{f}_M$ sera un estimateur du paramètre d'intérêt.

Les indicateurs statistiques permettant de mesurer la qualité de la prédiction sont également pondérés, selon le même procédé que celui de l'adaptation des algorithmes CART et *Random Forest*.

5. *Backtesting* et validation des algorithmes

Pour s'assurer de la performance, de la qualité et de la stabilité des modèles et des prédictions réalisées, le choix se porte dans un premier temps sur le *backtesting*. Cette pratique consiste à confronter les modèles à la réalité. La date d'observation des données est antérieure à la date à laquelle tous les sinistres sont clos. Le but de cette démarche est d'évaluer l'efficacité de la prédiction. Avec le *backtesting*, la stratégie à appliquer pour la période à venir est de supposer que ce qu'il s'est produit dans le passé se reproduira dans l'avenir. Cette hypothèse sera retenue comme prérequis pour une utilisation en situation réelle de prédiction de charge de sinistres, en partant du principe que les résultats issus du *backtesting* sont fiables pour les prédictions futures.

Habituellement, les échantillons de test servent également de validation des algorithmes. Cependant, l'objectif des modèles étant de prédire la charge ultime non connue au préalable des sinistres d'arrêts de travail, la simulation d'une mise en situation réelle est souhaitée. Ainsi dans un second temps, les modèles seront appliqués sur des données pour lesquelles tous les sinistres ne sont pas clos à la date d'inventaire et pour lesquels les règlements à l'ultime ne sont pas disponibles. Les résultats obtenus permettront de s'inscrire dans une démarche critique, non seulement sur les hypothèses retenues pour les cas de censure, mais aussi sur les enseignements transmis par les modèles.

Partie 3 – Périmètre de l'étude

1. Contexte

AG2R La Mondiale propose des accords de branches en santé et prévoyance pour près de 110 branches professionnelles. Adaptées à chaque convention collective nationale (CCN), les offres coconstruites avec les partenaires sociaux s'accompagnent d'actions ciblées spécifiques aux besoins de chaque secteur d'activité. Les ambitions qui animent les accords de branche sont de continuer à améliorer la protection sociale de chaque filière en proposant des garanties et des services toujours plus adaptés aux attentes de chaque profession et en renforçant les mécanismes de solidarité intergénérationnelle.

AG2R La Mondiale propose également des contrats de prévoyance performants pour les entreprises afin de sécuriser leurs salariés et leur famille tout en s'adaptant à leur environnement juridique. Ces contrats sur mesure permettent de répondre aux spécificités des entreprises : une couverture personnalisée leur est proposée, en fonction de leurs objectifs, leur statut, leur taille et leur structure. Ils bénéficient également d'un tarif préférentiel.

Parmi les contrats des branches du secteur interprofessionnel et les contrats sur mesure, l'un d'entre eux attire l'attention en termes d'arrêts de travail pour les raisons suivantes :

- Le taux d'absentéisme [9] atteint pour sa part 15% pour ce contrat contre 5,11% en moyenne nationale en 2019, selon le baromètre annuel sur l'absentéisme au travail du cabinet de conseil Ayming [15] ;
- En termes de chiffre d'affaires, ce contrat fait partie du top 10 des contrats des branches professionnelles et des entreprises. Son chiffre d'affaires représente plus de 3% du chiffre d'affaires total de ces contrats ;
- Depuis 2012, les garanties d'arrêts de travail – maintien de salaire, incapacité et invalidité – présentes dans son contrat de prévoyance collective n'ont pas évolué.

Dans ce mémoire, l'intérêt est porté au provisionnement « à la carte » que peuvent apporter les méthodes alternatives de provisionnement avec des méthodes d'apprentissage automatique pour ce contrat en particulier.

Avant d'explorer les données de ce contrat, les garanties d'arrêts de travail seront présentées. Le tableau ci-dessous décrit les différentes conditions, les franchises, les durées et les montants de prestations pour chacune des garanties. Le salaire de référence mentionné est égal au salaire brut (tranches A et B) soumis à cotisations et perçu par le salarié au cours des 12 derniers mois précédant l'arrêt de travail.

[9] Le taux d'absentéisme est le rapport du nombre d'heures d'absence sur le nombre d'heures de travail en théorie sur la période.

	Maintien de salaire	Incapacité	Invalidité
Conditions	Salarié ayant au moins 6 mois d'ancienneté dans la structure	- à l'issue de la période d'indemnisation du maintien de salaire - s'il s'agit d'un nouvel arrêt et que les droits du maintien de salaire sont épuisés : dans ce cas, une franchise de 3 jours s'applique.	Salarié ayant au moins 6 mois d'ancienneté dans la structure
Franchise	Indemnisation à partir du : - 4ème jour si la cause de l'arrêt est une maladie ou un accident de la vie privée - 1er jour si la cause de l'arrêt est un accident du travail ou une maladie professionnelle	Indemnisation après 30 jours d'arrêts de travail continus si l'ancienneté de 6 mois n'est pas atteinte	Aucune indemnisation si l'ancienneté de 6 mois n'est pas atteinte
Durée	- Pour une ancienneté de moins de 20 ans : 60 jours sur 12 mois consécutifs - Pour une ancienneté de plus de 20 ans : 90 jours sur 12 mois consécutifs	En tout état de cause, les prestations ne peuvent être versées au-delà du 1 095ème jour d'arrêt de travail (1 095 jours au total après l'arrêt initial)	Les prestations cessent si l'assuré passe à l'état valide, retraité ou décédé
Montant des prestations	90% de la 365ème partie du salaire de référence, (y compris les prestations brutes réelles ou reconstituées versées par la Sécurité sociale et l'éventuel salaire à temps partiel)	70% de la 365ème partie du salaire de référence, y compris les prestations brutes (réelles ou reconstituées de manière théorique)	- En cas d'invalidité 1ère catégorie : Versement d'une rente égale à 45 % du salaire de référence, y compris les prestations brutes de la Sécurité sociale - En cas d'un taux d'incapacité permanente professionnelle compris entre 33% (inclus) et 66% : Versement d'une rente égale à : (R x 3 N) / 2, avec R = rente invalidité 2ème catégorie N = taux d'incapacité permanente professionnelle - En cas d'invalidité 2ème, 3ème catégories ou taux d'incapacité permanente professionnelle supérieur ou égal à 66 % : Versement d'une rente égale à 75 % du salaire de référence, y compris les prestations brutes de la Sécurité sociale.

Tableau 1 : Garanties d'arrêts de travail pour le portefeuille de l'étude

2. Statistiques exploratoires

L'échantillon des arrêts de travail provient de 4 bases de gestion distinctes, avec une profondeur d'historique du 1er janvier 2012 au 31 décembre 2020 pour le contrat sélectionné.

Avertissement : à des fins de confidentialité, les résultats présentés dans ce mémoire ont été volontairement modifiés par l'application d'un coefficient multiplicateur. Ils restent toutefois comparables entre eux.

2.1. Bases de gestion

Toutes les extractions des bases de gestion s'effectuent à la date d'inventaire du 31 décembre 2020.

2.1.1. Base de gestion des prestations en arrêt de travail

Cette base de gestion permet l'accès aux règlements réalisés pour chaque sinistre d'arrêt de travail. Les variables présentes sont :

- Le numéro du sinistre ;
- Le numéro d'identification de l'assuré ;
- Le numéro du contrat de prévoyance ;
- Le risque (maintien de salaire, incapacité, invalidité) ;
- La date de survenance du sinistre (arrêt initial) ;
- La date de début d'indemnisation ;
- La date de fin d'indemnisation constatée à la date d'arrêté comptable ;
- Le montant de la charge de sinistre ;
- Le salaire trimestriel de référence de l'assuré ;
- La cause de l'arrêt de travail ;
- La date du règlement de la prestation ;
- La catégorie socio-professionnelle ;
- La franchise contractuelle appliquée.

2.1.2. Base de gestion des assurés

Cette base de gestion permet l'accès aux caractéristiques de chaque salarié assuré. Les variables présentes sont :

- Le numéro d'identification de l'assuré ;
- Le genre ;
- La date de naissance ;
- La date de décès ;
- Les dates de début et fin d'adhésion au contrat de prévoyance ;
- Le salaire trimestriel de référence.

2.1.3. Base de gestion des contrats de prévoyance

Cette base de gestion permet l'accès aux caractéristiques de chaque contrat de prévoyance. Les variables présentes sont :

- Le numéro de contrat ;

- Le numéro d'identité juridique ;
- La date d'adhésion ;
- La date de cessation ;
- Le numéro de SIREN.

2.1.4. Base de gestion des provisions mathématiques

Cette base de gestion permet l'accès aux provisions mathématiques calculées pour chaque sinistre connu et toujours ouvert pour les risques incapacité et invalidité au 31 décembre 2020. Les variables présentes sont :

- Le numéro de sinistre ;
- Le numéro d'identification de l'assuré ;
- Le genre de l'assuré ;
- La date de naissance de l'assuré ;
- La date de survenance du sinistre ;
- Les dates de début et fin d'adhésion au contrat de prévoyance ;
- Le taux technique non-vie appliqué.

Si le risque du sinistre est l'incapacité, les variables suivantes sont définies :

- Le montant des provisions pour le maintien en incapacité ;
- Le montant des provisions pour l'invalidité en attente.

Si le risque du sinistre est l'invalidité, les variables suivantes sont définies :

- Le montant de la provision pour maintien en invalidité ;
- La date théorique de fin de la garantie : elle représente la date de départ à la retraite de l'assuré.

Les tables du BCAC construites en 2013 sont utilisées pour calculer les provisions mathématiques en incapacité et en invalidité.

2.2. Création de la base de données des arrêts de travail et hypothèses préliminaires

Avec ces différentes bases de gestion, est créée une nouvelle base de données comprenant un sinistre par ligne, avec les valeurs clés que sont le numéro de sinistre, le risque et la date de survenance. Les dates de début et de fin d'indemnisation du sinistre sont calculées en prenant la valeur minimale de la date de début d'indemnisation et la valeur maximale de la date de fin d'indemnisation. Le montant des prestations est la somme des règlements effectués au titre du sinistre. Les caractéristiques des assurés, des contrats et des provisions dossier par dossier sont également intégrées à cette base de données.

2.2.1. Ajout de données calculées

Des données brutes disponibles permettent de calculer de nouvelles variables importantes dans le cadre de la modélisation des charges ultimes :

- **Age à la survenance** : l'âge à la survenance d'un sinistre est la différence de millésime entre la date de naissance de l'assuré et la date de survenance du sinistre.

- **Ancienneté au contrat** : l'ancienneté du salarié au sein du contrat à la survenance du sinistre est la différence de millésime entre sa date d'entrée dans l'entreprise et la date de survenance de l'arrêt.

- **Durée totale du sinistre** : la durée totale du sinistre est le nombre de jours entre la date de survenance du sinistre et la date de fin d'indemnisation du sinistre.

- **Durée indemnisée du sinistre** : la durée indemnisée du sinistre est le nombre de jours entre la date de début d'indemnisation et la date de fin d'indemnisation du sinistre.

- **Salaire annuel de référence** : le salaire annuel de référence est le salaire trimestriel de référence multiplié par 4.

2.2.2. Ajout de l'information sur l'effectif des entreprises

Le site www.sirene.fr [10] dispose des données publiques sur les entreprises. Pour cela, il faut fournir en entrée un fichier de numéros de SIREN des entreprises recherchées. En retour, le site génère un fichier avec les données caractéristiques en lien avec chaque SIREN, notamment l'effectif des entreprises. Cette donnée vient enrichir les caractéristiques des entreprises adhérentes à ce contrat.

2.2.3. Respect de la chronologie des sinistres

Dans le cas où un assuré remplit toutes les conditions pour bénéficier de chacune des garanties d'arrêts de travail, une chronologie est respectée au sujet du risque :

- Premièrement, il accède à la garantie de maintien de salaire ;
- Ensuite, une fois les droits épuisés, l'incapacité est déclenchée en relais du maintien de salaire ;
- Enfin, après 1 095 jours d'incapacité (depuis la date de survenance de l'arrêt initial), l'assuré bascule en invalidité.

Il est important de vérifier la cohérence chronologique des arrêts de travail. Pour ce faire, les données numéro de sinistre, risque et date de survenance permettent d'identifier si un sinistre en maintien de salaire, un sinistre en incapacité et un sinistre en invalidité ont leur arrêt initial en commun.

Si un sinistre en maintien de salaire est suivi d'un sinistre en incapacité, la date de fin d'indemnisation du maintien de salaire doit être inférieure à la date de début d'indemnisation de l'incapacité.

Si un sinistre en invalidité est le résultat de la fin d'indemnisation d'un sinistre en incapacité, alors la date de début d'indemnisation du sinistre en invalidité doit être supérieure à la date de fin d'indemnisation du sinistre en incapacité.

2.2.4. Gestion des doublons

Il y a une présence de doublons lorsque plusieurs sinistres sont identifiés avec les mêmes valeurs clés, notamment le numéro de sinistre, le numéro de l'assuré, la date de survenance, le risque et la charge de sinistre. Lorsqu'ils existent au sein de la base de données créée, un sinistre unique est retenu, en

[10] Le site http://www.sirene.fr est un service de l'INSEE (Institut National de la Statistique et des Etudes Economiques). Il regroupe des informations sur près de 28 millions d'établissements, en activité ou non.

requalifiant les dates de début et fin d'indemnisation, comme étant respectivement le minimum des dates de début d'indemnisation et le maximum des dates de fin d'indemnisation.

2.2.5. Gestion des rechutes

Nous définissons une rechute lorsqu'un arrêt d'une cause similaire survient au plus tard 2 mois après la fin de l'arrêt précédent.

Pour ce contrat, le taux de rechute observé est de moins de 2%. Par rapport au faible taux de rechute constaté, le choix de retenir un sinistre unique est acté, en requalifiant la date de début d'indemnisation comme étant la date du premier arrêt et la date de fin d'indemnisation comme étant la date de fin d'indemnisation du dernier arrêt, lorsque des rechutes ont été identifiées. On suppose donc que le nouveau sinistre s'apparente à un sinistre sans franchise, et présente ainsi une vision plus prudente en considérant l'état de l'assuré dès l'arrêt initial.

2.2.6. Gestion des censures

Un sinistre est censuré (ou non clos) lorsque des règlements futurs interviendront après la date d'inventaire (ou date d'arrêté comptable), qui correspond à la date de fin d'observation des sinistres.

La base de gestion des provisions mathématiques à l'arrêté comptable du 31 décembre 2020 permet d'identifier les sinistres en incapacité et invalidité connus dans le système de gestion, et qui sont considérés comme étant toujours en cours de paiement. En pratique, un grand nombre d'organismes assureurs a recours à une règle simple qui consiste à définir à dire d'expert les sinistres pour lesquels un provisionnement est requis. Ainsi, un sinistre en incapacité ou invalidité est considéré comme censuré lorsqu'il fait l'objet d'un paiement dans les N mois précédant la date d'inventaire. Pour le portefeuille étudié, les sinistres sont considérés comme étant toujours en cours de paiement lorsque le dernier règlement de prestations a été effectué dans les 4 mois précédant la date d'inventaire.

Il n'y a pas de règle précise pour qualifier la censure d'un sinistre en maintien de salaire. Pour conserver une cohérence pour l'ensemble des arrêts de travail, nous appliquons la même règle aux sinistres en maintien de salaire, incapacité et invalidité. Ainsi, un sinistre en maintien de salaire est considéré comme étant censuré lorsque le dernier règlement intervient pendant les 4 mois précédant la date d'arrêté comptable.

Dans le cas particulier du *backtesting*, l'information précise au sujet des sinistres clos et des sinistres censurés est disponible, puisque les données pour les années ultérieures le sont également. Ainsi, et seulement dans ce cas, un sinistre censuré est un sinistre pour lequel il existe des paiements de prestations *a posteriori* de la date d'arrêté comptable.

A ce stade, 683 720 sinistres survenus entre 2012 et 2020 sont disponibles pour l'analyse.

Risque	Nombre de lignes	Proportion
Maintien de salaire	561 447	82%
Incapacité	115 161	17%
Invalidité	6 662	1%
TOTAL	**683 720**	**100%**

Tableau 2 : Volumétrie des arrêts de travail disponibles pour l'étude

2.3. Observations majeures révélées par les statistiques descriptives

Les statistiques exploratoires présentées ci-après ont permis d'arrêter des choix sur les données finales retenues.

2.3.1. Année de survenance 2020

En suivant la répartition des arrêts de travail par mois pour chacun des risques, l'année de survenance 2020 se comporte très différemment de toutes les autres années de survenance du portefeuille étudié, et ceci, quel que soit le risque. La crise sanitaire liée à la pandémie du COVID-19 est la cause du caractère exceptionnel des sinistres survenus en 2020.

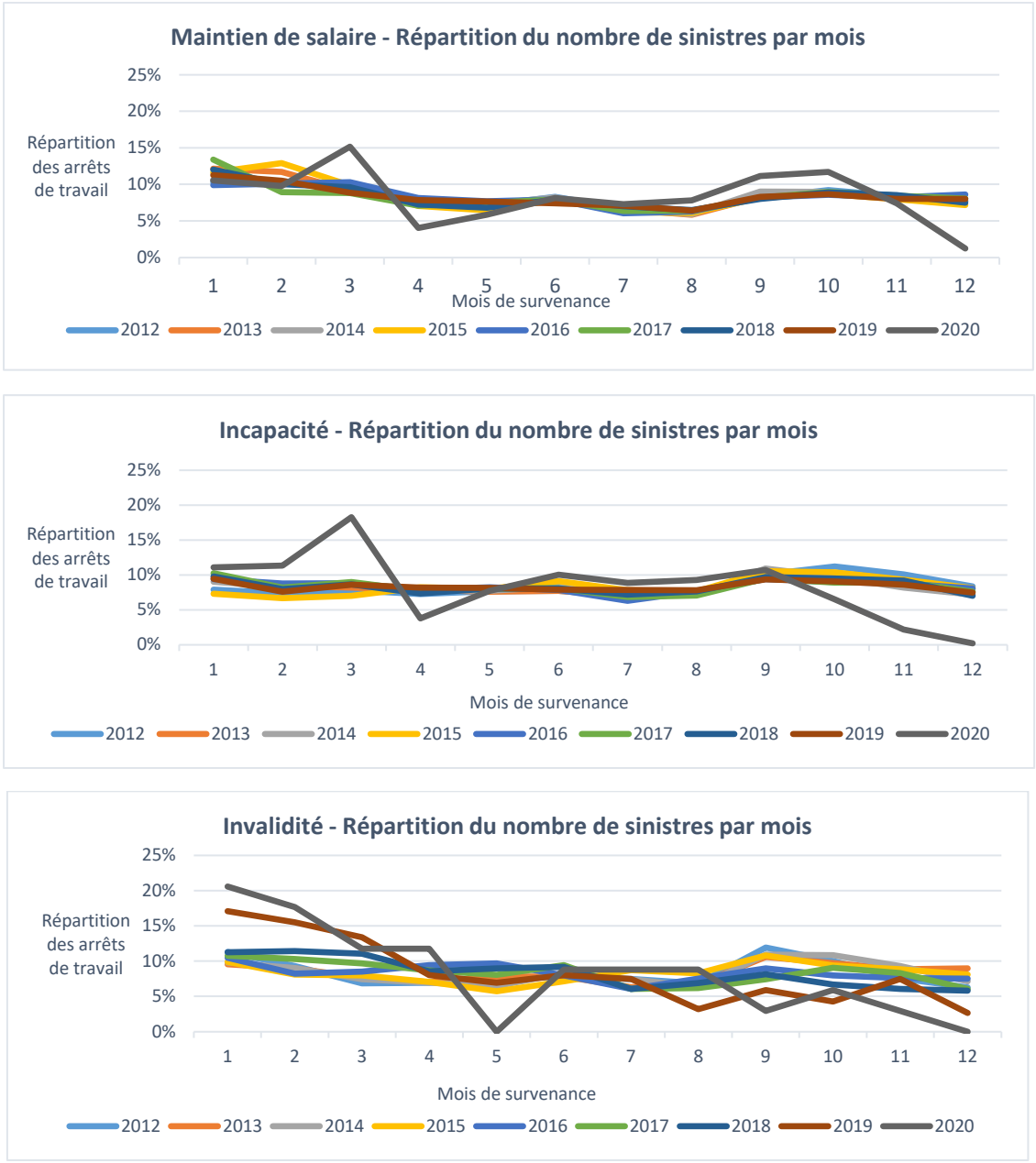


Figure 6 : Répartition du nombre de sinistres par risque et par mois de survenance

Au regard de ces éléments, et pour assurer la qualité de prédiction de la modélisation, nous choisissons d'exclure l'année de survenance 2020 de l'étude.

2.3.2. *Risque « invalidité »*

Ce risque représente moins de 2% de l'ensemble des arrêts de travail et fait partie des risques pour lesquels la censure a le taux le plus élevé, eu égard à la durée contractuelle du risque (à la suite de l'incapacité et jusqu'à la date de départ à la retraite, ou du décès de l'assuré).

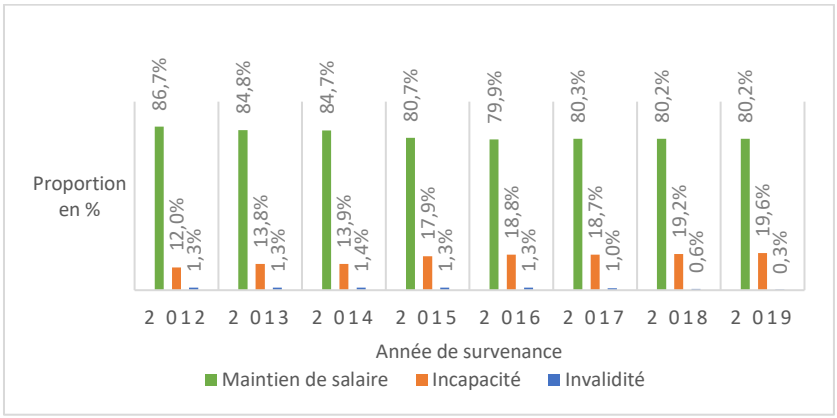


Figure 7 : Proportion des arrêts de travail par risque

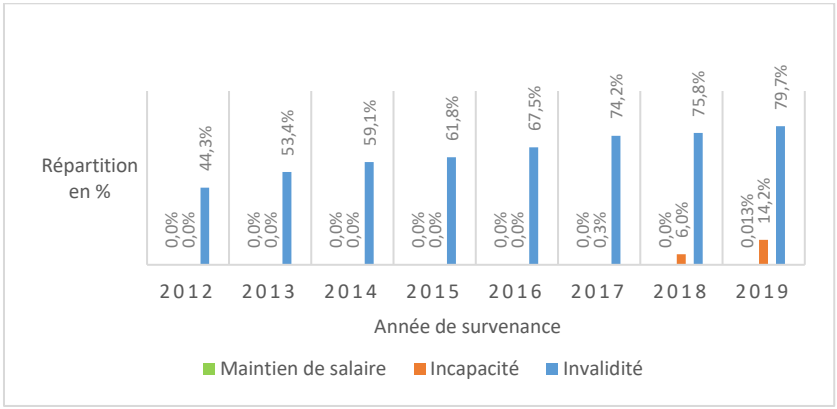


Figure 8 : Arrêts de travail et censure au 31 décembre 2020

Par ailleurs, le taux de censure des sinistres en invalidité survenus en 2012 est de plus de 44%. Ce taux augmente au fil des années de survenance pour atteindre près de 80% en 2019. Le recul sur l'évolution des sinistres en invalidité est très faible.

A ce stade, afin de garantir la performance et la robustesse des modèles, nous choisissons d'exclure le risque « invalidité » du périmètre de l'étude.

Le périmètre regroupant les sinistres survenus de 2012 à 2019 à la date d'inventaire du 31 décembre 2020, hors risque invalidité, se décompose de la manière suivante :

Risque	Nombre de lignes	Proportion
Maintien de salaire	504 784	83%
Incapacité	103 402	17%
TOTAL	**608 186**	**100%**

Tableau 3 : Volumétrie des sinistres en maintien de salaire et incapacité, survenus entre 2012 et 2019, vision à fin 2020

Les variables présentes au sein du portefeuille sont analysées par la suite.

2.4. Analyse des variables

Des études statistiques sont disponibles sur la population sous risque. Etant donné le caractère confidentiel du contrat, elles ne pourront être exposées dans ce mémoire. Toutefois, les résultats des statistiques exploratoires des arrêts de travail du portefeuille étudié seront comparés à ceux de la population sous risque lorsque ces derniers permettront l'interprétation.

2.4.1. Variables qualitatives

La proportion des arrêts de travail en fonction des variables qualitatives est explorée, afin de relever des caractéristiques spécifiques du portefeuille ou d'éventuelles anomalies :

- **Le genre**

Modalité	Libellé	Pourcentage moyen par année de survenance
F	Femme	98,1%
M	Homme	1,9%

Tableau 4 : Genre – analyse des valeurs qualitatives

La population concernée par les arrêts de travail en maintien de salaire et en incapacité est très majoritairement féminine dans le cadre de ce contrat. En effet, pour chacune des années de survenance, plus de 98% des personnes en arrêt de travail sont des femmes. Au sein de la population sous risque, on note plus de 95% de salariées féminines. Cette observation est le résultat de la structure de la population sous risque très typée, et composée principalement de femmes.

- **La situation familiale**

Modalité	Libellé	Pourcentage moyen par année de survenance
C	Célibataire	10,5%
M	Marié(e)	89,4%
V	Veuf(ve)	0,1%

Tableau 5 : Situation familiale – analyse des valeurs qualitatives

Pour les risques en maintien de salaire et incapacité, les personnes concernées sont en moyenne à plus de 89% en couple, contre près de 11% pour les personnes seules. Cette variable étant recueillie par les équipes de gestion au moment de la déclaration du sinistre, la comparaison ne peut être réalisée par rapport à la population sous risque.

- **La catégorie socio-professionnelle**

Modalité	Libellé	Pourcentage moyen par année de survenance
CAD	Cadre	0,8%
NCA	Non-Cadre	99,2%

Tableau 6 : Catégorie socio-professionnelle – analyse des valeurs qualitatives

Pour chacune des années de survenance, plus de 99% des personnes en arrêt de travail sont des salariés non-cadres, quel que soit le risque. La structure de la population sous risque, constituée de plus de 96% de personnel non-cadre, explique également cette répartition en arrêt de travail.

- **Les causes d'arrêt de travail**

Les observations sont présentées en fonction du risque et de la cause :

		Pourcentage moyen par année de survenance	
Modalité	**Libellé**	**Maintien de salaire**	**Incapacité**
1	Maladie	88,3%	92,1%
2	Accident du travail	11,5%	7,8%
3	Maladie professionnelle	0,01%	0,02%
4	Accident de la vie privée	0,01%	0,01%

Tableau 7 : Cause d'un arrêt de travail – analyse des valeurs qualitatives

La maladie reste la principale cause d'arrêt de travail, plus élevée en incapacité qu'en maintien de salaire. L'accident de travail est la deuxième cause d'arrêt de travail la plus importante et représente plus de 11% pour le maintien de salaire, et près de 8% pour l'incapacité. Les maladies professionnelles et les accidents vie privée représentent moins de 0,03%, quel que soit le risque.

- **La franchise contractuelle**

La franchise appliquée au regard des garanties contractuelles est analysée par risque et par modalité :

	Pourcentage moyen par année de survenance	
Modalité	**Maintien de salaire**	**Incapacité**
0 jour	11,7%	80,7%
3 jours	88,3%	16,0%
30 jours		3,3%

Tableau 8 : Franchise contractuelle – analyse des valeurs qualitatives

Pour les arrêts en maintien de salaire, la franchise de 3 jours est largement appliquée. Elle est mise en œuvre dans le cas d'un arrêt pour cause de maladie ou d'accident de la vie privée. Pour les accidents de travail ou les maladies professionnelles, aucune franchise (0 jour) n'est appliquée. A titre de comparaison, la loi de mensualisation applique une franchise de 7 jours pour les salariés du secteur privé. La garantie maintien de salaire du portefeuille étudié prévoit donc des dispositions plus favorables pour ses salariés.

Les arrêts en incapacité pour lesquels il n'y a aucune application de franchise sont majoritaires, de plus de 80%. Ils correspondent aux arrêts en relais du maintien de salaire, ou encore pour cause d'accidents de travail ou de maladies professionnelles. La franchise de 3 jours est appliquée lorsque les droits à la garantie « maintien de salaire » sont épuisés. Environ 16% des sinistres illustrent ce cas. Enfin, la franchise de 30 jours est appliquée si l'ancienneté du salarié au sein du contrat de prévoyance est inférieure à 6 mois. Plus de 3% des sinistres sont concernés par cette dernière franchise.

- **L'effectif des entreprises**

La proportion des sinistres en fonction de l'effectif de l'entreprise est présentée dans le tableau ci-dessous :

Modalité	Pourcentage moyen par année de survenance
0 à 2 salariés	1,3%
3 à 5 salariés	0,4%
6 à 10 salariés	1,4%
10 à 19 salariés	10,1%
20 à 49 salariés	29,9%
50 à 99 salariés	14,9%
100 à 199 salariés	10,0%
200 à 249 salariés	3,3%
250 à 499 salariés	10,8%
500 à 999 salariés	9,4%
1000 à 1999 salariés	1,0%
2000 à 4999 salariés	2,3%
Non disponible en consultation au 31/12/2020	5,1%
Valeur manquante	0,2%

Tableau 9 : Effectif d'une entreprise – analyse des valeurs qualitatives

Les salariés des entreprises pour lesquelles l'effectif se situe entre 20 et 49 salariés sont les plus présents en arrêt de travail, et représentent environ 30% de la population sinistrée. Les statistiques confidentielles disponibles pour la population sous risque indiquent que les entreprises de 20 à 49 salariés sont les plus représentées avec 32% des salariés. On en déduit que la structure de la population sous risque justifie la présence relativement importante des salariés d'entreprises de 20 à 49 salariés.

La valeur « Non disponible en consultation au 31/12/2020 » confirme que le SIREN a bien existé. Toutefois, les données ne peuvent être récupérées en consultation depuis le site www.sirene.fr du fait d'une date de cessation d'activité ou de dissolution trop ancienne de l'entreprise. Ne s'agissant pas d'une anomalie de SIREN, ces données seront conservées dans le cadre de l'étude.

Des effectifs non renseignés sont également relevés. Il s'agit d'entreprises pour lesquelles le SIREN n'a pas été retrouvé sur le site. Ces effectifs ressortent en anomalie et représentent moins de 1% de l'échantillon de données. Face au faible pourcentage de données manquantes, les sinistres liés aux données manquantes seront exclus de l'étude.

La section suivante est consacrée à l'analyse des variables quantitatives du portefeuille.

2.4.2. Variables quantitatives

L'analyse des variables quantitatives est effectuée par quartile. Si elle révèle des anomalies ou des valeurs extrêmes, elle est complétée par une analyse par percentile pour un arbitrage autour de leur suppression. En effet, dans un souci de qualité de prédiction, il faut s'assurer de la faible présence de ces valeurs extrêmes au sein de l'échantillon de données, avant de les écarter du périmètre de l'étude.

Les variables quantitatives sont :

- **L'âge à la survenance d'un arrêt de travail**

Pour l'étude, une population active, âgée de 18 à 67 ans, âge à partir duquel la retraite à taux plein est automatiquement acquise, est retenue.

Une analyse par quartile est réalisée sur le portefeuille de l'étude, afin d'être comparée à la population sous risque.

		Minimum	1er quartile	Médiane	Moyenne	3e quartile	Maximum	Ecart-type
Maintien de salaire	**F**	18	36	47	45	54	67	11,2
	H	18	34	45	44	53	67	11,5
	Total	**18**	**36**	**47**	**45**	**54**	**67**	**11,2**
Incapacité	**F**	18	36	47	45	54	67	11,4
	H	18	37	47	45	54	67	11,2
	Total	**18**	**36**	**47**	**45**	**54**	**67**	**11,4**

Tableau 10 : Age à la survenance – analyse par quartile

L'analyse des quartiles révélant la même répartition, l'analyse se poursuit sans distinction des risques maintien de salaire et incapacité pour le tableau suivant.

Maintien de salaire et Incapacité	[18 - 36 ans] **25%**	[37 - 47 ans] **25%**	[48 - 54 ans] **25%**	[55 - 67 ans] **25%**
Population sous risque en moyenne	**27%**	**36%**	**15%**	**22%**

Tableau 11 : Age à la survenance – comparaison avec la population sous risque

Malgré une distribution équivalente de l'âge à la survenance quel que soit le risque, les différences sont en général peu marquées par genre. On remarque ensuite que les salariés de moins de 37 ans sont plus concernés par les arrêts de travail dans la population sous risque que ceux entre 37 et 47 ans. A partir de 48 ans, plus l'âge augmente, plus les salariés sont enclins à être en arrêt de travail. Cette hausse s'explique par un état de santé des salariés qui se détériore avec l'âge.

- **L'ancienneté du salarié au sein du contrat de prévoyance à la survenance**

Une ancienneté comprise entre 0 et 50 ans est retenue pour le périmètre de l'étude, si on suppose au maximum une carrière entière de 18 à 67 ans réalisée dans la même entreprise pour un salarié.

	Minimum	1er quartile	Médiane	Moyenne	3e quartile	Maximum	Écart-type
Maintien de salaire	0	2	6	7,9	11	50	7
Incapacité	0	2	6	7,4	11	48	6,7

Tableau 12 : Ancienneté du salarié au sein du contrat, à la survenance – analyse par quartile

Les distributions par risque sont très proches. En moyenne, les salariés sont présents entre 7 et 8 ans environ à la survenance d'un sinistre d'arrêt de travail. Les statistiques confidentielles de la population sous risque montrent que les femmes restent en moyenne plus de 7 ans dans une entreprise. On aurait donc tendance à penser que les salariés avec une ancienneté plus importante seraient plus sujets à être en arrêts, car plus fragilisés par le travail. Cependant, 25% des salariés en arrêts de travail ont 2 ans d'ancienneté ou moins. Ce constat va dans le sens inverse des salariés avec une ancienneté plus importante.

- **Le salaire annuel de référence**

	Minimum	1er quartile	Médiane	Moyenne	3e quartile	Maximum	Écart-type
Maintien de salaire	0 €	4 120 €	5 028 €	5 141 €	5 999 €	112 222 €	1 820 €
Incapacité	0 €	4 071 €	4 956 €	5 057 €	5 914 €	45 954 €	1 762 €

Tableau 13 : Salaire de référence annuel – analyse par quartile

Les salaires moyens annuels varient entre 5 057 € et 5 141 €. Le salaire maximal présent dans l'échantillon est de 112 222 €. En pratique, un salaire est considéré comme étant extrême lorsqu'il dépasse 4 fois le plafond annuel de la Sécurité sociale (PASS)[11]. Le PASS le plus faible sur l'historique de l'échantillon est celui de l'année 2012, et sa valeur était de 36 372 €. Le salaire maximal observé dans la base de données n'excède pas 4 fois le PASS le plus faible de l'échantillon. Cette valeur maximale est donc conservée.

Cependant, des salaires de référence annuels nuls sont identifiés, et considérés comme étant des valeurs aberrantes. L'analyse se poursuit alors par percentile afin de s'assurer du pourcentage de sinistres concernés par cette valeur.

	Minimum	1er percentile
Maintien de salaire	0 €	1 246 €
Incapacité	0 €	1 131 €

Tableau 14 : Salaire de référence annuel – extrait de l'analyse par percentile : minimum et 1er percentile

Les salariés en arrêt de travail avec un salaire annuel de référence nul représentent moins de 1% de l'échantillon de données. Du fait de la faible présence dans l'échantillon, les sinistres avec des salaires de référence nuls sont supprimés du périmètre de l'étude.

D'après l'analyse par quartile, plus de 75% des salariés en arrêts de travail perçoivent un salaire annuel de plus de 5 900 €, alors que les statistiques confidentielles de la population sous risque indiquent une rémunération annuelle moyenne de 10 800 €. Les salariés touchés par les arrêts de travail sont ceux pour lesquels la rémunération est relativement faible. Ce constat laisse supposer des salaires peu élevés, qui pourraient s'apparenter à des emplois à temps partiel, ou encore à des cumuls d'emplois. La précarité de l'emploi pourrait être l'explication des salaires relativement bas des salariés en arrêts de travail.

- **Le montant des prestations réglées**

Les prestations réglées au titre des sinistres figurant dans les bases de gestion représentent en moyenne 386 € pour le maintien de salaire et 1 173 € pour l'incapacité.

	Minimum	1er quartile	Médiane	Moyenne	3e quartile	Maximum	Ecart-type
Maintien de salaire	0,01 €	68 €	186 €	386 €	493 €	35 480 €	529 €
Incapacité	0,01 €	104 €	381 €	1 173 €	1 227 €	203 023 €	2 211 €

Tableau 15 : Prestations réglées au titre d'un sinistre – analyse par quartile

[11] Le plafond annuel de la Sécurité sociale (PASS) est un montant fixé par l'Assurance maladie, en fonction de l'évolution annuelle des salaires. Il s'agit du montant de référence pour le calcul de nombreuses cotisations sociales. Le PASS sert de base pour calculer le montant maximal de prestations sociales comme les pensions d'assurance vieillesse du Régime général, les indemnités journalières pour cause de maladie, d'accident du travail, de maternité, de paternité et pour les pensions d'invalidité. Le calcul des plafonds de déductibilité de certaines primes d'assurance à des contrats de prévoyance ou de retraite se base également sur le PASS.

Pour vérifier l'existence des valeurs extrêmes pour les prestations réglées, notamment pour les prestations maximales, une analyse par percentile pour les prestations réglées est réalisée :

	99e percentile	Maximum
Maintien de salaire	2 330 €	35 480 €
Incapacité	9 345 €	203 023 €

Tableau 16 : Prestations réglées au titre d'un sinistre – extrait de l'analyse par percentile : 99e percentile et maximum

Parmi les sinistres en incapacité, 2 d'entre eux présentent des valeurs très élevées, supérieures à 100 000 €. Sachant que le salaire de référence annuel maximal observé pour les arrêts de travail incapacité de l'échantillon est de 45 954 €, et compte tenu des garanties contractuelles (en incapacité, l'indemnité versée représente 70% du salaire annuel de référence sur la durée de l'état en incapacité), le montant maximal attendu pour les prestations versées serait d'environ 97 000 €. Ainsi, les montants constatés apparaissent très élevés et incohérents avec les garanties. Comme ils représentent moins de 1% de l'ensemble des sinistres, et pour garantir la qualité des prédictions, les sinistres dont la charge est supérieure à 97 000 € sont supprimés de l'échantillon de l'étude.

- **La durée totale des sinistres**

La durée totale d'un sinistre représente le nombre de jours entre la date de survenance du sinistre à l'origine de l'arrêt et la date de fin d'indemnisation du sinistre.

	Minimum	1er quartile	Médiane	Moyenne	3e quartile	Maximum	Ecart-type
Maintien de salaire	1	7	13	25	33	366	28
Incapacité	1	58	115	203	250	1 095	228

Tableau 17 : Durée totale (en jours) des sinistres – analyse par quartile

Supérieure ou égale à un jour, la durée totale d'un sinistre d'arrêt de travail représente en moyenne 25 jours pour le maintien de salaire (366 jours au maximum) et 203 jours pour l'incapacité (1 095 jours au maximum).

- **La durée indemnisée des sinistres**

La durée indemnisée d'un sinistre représente le nombre de jours entre la date de début de période d'indemnisation du sinistre au titre du risque, et la date de fin d'indemnisation du sinistre.

	Minimum	1er quartile	Médiane	Moyenne	3e quartile	Maximum	Ecart-type
Maintien de salaire	1	4	11	22	31	364	28
Incapacité	1	18	61	155	188	1 095	217

Tableau 18 : Durée indemnisée (en jours) des sinistres – analyse par quartile

Supérieure ou égale à 1 jour et inférieure ou égale à la durée totale d'un sinistre, la durée indemnisée représente en moyenne 22 jours pour le maintien de salaire (364 jours au maximum) et 155 jours pour l'incapacité (1 095 jours au maximum). Les écarts constatés entre les durées totale et indemnisée s'expliqueraient par l'application de la franchise ou encore l'ancienneté au sein du contrat de prévoyance.

Les statistiques exploratoires du portefeuille révèlent ainsi une population très typée avec majoritairement des femmes, principalement non-cadres et vivant en couple, avec des salaires annuels peu élevés.

3. Sélection du périmètre pour la modélisation de la charge ultime

3.1. Echantillons de données à différentes dates d'arrêté comptable

3.1.1. Backtesting et données arrêtées au 31 décembre 2017

Pour entrainer et calibrer les algorithmes de *machine learning* dans une démarche de *backtesting*, les arrêts de travail survenus entre 2012 et 2017 et observés à la date d'inventaire du 31 décembre 2017 sont sélectionnés. Les sinistres avec des années de survenance entre 2012 et 2017, mais non connus de l'organisme assureur à cette date ne sont pas pris en compte. Les durées maximales contractuelles des sinistres maintien de salaire et incapacité étant respectivement de 12 mois glissants et de 1 095 jours, nous pouvons considérer que les sinistres seront clos à l'arrêté comptable du 31 décembre 2020, soit 1 095 jours après leur survenance. La décomposition des sinistres survenus entre 2012 et 2017 à la date d'arrêté comptable du 31 décembre 2017 est la suivante :

Risque	Nombre de lignes	Proportion
Maintien de salaire	365 565	84%
Incapacité	68 364	16%
TOTAL	**433 929**	**100%**

Tableau 19 : Echantillon des arrêts de travail survenus de 2012 à 2017, vision à fin 2017

L'échantillon de *backtesting* est composé d'une base de sinistres avec la plus grande profondeur d'historique (6 ans) et pour lesquels la charge ultime est disponible. Le rapport entre le nombre de sinistres non clos et le nombre total de sinistres pour chaque risque conduit à un taux de censure de l'ordre de 1% en maintien de salaire et 12% en incapacité.

3.1.2. Echantillon de validation et données arrêtées au 31 décembre 2019

Pour mesurer la pertinence des résultats et l'adaptabilité des modèles, il est recommandé de les appliquer à un échantillon de sinistres différent de celui ayant servi pour le *backtesting*. Les sinistres survenus entre 2018 et 2019, avec les données arrêtées au 31 décembre 2019, sont retenus.

Ce périmètre se décompose comme suit :

Risque	Nombre de lignes	Proportion
Maintien de salaire	108 622	82%
Incapacité	24 441	18%
TOTAL	**133 063**	**100%**

Tableau 20 : Echantillon des arrêts de travail survenus de 2018 à 2019, vision à fin 2019

Les sinistres pour lesquels la dernière indemnisation intervient dans les 4 mois précédant la date d'inventaire sont considérés comme censurés (cf. section 2.2.6). Dans le cadre de cette hypothèse de censure des 4 mois, le taux de censure par risque, ou rapport entre le nombre de sinistres non clos et le nombre total de sinistres pour chaque risque, est de 12% en maintien de salaire et 31% en incapacité.

Avec cet échantillon, toute la visibilité n'est pas disponible sur les règlements futurs potentiels jusqu'à l'ultime, c'est-à-dire jusqu'au 31 décembre 2022, en considérant que tous les sinistres sont clos au bout de 3 ans au maximum. Toutefois, ayant à disposition les données à l'arrêté comptable du 31 décembre 2020 pour les sinistres de cet échantillon, les règlements réalisés pendant l'année 2020 pourront servir de base de comparaison à horizon d'un an avec les résultats de la modélisation de la charge ultime. Certes, ils ne seront pas complets, mais permettront d'analyser les résultats ainsi que la discussion.

Dans la section suivante, les données sont retraitées pour la mise en œuvre des algorithmes, en décrivant les méthodologies appliquées de manière à conserver la cohérence des montants de prestations qui peuvent s'écouler sur plusieurs années.

3.2. Prise en compte de l'inflation

Pour garder une cohérence en termes de montants au sein de l'échantillon afin de prédire la charge ultime des arrêts de travail dont la durée peut s'étaler sur 3 ans, il convient d'introduire l'inflation. L'INSEE définit l'inflation comme « la perte du pouvoir d'achat de la monnaie qui se traduit par une augmentation générale et durable des prix ». L'INSEE la distingue de l'augmentation du coût de la vie, en considérant la perte de valeur comme étant un phénomène qui impacte l'économie nationale dans son ensemble (ménages, entreprises, etc.).

Au sein des échantillons de l'étude, plusieurs montants sont à analyser : les règlements des sinistres, les provisions mathématiques et les salaires annuels de référence.

3.2.1. Règlement de prestations pour le maintien de salaire et l'incapacité

Afin de pouvoir utiliser les données de prestations en arrêt de travail sans être tributaire des années de règlements, il convient de capitaliser les montants des prestations réglées au titre d'un sinistre. Les indemnités journalières évoluent avec les revalorisations du salaire minimum interprofessionnel de croissance[12] (SMIC). Ainsi, l'indice de capitalisation du SMIC horaire est retenu pour mettre au même niveau les montants de prestations.

Son évolution est présentée dans le tableau ci-après :

Année	Smic horaire brut	Revalorisation par rapport à l'année précédente	Date de parution au Journal Officiel
2020	10,15 €	1,2%	19/12/2019
2019	10,03 €	1,5%	20/12/2018
2018	9,88 €	1,2%	21/12/2017
2017	9,76 €	0,9%	23/12/2016
2016	9,67 €	0,6%	18/12/2015
2015	9,61 €	0,8%	22/12/2014
2014	9,53 €	1,1%	19/12/2013
2013	9,43 €	0,3%	21/12/2012
2012	9,40 €	2,0%	29/06/2012
2012	9,22 €		23/12/2011

Tableau 21 : Indice de capitalisation du SMIC entre 2012 et 2020 en France, hors Mayotte

[12] Le SMIC horaire brut en euros est apprécié à la date d'entrée en vigueur du nouveau taux. Il peut donc y avoir un changement des taux en cours d'année.

Disposant de la vision des règlements jusqu'au 31 décembre 2020 pour les données du *backtesting*, il est pertinent d'inflater les montants jusqu'à la fin de l'année 2020. Les prestations réglées entre 2018 et 2020 (RBNS) seront également inflatées de cette manière.

$$Règlement_{t\ inflaté} = Règlement_t \times \prod_{k=t+1}^{2020} (1 + revalorisation\_smic_k)$$

Tous les règlements de prestations sont ainsi inflatés jusqu'en 2020. Une exception est faite pour les règlements réalisés en 2012 : du fait de la double évolution du SMIC en 2012, les règlements du premier semestre 2012 seront inflatés sur 6 mois au taux de 2%, puis annuellement jusqu'en 2020.

Concernant l'échantillon de validation, il aurait fallu émettre des hypothèses d'inflation pour les sinistres qui seront clos au-delà de l'année 2020. Ce point est noté comme étant un axe d'amélioration de l'étude.

3.2.2. Provisions mathématiques pour l'incapacité

Pour inflater les provisions mathématiques, il faut considérer d'une part le taux technique à appliquer, et d'autre part la revalorisation des rentes à prendre en compte.

Voici un récapitulatif des taux techniques appliqués dans le cadre de ce contrat :

Année	Taux technique annuel
2020	0%
2019	0,365%
2018	0,5%
2017	0,5%

Tableau 22 : Taux technique annuel appliqué au calcul des provisions mathématiques en incapacité pour le portefeuille étudié

La formule usuelle pour le calcul des provisions mathématiques en incapacité est la suivante :

$$PM_{incap\ en\ cours_{i,t}} = Rente_{mensuelle,t} \times \sum_{j=c+1}^{d} \frac{1}{(1+i_t)^{\left(\frac{j-a}{12}\right)}} \times \frac{l_{INC_{x,j}}}{l_{INC_{x,a}}}$$

avec :

- x, l'âge d'entrée en incapacité, en années ;
- a, l'ancienneté en incapacité (durée en mois entre la date d'entrée en incapacité et la date de fin du contrat, comprise entre 0 et 35 mois) ;
- b, la franchise éventuelle, durée en mois entre la date d'entrée en incapacité et la date d'effet des garanties ;
- $c = max(a\,;b)$;
- d, la durée en mois entre la date d'entrée en incapacité et la date de fin du contrat (en général, égale à 36 mois) ;
- i_t, le taux technique annuel de l'année t ;
- t, l'année de l'arrêté comptable ;
- $l_{INC_{x,j}}$, l'effectif des personnes entrées en incapacité à l'âge x et toujours en incapacité au terme de j mois, déterminé grâce à une loi de maintien en incapacité ;
- $Rente_{mensuelle,t}$, la rente mensuelle versée pour l'incapacité l'année t.

Notons $\alpha_{i,t} = \sum_{j=c+1}^{d} \frac{1}{(1+i_t)^{(\frac{j-a}{12})}} \times \frac{l_{INC_{x,j}}}{l_{INC_{x,a}}}$, le coefficient d'actualisation pour une rente mensuelle d'1€, au taux technique annuel i_t, à l'arrêté comptable du 31 décembre de l'année t.

Notons $\alpha_{i,2020} = \sum_{j=c+1}^{d} \frac{1}{(1+i_{2020})^{(\frac{j-a}{12})}} \times \frac{l_{INC_{x,j}}}{l_{INC_{x,a}}}$, le coefficient d'actualisation pour une rente mensuelle d'1€, au taux technique annuel i_{2020}, à l'arrêté comptable du 31 décembre 2020.

En considérant que la rente évolue avec l'indice de revalorisation du SMIC, la provision mathématique inflatée à l'année 2020 devient :

$$PM_{incap\ en\ cours\ inflatée} = (Rente_{mensuelle,t} \times 12) \times \prod_{k=t+1}^{2020} (1 + revalorisation\_smic_k) \times \frac{\alpha_{i_{2020}}}{\alpha_{i_t}}$$

Les salariés du portefeuille étudié sont âgés de 18 à 67 ans. Or, la loi de maintien en incapacité du BCAC est disponible pour les âges compris entre 20 et 65 ans. Pour cette étude, on considère que les probabilités de maintien en incapacité à 20 ans sont également observées à 18 et 19 ans. De même, on considère des probabilités de maintien en incapacité identiques de 65 à 67 ans.

3.2.3. Salaire annuel de référence

Avant d'appliquer l'inflation au salaire de référence, il faut s'assurer que les salaires de ce contrat ont réellement été revalorisés annuellement. Les statistiques issues de nos recherches indiquent une non-évolution des salaires depuis une dizaine d'années. Pour ces raisons, il est décidé de ne pas impacter les salaires annuels de référence avec les indices de revalorisation du SMIC.

Le chapitre suivant aborde les triangles de règlements cumulés ainsi que les provisions pour les échantillons de sinistres du périmètre de l'étude.

4. Provisionnement

4.1. Triangles de règlements cumulés

Les triangles de règlements cumulés sont présentés pour les échantillons qui serviront à la modélisation.

Nous choisissons de déterminer les coefficients de passage avec la méthode de *Chain Ladder* en cohérence avec le choix méthodologique réalisé à date sur ce portefeuille.

4.1.1. Echantillon de backtesting : Sinistres survenus entre 2012 et 2017, à l'arrêté comptable du 31 décembre 2017

Les triangles de règlements cumulés et les coefficients de passage de *Chain Ladder* sont déclinés pour un historique de 6 années. Dans un premier temps, les triangles avec les montants inflatés sont construits. Ils seront comparés aux triangles issus des montants non inflatés, afin de vérifier la stabilité des cadences de règlements. Les coefficients de passage de *Chain Ladder* sont calculés pour permettre d'avoir une estimation de l'ultime.

- Pour le maintien de salaire :

	Années de développement					
Survenance	A1	A2	A3	A4	A5	A6
2012	21 680 907 €	30 081 188 €	30 431 470 €	30 431 470 €	30 431 470 €	30 431 470 €
2013	21 857 319 €	30 334 031 €	30 635 628 €	30 635 628 €	30 635 628 €	
2014	22 028 435 €	30 147 008 €	30 433 140 €	30 433 140 €		
2015	18 647 850 €	23 894 687 €	24 174 114 €			
2016	17 302 529 €	22 787 272 €				
2017	15 979 007 €					

Tableau 23 : Maintien de salaire – triangle des règlements inflatés cumulés, vision à fin 2017

	Années de développement				
	A1 – A2	A2 – A3	A3 – A4	A4 – A5	A5 – A6
Coefficient de passage *Chain Ladder*	1,352	1,011	1,001	1,000	1,000

Tableau 24 : Maintien de salaire – coefficients de passage de *Chain Ladder* avec inflation, vision à fin 2017

Pour les sinistres en maintien de salaire, le premier coefficient de développement de *Chain Ladder*, de 1,352, est élevé par rapport aux coefficients suivants. Ce constat corrobore la durée maximale contractuelle des sinistres en maintien de salaire, qui s'étale sur 12 mois consécutifs.

Les résultats avec les règlements non inflatés sont les suivants :

	Années de développement					
Survenance	A1	A2	A3	A4	A5	A6
2012	20 024 504 €	27 828 902 €	28 157 788 €	28 157 788 €	28 157 788 €	28 157 788 €
2013	20 306 849 €	28 265 772 €	28 551 323 €	28 551 323 €	28 551 323 €	
2014	20 682 856 €	28 369 504 €	28 642 105 €	28 642 105 €		
2015	17 655 747 €	22 654 459 €	22 923 149 €			
2016	16 484 281 €	21 758 280 €				
2017	15 365 036 €					

Tableau 25 : Maintien de salaire – triangle des règlements non inflatés cumulés, vision à fin 2017

	Années de développement				
	A1 – A2	A2 – A3	A3 – A4	A4 – A5	A5 – A6
Coefficient de passage *Chain Ladder*	1,354	1,011	1,001	1,000	1,000

Tableau 26 : Maintien de salaire – coefficients de passage de *Chain Ladder* sans inflation, vision à fin 2017

Les coefficients de passage de *Chain Ladder* sont proches, que l'inflation ait été ou non prise en compte pour les règlements. C'est la confirmation d'une stabilité des cadences de règlements.

- Pour l'incapacité :

	Années de développement					
Survenance	A1	A2	A3	A4	A5	A6
2012	2 843 560 €	10 232 563 €	12 952 301 €	12 952 301 €	12 952 301 €	12 952 301 €
2013	3 634 068 €	11 759 711 €	14 803 969 €	14 803 969 €	14 803 969 €	
2014	3 823 681 €	12 330 011 €	15 456 159 €	15 456 159 €		
2015	4 749 021 €	12 760 338 €	15 289 284 €			
2016	5 209 606 €	13 650 323 €				
2017	4 719 325 €					

Tableau 27 : Incapacité – triangle des règlements inflatés cumulés, vision à fin 2017

	Années de développement				
	A1 – A2	A2 – A3	A3 – A4	A4 – A5	A5 – A6
Coefficient de passage *Chain Ladder*	2,998	1,243	1,001	1,000	1,000

Tableau 28 : Incapacité – coefficients de passage de *Chain Ladder* avec inflation, vision à fin 2017

Pour les sinistres en incapacité, le premier coefficient de développement de 2,998 est nettement plus élevé que celui des sinistres en maintien de salaire. La durée contractuelle des sinistres incapacité de 3 ans explique ce coefficient élevé.

Les résultats avec les règlements non inflatés sont les suivants :

	Années de développement					
Survenance	A1	A2	A3	A4	A5	A6
2012	2 632 382 €	9 497 239 €	12 050 845 €	12 050 845 €	12 050 845 €	12 050 845 €
2013	3 376 282 €	11 005 580 €	13 887 878 €	13 887 878 €	13 887 878 €	
2014	3 590 116 €	11 643 893 €	14 622 204 €	14 622 204 €		
2015	4 496 364 €	12 128 821 €	14 560 595 €			
2016	4 963 240 €	13 079 634 €				
2017	4 537 992 €					

Tableau 29 : Incapacité – triangle des règlements non inflatés cumulés, vision à fin 2017

	Années de développement				
	A1 – A2	A2 – A3	A3 – A4	A4 – A5	A5 – A6
Coefficient de passage *Chain Ladder*	3,009	1,245	1,001	1,000	1,000

Tableau 30 : Incapacité – coefficients de passage de *Chain Ladder* sans inflation, vision à fin 2017

Comme pour les sinistres en maintien de salaire, les coefficients de passage de *Chain Ladder* sont proches en incapacité, avec ou sans inflation. Ce constat confirme une stabilité des cadences de règlements.

4.1.2. *Echantillon de validation : Sinistres survenus entre 2018 et 2019, à l'arrêté comptable du 31 décembre 2019*

Les triangles de règlements cumulés et les coefficients de passage de *Chain Ladder* sont déclinés pour un historique de 2 années.

- Pour le maintien de salaire :

	Années de développement	
Survenance	A1	A2
2018	16 101 866 €	21 737 413 €
2019	16 199 160 €	

Tableau 31 : Maintien de salaire – triangle des règlements cumulés, avec inflation, vision à fin 2019

	Années de développement	
Survenance	A1	A2
2018	15 673 540 €	21 242 461 €
2019	16 007 642 €	

Tableau 32 : Maintien de salaire – triangle des règlements cumulés, sans inflation, vision à fin 2019

En se référant au triangle de règlements de l'échantillon de *backtesting*, nous en déduisons que les sinistres survenus en 2018 et observés à la fin de l'année 2019 ont atteint près de 99% de leur charge ultime. Avec une survenance en 2019 en revanche, environ 73% de leur charge ultime est atteinte.

- Pour l'incapacité :

	Années de développement	
Survenance	A1	A2
2018	5 445 477 €	14 322 594 €
2019	5 744 296 €	

Tableau 33 : Incapacité – triangle des règlements cumulés, avec inflation, vision à fin 2019

	Années de développement	
Survenance	A1	A2
2018	5 300 622 €	14 072 788 €
2019	5 676 383 €	

Tableau 34 : Incapacité – triangle des règlements cumulés, sans inflation, vision à fin 2019

En prenant comme référence le triangle de règlements de l'échantillon de *backtesting*, nous en déduisons que les sinistres survenus en 2018 et observés à la fin de l'année 2019 ont atteint environ 81% de leur charge ultime. Pour une survenance en 2019 en revanche, près de 27% de leur charge ultime est atteinte.

4.2. Incapacité et provisions mathématiques

Les provisions mathématiques concernent les sinistres en incapacité et correspondent aux prestations dues pour une durée maximale de 1 095 jours. Pour l'incapacité, les provisions mathématiques à constituer viennent s'ajouter aux provisions RBNS.

Pour l'échantillon de *backtesting* :

Les provisions mathématiques à l'arrêté comptable du 31 décembre 2017 s'élèvent à 19 187 747 €, avec prise en compte de l'inflation.

	Minimum	1er quartile	Médiane	Moyenne	3e quartile	Maximum
PM 2017 inflatées	5 €	1 927 €	2 888 €	3 023 €	3 880 €	36 910 €
PM 2017 non inflatées	5 €	1 854 €	2 777 €	2 907 €	3 731 €	35 482 €

Tableau 35 : Incapacité – Provisions mathématiques, vision à fin 2017

La distribution des provisions mathématiques indique une moyenne d'environ 3 000 € par sinistre.

Pour l'échantillon de validation :

Les provisions mathématiques à l'arrêté comptable du 31 décembre 2019 s'élèvent à 17 041 358 €, avec prise en compte de l'inflation.

	Minimum	1er quartile	Médiane	Moyenne	3e quartile	Maximum
PM 2019 inflatées	4 €	295 €	2 336 €	2 343 €	3 547 €	35 022 €
PM 2019 non inflatées	4 €	291 €	2 303 €	2 309 €	3 496 €	34 509 €

Tableau 36 : Incapacité – Provisions mathématiques, vision à fin 2019

La distribution des provisions mathématiques indique une moyenne autour de 2 300 € par sinistre.

4.3. Récapitulatif des provisions

4.3.1. Echantillon de backtesting

Pour les sinistres survenus entre 2012 et 2017, les provisions inscrites dans le tableau ci-dessous sont le reflet de la réalité, car la visibilité des règlements s'étend jusqu'au 31 décembre 2020.

	Maintien de salaire	Incapacité	Total
RBNS à l'ultime	874 492 €	9 342 488 €	10 216 980 €
PM maintien en incapacité		19 187 747 €	19 187 747 €

Tableau 37 : Sinistres survenus entre 2012 et 2017 – Provisions constatées à l'ultime

Les provisions totales associées à cet échantillon se décomposent de la manière suivante :

- Près de 35% représentent les RBNS dont :
  - 3% pour le maintien de salaire ;
  - 32% pour l'incapacité ;
- 65% font référence aux provisions mathématiques pour les sinistres en incapacité.

Les provisions RBNS sont observées à chaque date d'inventaire et à l'ultime pour l'échantillon de *backtesting* :

	2018	2019	2020
Maintien de salaire	97%	99%	100%
Incapacité	80%	98%	100%

Tableau 38 : Evolution des provisions RBNS cumulées, entre 2017 et 2020

A horizon d'un an, 97% des provisions RBNS sont déjà réglées pour les sinistres en maintien de salaire, contre 80% pour les sinistres en incapacité. A horizon de 2 ans, 2% de provisions RBNS complémentaires sont réglés pour les sinistres en maintien de salaire, alors que pour les sinistres en incapacité, les provisions complémentaires s'élèvent à 18%. A horizon de 3 ans, les sinistres en maintien de salaire et en incapacité sont réglés à l'ultime.

Les provisions mathématiques des sinistres à la date d'arrêté comptable du 31 décembre 2017 sont observées. Avec la même liste de sinistres que précédemment, les provisions mathématiques sont observées dans les bases de gestion à fin 2018, fin 2019 et fin 2020. Le tableau ci-après décrit l'évolution des provisions mathématiques de maintien en incapacité à ces différentes dates d'inventaire.

	Evolution 2017/2018	Evolution 2017/2019	Evolution 2017/2020
Incapacité	-84%	-99%	-100%

Tableau 39 : Incapacité – Evolution des provisions mathématiques en incapacité entre 2017 et 2020

On remarque que les provisions mathématiques en incapacité deviennent nulles au terme de 3 ans, durée maximale de l'incapacité.

4.3.2. Echantillon de validation

Pour les sinistres survenus entre 2018 et 2019, les projections des provisions RBNS sont estimées à partir des coefficients de passage de *Chain Ladder*, sous réserve de non prise en compte des sinistres tardifs et sont présentées dans le tableau ci-dessous. Les règlements étant disponibles jusqu'au 31 décembre 2020, cette information est également indiquée.

	Maintien de salaire	Incapacité	Total
Provisions RBNS à l'ultime	6 210 302 €	19 125 228 €	25 335 530 €
Provisions RBNS - vision à 1 an (31 décembre 2020)	428 215 €	8 043 840 €	8 472 055 €
PM maintien en incapacité		17 041 358 €	17 041 358 €

Tableau 40 : Echantillon de validation – Provisions estimées à l'ultime, vision à fin 2019

A l'issue des statistiques exploratoires, le périmètre est finalisé pour aborder la modélisation dans la partie suivante.

Partie 4 – Application des méthodes d'apprentissage et résultats

1. Mise en œuvre des méthodes d'apprentissage

1.1. *Backtesting*, échantillons d'apprentissage et de test

Pour mettre en œuvre les algorithmes d'apprentissage automatique, le *backtesting* va s'appuyer sur la réalité :

- En positionnant une date d'observation dans le passé pour l'ensemble des sinistres observés ;
- En disposant de la visibilité de tous les règlements des sinistres concernés, jusqu'à une date où tous les arrêts de travail sont clos.

Pour ce faire, les données qui serviront à réaliser la modélisation sont les sinistres survenus entre 2012 et 2017, observés au 31 décembre 2017. Voici les étapes mises en œuvre :

- Extraction des sinistres en maintien de salaire et incapacité survenus entre le 1er janvier 2012 et le 31 décembre 2017. La date d'observation correspond à la date d'arrêté comptable du 31 décembre 2017 ;
- Vérification de la clôture des sinistres au plus tard le 31 décembre 2020 ;
- Vérification de l'absence de prise en compte des tardifs dans l'extraction.

A ce stade, une base de sinistres en maintien de salaire et incapacité est constituée pour une mise en œuvre des algorithmes de *machine learning*. Les caractéristiques pour chaque sinistre sont disponibles, notamment celles liées au salarié, au sinistre lui-même et à l'information du code de clôture. Les montants des prestations réglées ainsi que la charge à l'ultime sont également disponibles. 433 929 sinistres, dont 365 565 en maintien de salaire et 68 364 en incapacité sont répertoriés.

Le taux de censure est d'environ 1% en maintien de salaire et 12% en incapacité. C'est sur cette base que sont calculés les poids IPCW, décrits dans la section 3.2.2 de la partie 2.

Plusieurs phases sont nécessaires pour la mise en œuvre d'un algorithme en *machine learning :*

- La première est la phase d'apprentissage (ou *train*). Elle consiste à fournir un maximum d'information sur la base d'étude de l'algorithme afin qu'il puisse s'entraîner. En particulier, sa volumétrie doit être suffisamment importante. Il est d'usage de retenir entre 70% et 80% des données disponibles ;
- La seconde est la phase de test, au cours de laquelle on s'assure de n'être ni en sur-apprentissage, ni en sous-apprentissage. Elle est réalisée avec des données différentes de celles de l'apprentissage.

La base d'apprentissage sert d'*input* à l'algorithme et correspond à 70% de la base de *backtesting* pour cette étude. Il s'agit d'un échantillon tiré aléatoirement et sans remise. Avec ces données, l'algorithme retient des enseignements concernant les variables les plus discriminantes qui auront influé sur la qualité de la prédiction.

La base de test sert à valider le calibrage et la qualité de prédiction des algorithmes par le biais des indicateurs statistiques de performance WRSE et R^2_W décrits dans la section 3.4 de la partie 2. Il s'agit du complémentaire de la base d'apprentissage. La base de test représente 30% de la base de *backtesting*.

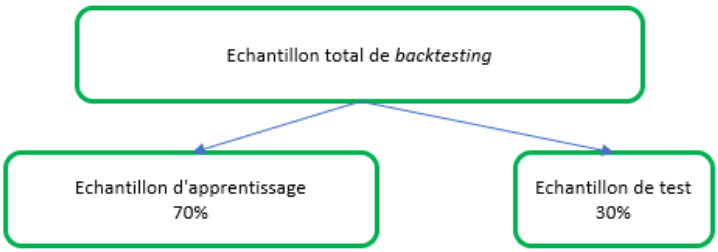


Figure 9 : Découpage de la base de données pour le *backtesting*

1.2. Choix des variables

Le tableau suivant décrit les variables retenues pour les données de *backtesting* :

N°	Variable	Commentaire
1	Genre	Homme, Femme
2	Situation familiale	Personne seule ou en couple
3	Ancienneté contractuelle	Ancienneté du salarié au sein du contrat de prévoyance, arrêt de travail
4	Salaire de référence annuel	En euros
5	CSP	Catégorie socio-professionnelle
6	Effectif	Effectif de l'entreprise du salarié
7	Risque	Maintien de salaire ou Incapacité
8	Prestations réglées	Prestations déjà réglées à la date d'inventaire, y compris inflation
9	Cause	Maladie, Accident de travail, Autre
10	Franchise contractuelle	0, 3 ou 30j selon les cas
11	Mois de survenance	Mois de survenance de l'arrêt initial
12	Age de survenance	Age à la survenance de l'arrêt initial
13	Durée totale	Durée totale de l'arrêt initial (en jours)
14	Code de clôture	Etat clos ou non clos d'un sinistre

Tableau 41 : Variables prises en compte pour la modélisation

La variable réponse à prédire par les modèles est la charge ultime inflatée.

Les charges inflatées de sinistres présentent l'avantage de s'affranchir des évolutions d'une année à l'autre. Ainsi, l'intégration des variables de type « date » au sein des variables explicatives ne présente pas d'utilité. En effet, le modèle construit est censé être capable de prédire une charge ultime à un arrêté comptable ultérieur. Il risque toutefois d'être limité dans la prédiction s'il ne retrouve pas de dates identiques ou proches de celles utilisées lors de la phase d'apprentissage. Néanmoins, pour compenser la perte d'information de la temporalité, la variable « durée » est conservée, et intègre indirectement cette notion. La variable « mois de survenance » est également intégrée aux variables explicatives, avec pour objectif de vérifier l'existence d'une saisonnalité.

La partie suivante présente le calibrage ou *tuning* des algorithmes.

1.3. Calibrage des algorithmes

Le calibrage consiste à ajuster les hyperparamètres des algorithmes. Un nombre important de combinaisons de paramètres sera testé afin de comparer les performances pour déduire le meilleur paramétrage. Le calibrage est réalisé sous R.

1.3.1. CART pondéré

C'est avec le package *rpart* que l'algorithme CART pondéré est implémenté sous R. Il se rapproche dans de nombreux détails à ce qui a été établi par Breiman et al. (1984) [6].

Pour mettre en œuvre cet algorithme, plusieurs paramètres sont à ajuster :

- *minsplit* : il s'agit du nombre minimal d'individus à chaque nœud terminal. Il est maintenu par défaut à 30 individus par feuille ;
- *cp* ou *complexity parameter* : plus la complexité de l'arbre est petite, plus l'arbre peut être grand (nombre important de nœuds). Plus il est grand, plus la complexité est pénalisée. Sa valeur par défaut est de 1% ;
- *method* : dans le cadre d'une régression, la méthode « *anova* » est utilisée pour la construction de l'arbre. Lorsque ce paramètre n'est pas renseigné, l'algorithme se charge de détecter la méthode la plus appropriée à la donnée à prédire ;
- *weights* : les poids IPCW construits avec l'estimateur de Kaplan-Meier sont également renseignés en paramètre pour la prise en compte des données censurées.

Aux résultats de ce modèle, seront comparés ceux issus des algorithmes *Random Forest* et *Gradient Boosting* pondérés.

1.3.2. Random Forest pondéré

C'est avec le package *randomforest* que l'algorithme pondéré du même nom est implémenté sous R. Pour fixer les paramètres, de nombreuses itérations ont été nécessaires. Voici ci-dessous les paramètres ayant servi au calibrage de cet algorithme.

- *ntree* : il s'agit de fixer le nombre d'arbres maximal à construire par l'algorithme. Après avoir fait varier le paramètre *ntree* entre 200 et 1 000, la construction de 200 arbres a été retenue ;
- *mtry* est le nombre de variables échantillonnées au hasard comme candidats à chaque division. Pour la régression, il est recommandé de fixer ce paramètre à la partie entière de $\frac{p}{3}$, p étant le nombre de variables explicatives. Le calibrage de ce paramètre peut être challengé, en divisant ou en multipliant cette valeur par 2. La valeur de *mtry* positionnée à 8 est celle qui donne les meilleurs résultats ;
- *nodesize* représente la taille minimale des nœuds. Après plusieurs essais entre 2 et 5, la valeur 5 est retenue (également la valeur par défaut pour la régression) ;
- *maxnodes* est le nombre maximal de nœuds terminaux que les arbres de la forêt peuvent avoir. Sinon, les arbres sont cultivés au maximum possible (sous réserve des limites de la taille des nœuds). Si la valeur est supérieure au maximum possible, un avertissement est émis. La valeur optimale dans cette étude est fixée à 50, avec des essais entre 10 et 50 ;
- *weights* : comme pour CART pondéré, les poids IPCW construits avec l'estimateur de Kaplan-Meier sont renseignés en paramètre pour tenir compte des données censurées.

1.3.3. Gradient Boosting pondéré

C'est avec le package *gbm* que l'algorithme *Gradient Boosting* pondéré est implémenté sous R. Afin d'approcher une estimation optimale, de nombreuses itérations ont également été mises en œuvre, pour sélectionner un ensemble d'hyperparamètres jugés importants en fonction des données de cette étude :

- *distribution* : elle est positionnée à « *gaussian* », s'agissant d'un problème de régression ;
- *n.trees* est le nombre d'arbres de régression construits par l'algorithme. Il est fixé à 2 000,

après une série de tests entre 1 000 et 10 000 ;
- *shrinkage* : ce paramètre essentiel est également connu sous le nom de taux d'apprentissage ou de réduction de la taille des pas. Un taux d'apprentissage plus faible nécessite généralement plus d'arbres. C'est la valeur optimale de 5% qui a été retenue, parmi des valeurs testées entre 0,1% et 10% ;
- *n.minobsinnode* représente le nombre minimum d'observations dans les nœuds terminaux des arbres. Il est fixé à 5, à la suite d'essais allant de 2 à 5 ;
- *interaction.depth* est le paramètre qui représente la profondeur maximale de chaque arbre, c'est-à-dire le plus haut niveau d'interactions de variables autorisé. Il est fixé à 6, parmi des essais allant de 2 à 10 ;
- *cv.folds* : il s'agit du nombre de validation croisée à effectuer pour l'optimisation des paramètres. Il est fixé à 3, après une série de tests entre 2 et 5 ;
- *weights* : comme pour CART et *Random Forest* pondérés, les poids IPCW construits avec l'estimateur de Kaplan-Meier sont introduits comme paramètre de l'algorithme *Gradient Boosting*.

A la suite du découpage de l'échantillon du *backtesting*, du calcul des poids IPCW, puis du calibrage des algorithmes avec l'échantillon d'apprentissage, les enseignements issus des modèles sont analysés.

2. Application des algorithmes et interprétation des résultats

Les 3 algorithmes CART, *Random Forest* et *Gradient Boosting* sont appliqués aux échantillons sous leur forme adaptée aux données censurées. Chaque algorithme présente l'avantage d'indiquer les variables les plus influentes lors de la phase de construction des modèles. Ce chapitre sera consacré non seulement à l'analyse de ces variables discriminantes, mais également à la comparaison entre les valeurs prédites et réelles des charges ultimes et aux indicateurs de performance. La qualité de prédiction pourra ainsi être mesurée, avec pour objectif la sélection du modèle optimal.

2.1. Variables d'importance

Quel que soit l'algorithme mis en œuvre pendant la phase d'apprentissage, le montant des prestations réglées est classé comme étant la variable la plus discriminante dans la construction du modèle.

Le code de clôture d'un sinistre apparaît comme étant une variable influente et est classé en 2ème ou 3ème position, selon le modèle.

Le montant de salaire annuel de référence semble contenir de l'information pertinente dans le cadre de la prédiction de la charge ultime et figure parmi les 5 variables les plus influentes pour chaque modèle.

Voici ci-dessous les variables les plus influentes, par ordre de positionnement, pour l'application de chaque algorithme :

	CART pondéré	*Random Forest* pondéré	*Gradient Boosting* pondéré
1	**Montant des prestations réglées**	**Montant des prestations réglées**	**Montant des prestations réglées**
2	Durée totale du sinistre	Durée totale du sinistre	**Code de clôture**
3	**Code de clôture**	**Code de clôture**	**Salaire annuel de référence**
4	**Salaire annuel de référence**	Risque	**Mois de survenance**
5	Ancienneté au sein du contrat de prévoyance	**Salaire annuel de référence**	**Effectif de l'entreprise**
6	Age à la survenance	Classe socio-professionnelle	Risque
7	Cause de l'arrêt	**Franchise contractuelle appliquée**	Durée totale du sinistre
8		Cause de l'arrêt	Ancienneté au sein du contrat de prévoyance
9		Mois de survenance	Age à la survenance
10		Age à la survenance	Franchise contractuelle appliquée
11		Situation familiale	Cause de l'arrêt
12		Genre	Genre
13		Effectif de l'entreprise	Classe socio-professionnelle
14		Ancienneté au sein du contrat de prévoyance	Situation familiale

Tableau 42 : Variables explicatives par degré d'importance

Plus spécifiques à chaque algorithme, certaines variables devraient être considérées comme étant influentes :

- L'algorithme CART pondéré détecte comme explicative la variable « ancienneté du salarié au sein du contrat de prévoyance ». L'octroi des garanties intervient après 6 mois d'ancienneté du salarié pour le portefeuille étudié. A titre de comparaison, l'article L. 1226-1 du Code du Travail spécifie la condition d'une année d'ancienneté dans l'entreprise par le salarié pour pouvoir bénéficier de la garantie maintien de salaire de la Sécurité sociale. Ainsi, les conditions d'ancienneté sont plus favorables pour le contrat étudié. L'âge à la survenance fait également partie des variables explicatives les plus pertinentes pour cet algorithme. Toutefois, ces variables sont considérées avec prudence en raison de l'instabilité de l'arbre CART.
- Pour l'algorithme *Random Forest* pondéré, la franchise contractuelle est un facteur explicatif important. Le contrat du portefeuille étudié dispose de franchises relativement courtes, de 0 à 3 jours pour le maintien de salaire, et 30 jours au maximum pour l'incapacité si le salarié ne remplit pas les conditions d'ancienneté du contrat. A titre comparatif, la Loi de mensualisation spécifie une franchise de 7 jours pour les maladies ou accidents de la vie privée. Les conditions du contrat étudié sont donc plus favorables pour accéder aux garanties d'arrêts de travail.
- Le *Gradient Boosting* pondéré accorde de l'importance au mois de survenance de l'arrêt ainsi qu'à l'effectif de l'entreprise. Ce modèle tiendrait compte de la structure de l'entreprise. Cela corrobore les observations issues des statistiques exploratoires sur l'effectif des entreprises. Les petites et moyennes entreprises, de 20 à 49 salariés sont majoritairement représentées tant au sein de la population sous risque qu'au sein de la population sinistrée. Le mois de survenance étant également une variable impactante, il faudrait tenir compte de la saisonnalité des arrêts de travail.

Une fois les variables pertinentes analysées, les résultats issus de l'échantillon de test sont comparés aux valeurs réelles de la charge ultime des sinistres.

2.2. Comparaison des charges ultimes réelles et prédites

La comparaison s'effectue sur l'ensemble des sinistres, puis l'analyse se poursuit en fonction de leur code de clôture (clos ou non clos).

2.2.1. Tous les sinistres

Les graphiques ci-dessous illustrent la comparaison entre les charges ultimes réelles et prédites, pour l'ensemble de l'échantillon de test, et pour chaque algorithme pondéré mis en œuvre :

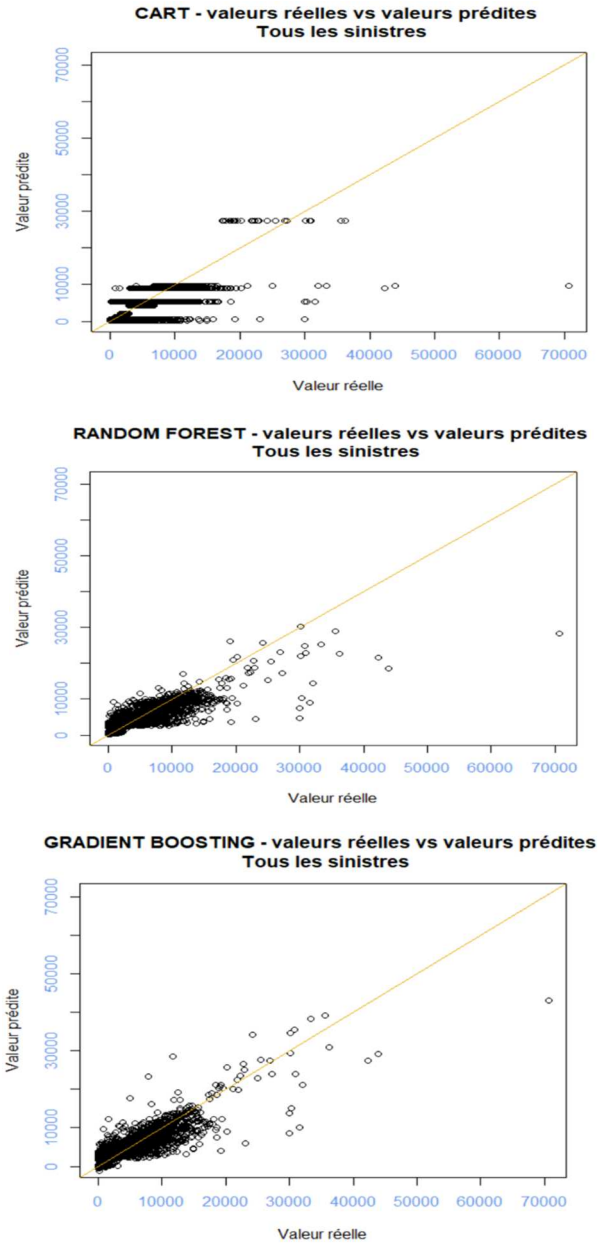


Figure 10 : Tous les sinistres – Comparaison entre charges ultimes réelles et prédites

Les résultats observés pour l'algorithme CART pondéré semblent peu satisfaisants. Un manque de précision est constaté, avec une prédiction qui s'effectue par paliers. De plus, les points sont dispersés de manière importante de part et d'autre de la diagonale.

Une nette amélioration est observée pour l'algorithme *Random Forest* pondéré. Les points se rapprochent plus de la diagonale. On détecte cependant un décrochage des points du côté droit de la

diagonale lorsque les montants des sinistres augmentent. Au fur et à mesure que les charges ultimes évoluent à la hausse, on note un manque de précision. Toutefois ce constat est contenu car les points les plus éloignés de la diagonale sont peu nombreux.

L'amélioration se poursuit avec *Gradient Boosting* pondéré. On note une correction du décrochage observé pour l'algorithme *Random Forest* pondéré. De même, peu de points s'éloignent nettement de la diagonale.

2.2.2. Sinistres clos

Les résultats ci-dessous montrent le fonctionnement des algorithmes pondérés pour les sinistres clos :

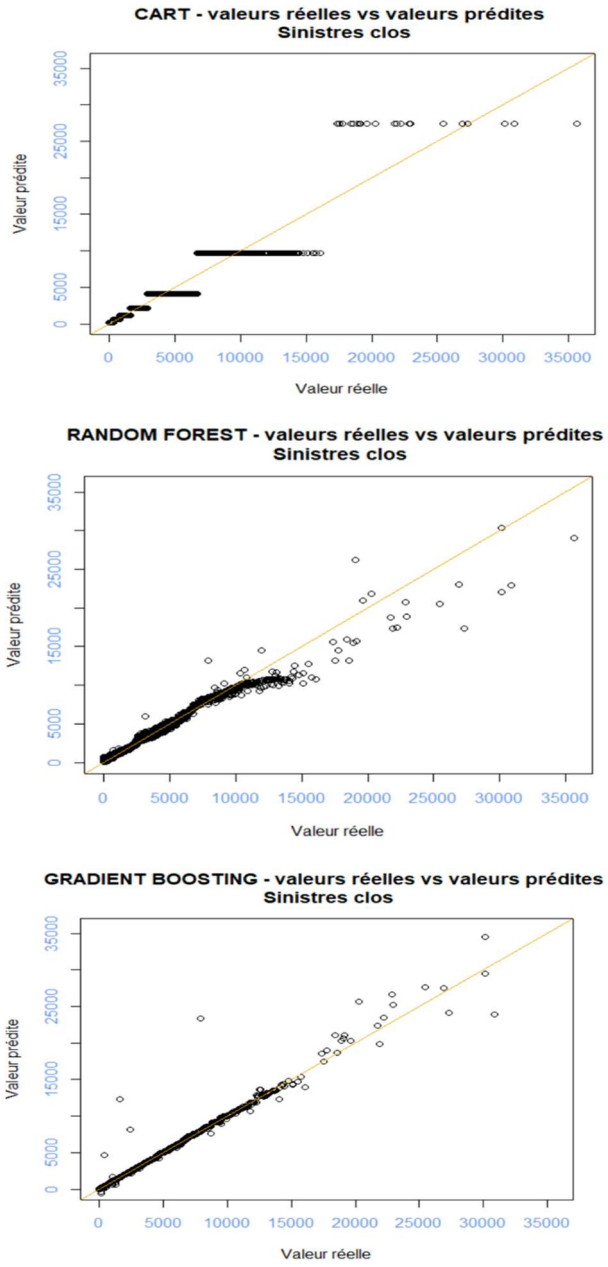


Figure 11 : Sinistres clos – Comparaison entre charges ultimes réelles et prédites

On constate globalement que les résultats pour les sinistres clos sont très nettement meilleurs que pour l'ensemble des sinistres :

- L'algorithme CART pondéré, malgré une amélioration des prédictions pour les sinistres clos, continue d'estimer les charges ultimes par paliers. La précision fait défaut lorsque les charges de sinistres croissent.
- Pour l'algorithme *Random Forest* pondéré, les points se concentrent autour de la diagonale. La dispersion devient visible pour les charges de sinistres plus importantes. Cet algorithme a donc tendance à sous-estimer les valeurs élevées.
- Enfin, pour l'algorithme *Gradient Boosting* pondéré, les résultats sont plus précis que ceux de *Random Forest* pondéré. La dispersion concerne peu de points, associés aux charges les plus importantes.

2.2.3. Sinistres non clos

L'analyse se termine avec les résultats pour les sinistres non clos :

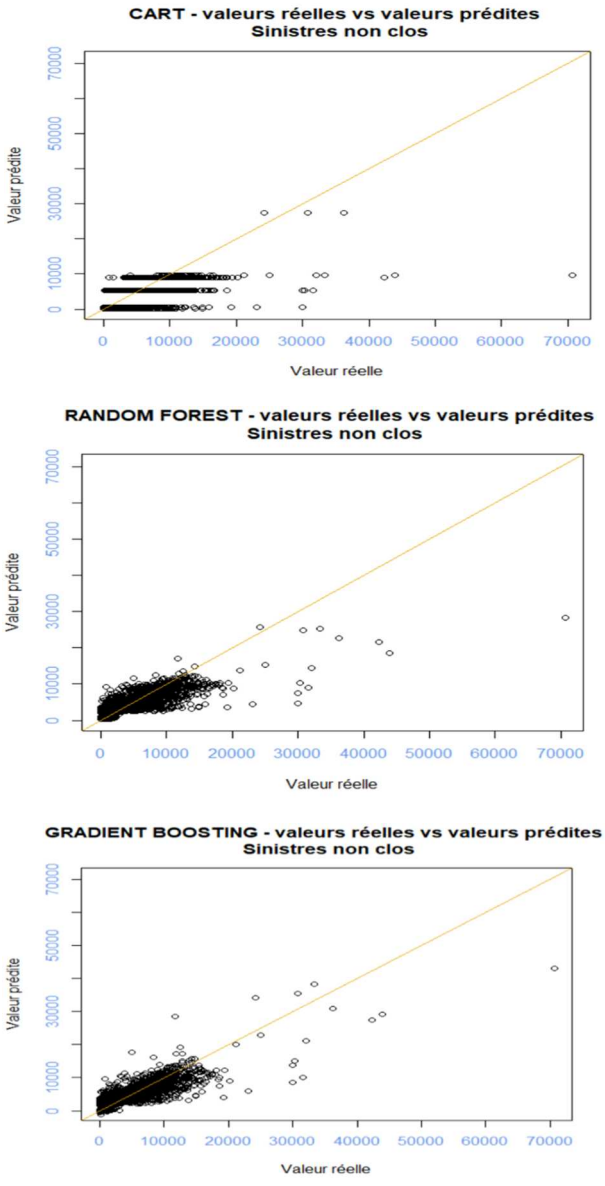


Figure 12 : Sinistres non clos – Comparaison entre charges ultimes réelles et prédites

Tous les algorithmes mis en œuvre ont tendance à réaliser une meilleure prédiction pour les sinistres clos, comparativement aux résultats sur les sinistres non clos :

- En effet, l'algorithme CART pondéré poursuit une prédiction par paliers, avec une majorité de points du côté droit de la diagonale. Cette sous-estimation des charges ultimes pour lesquelles un provisionnement est à réaliser rend le modèle peu prudent.
- L'algorithme *Random Forest* pondéré donne de meilleurs résultats. On retrouve cependant une sous-estimation des charges de sinistres lorsqu'elles croissent.
- Enfin, pour l'algorithme *Gradient Boosting* pondéré, une amélioration est observée par rapport à *Random Forest* pondéré.

A première vue, l'algorithme *Gradient Boosting* pondéré semble se démarquer des autres modèles en réalisant une meilleure prédiction. Il convient de confirmer cette interprétation avec l'appui des indicateurs de performance.

2.3. Indicateurs de performance

Les indicateurs statistiques de performance retenus pour l'estimation de la charge ultime des sinistres sont WRSE (*Weighted Relative Squared Error*) et R^2_w (coefficient de détermination pondéré) adaptés pour les variables à grande amplitude. Voici les résultats obtenus avec l'échantillon de test pour chaque algorithme pondéré implémenté :

	Tous les sinistres			Sinistres clos			Sinistres non clos		
	CART	***Random Forest***	***Gradient Boosting***	**CART**	***Random Forest***	***Gradient Boosting***	**CART**	***Random Forest***	***Gradient Boosting***
WRSE	18,2%	7,8%	5,4%	6,8%	1,3%	0,5%	37,4%	18,7%	13,7%
R^2w	81,8%	92,2%	94,6%	93,2%	98,7%	99,5%	62,6%	81,3%	86,3%

Tableau 43 : Indicateurs statistiques WRSE et R^2w issus de l'échantillon de test

Les indicateurs WRSE et R^2w viennent corroborer les observations constatées lors de la comparaison des valeurs réelles et prédites de la charge ultime. Ils confirment le fait que l'algorithme *Gradient Boosting* pondéré mis en œuvre est le plus performant des 3 algorithmes appliqués aux données *backtestées*.

2.4. Analyse des résultats

L'interprétation de la prédiction de la charge ultime des sinistres est abordée, d'une part pour l'ensemble des sinistres, et d'autre part en considérant la charge moyenne des sinistres, en fonction des variables explicatives les plus pertinentes.

2.4.1. Analyse des résultats globaux

Une première analyse de la somme des charges ultimes prédites est réalisée pour tous les sinistres.

	Valeur réelle	CART pondéré		*Random Forest* pondéré		*Gradient Boosting* pondéré	
		Prédiction	Ecarts observés	Prédiction	Ecarts observés	Prédiction	Ecarts observés
Tous les sinistres	78 423 690 €	77 244 845 €	-1,5%	78 502 709 €	0,1%	78 472 499 €	0,1%
Sinistres clos	65 146 608 €	67 559 356 €	3,7%	65 657 022 €	0,8%	65 248 791 €	0,2%
Sinistres non clos	13 277 082 €	9 685 489 €	-27,1%	12 845 687 €	-3,2%	13 223 709 €	-0,4%

Tableau 44 : Echantillon de test – Charge ultime globale ventilée par code de clôture

Les résultats issus de l'algorithme du *Gradient Boosting* pondéré viennent confirmer les premiers constats en termes de qualité de prédiction. En effet, que les sinistres soient considérés dans leur ensemble ou en fonction de leur code de clôture, les résultats se rapprochent de la réalité avec des écarts faibles par rapport aux valeurs réelles et aux résultats des autres algorithmes. On note une légère sous-estimation des sinistres non clos par *Gradient Boosting* pondéré, avec un écart de -0,4 points par rapport aux valeurs réelles.

Pour CART pondéré, on constate une surestimation des sinistres clos et une forte sous-estimation des sinistres non clos.

Enfin, pour l'algorithme *Random Forest pondéré*, la surestimation des sinistres clos et la sous-estimation des sinistres non clos sont présentes, mais elles sont moins marquées que celles constatées avec l'algorithme CART pondéré.

La comparaison des charges ultimes moyennes de sinistres est maintenant exposée.

2.4.2. Analyse de la charge moyenne de sinistres

Les résultats sont analysés pour l'ensemble des sinistres, puis par risque.

2.4.2.1. Par risque et code de clôture des sinistres

La première analyse consiste à observer la charge ultime moyenne des sinistres. Les résultats ventilés par code de clôture sont présentés ci-après :

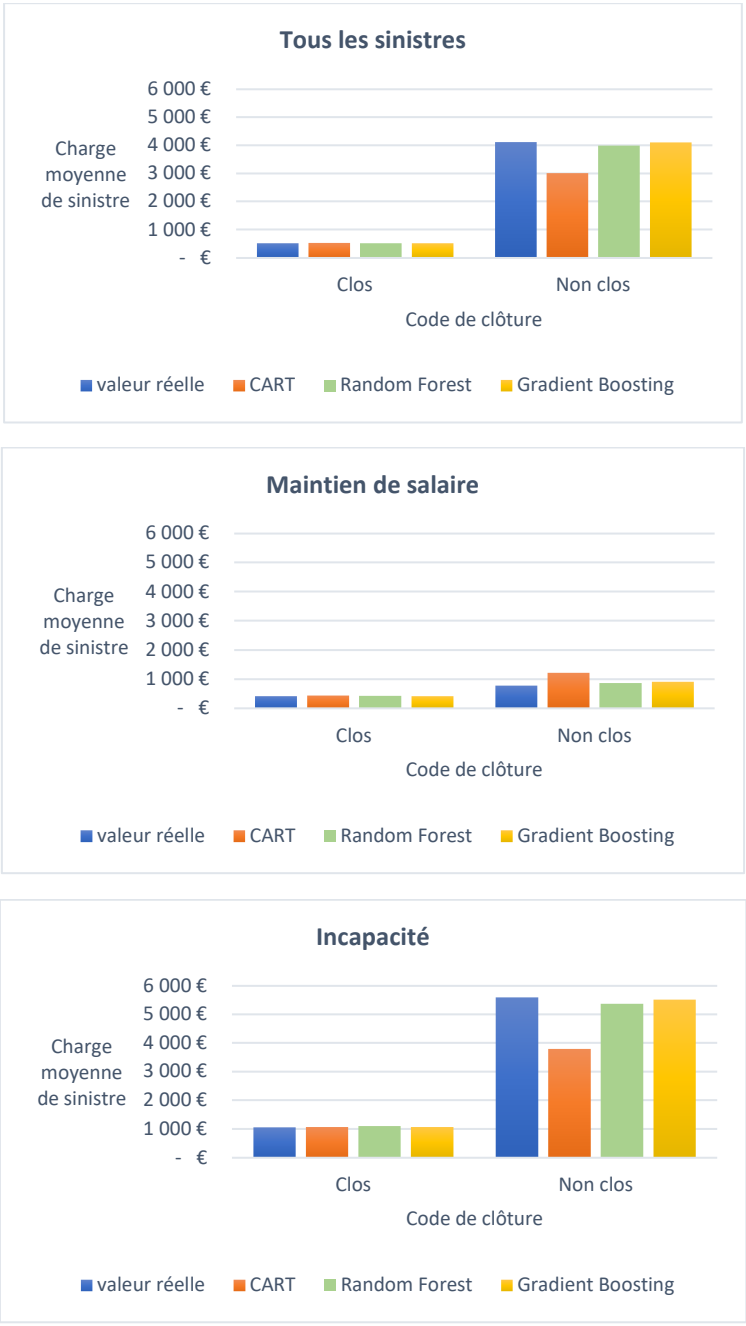


Figure 13 : Echantillon de test – Charge moyenne des sinistres ventilée par risque et code de clôture

Pour l'ensemble des sinistres, on observe une prédiction très proche de la réalité pour les sinistres clos. En revanche pour la charge ultime des sinistres non clos, *Gradient Boosting* et *Random Forest* pondérés réalisent une prédiction proche de la valeur réelle de la charge ultime moyenne, tandis que CART tend à la sous-estimer.

Toutefois, ces résultats globaux masquent des disparités lorsqu'on se focalise sur l'analyse par risque. C'est de nouveau l'algorithme CART pondéré qui réalise les prédictions les moins fiables : il surestime les sinistres en maintien de salaire en cours de paiement. A l'inverse, les sinistres en incapacité ont une prédiction en deçà des valeurs réelles avec le même algorithme lorsqu'ils sont non clos. *Gradient Boosting* et *Random Forest* pondérés présentent des résultats plus acceptables, quels que soient le risque et le code de clôture.

2.4.2.2. Par risque et mois de survenance

Pour répondre à la question « existe-t-il une saisonnalité pour les arrêts de travail du contrat étudié ? », l'observation des charges moyennes des sinistres par mois de survenance et par risque est réalisée :

Figure 14 : Charge moyenne des sinistres ventilée par risque et mois de survenance

Il semble y avoir une saisonnalité pour les arrêts de travail du portefeuille étudié. Il existe une tendance que tous les algorithmes suivent dans leur prédiction. Les pics de sinistralité se situent en mai et septembre pour le maintien de salaire. Ils s'observent en mars, juin et août pour l'incapacité. Cette saisonnalité ne semble toutefois pas corrélée strictement à la météorologie et aux maladies saisonnières.

- Pour le maintien de salaire, le mois de survenance et la charge de travail semblent corrélés. Pour les mois présentant peu de jours travaillés, une surcharge de travail ponctuelle pourrait générer du stress pour les salariés présents à cette période ;
- Pour l'incapacité, les mois où les pics sont observés marqueraient une période de fatigue liée au changement de saison ou à une période de surcharge de travail (surcharge mentale entre vie professionnelle et vie familiale).

Comme précédemment, on note la sous-estimation des sinistres en incapacité et la surestimation des sinistres en maintien de salaire pour les prédictions réalisées avec l'algorithme CART pondéré.

Pour plus de détails, l'annexe B présente les résultats complémentaires sur la charge moyenne des sinistres croisée avec d'autres variables explicatives.

A ce stade de l'étude, nous retenons *Gradient Boosting* pondéré comme étant l'algorithme avec lequel les résultats optimaux sont obtenus.

A travers un échantillon de validation, nous souhaitons nous assurer du fait que cet algorithme s'adapte à des données nouvelles, n'ayant servi ni à la phase d'apprentissage, ni à la phase de test.

3. Validation des algorithmes

3.1. Echantillon de validation

Pour valider les modèles mis en œuvre, les sinistres survenus en 2018 et 2019 et observés à la date d'inventaire du 31 décembre 2019 ont été sélectionnés. Pour être le plus proche de la réalité, des hypothèses ont été émises pour les codes de clôture, en simulant une fin d'observation à la date d'inventaire. Ainsi, tous les sinistres pour lesquels il existe un règlement pendant les quatre mois précédant la date d'inventaire sont considérés comme non clos. Le taux de censure avec cette hypothèse est de 12% en maintien de salaire et 31% en incapacité (cf. section 3.1.2 de la partie 3). C'est cette démarche qui sera appliquée lorsque le modèle sera mis en œuvre sur des données pour lesquelles il n'y aura aucune observation ultérieure à la date d'arrêté comptable.

Toutefois, la vision à fin 2020 des règlements des sinistres est disponible. Ce point permettra de challenger les résultats obtenus, même s'il s'agit d'une observation incomplète.

3.2. Résultats par code de clôture des sinistres

Les résultats présentés sont analysés par code de clôture des sinistres.

3.2.1. Sinistres clos

Analyser les résultats des sinistres clos revient à s'assurer que, non seulement les algorithmes s'adaptent à des nouvelles données, mais aussi à réaliser du *backtesting* sur les sinistres clos à la date d'observation.

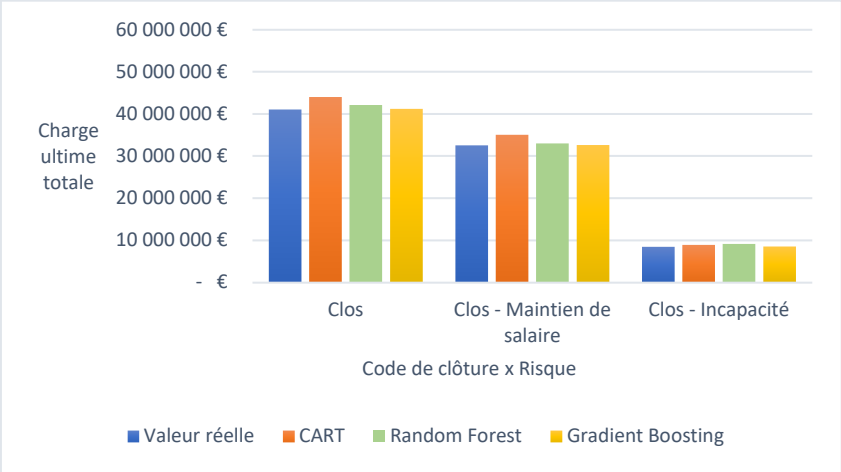


Figure 15 : Echantillon de validation - Charge ultime estimée pour les sinistres clos

Excepté pour CART pondéré qui surestime les sinistres clos, les modèles semblent s'adapter correctement aux nouvelles données de l'échantillon de validation lorsqu'il n'y a pas de censure.

3.2.2. Sinistres non clos

L'intérêt de l'étude réside dans le calcul de la charge ultime pour les sinistres non clos afin d'en déduire le provisionnement à réaliser.

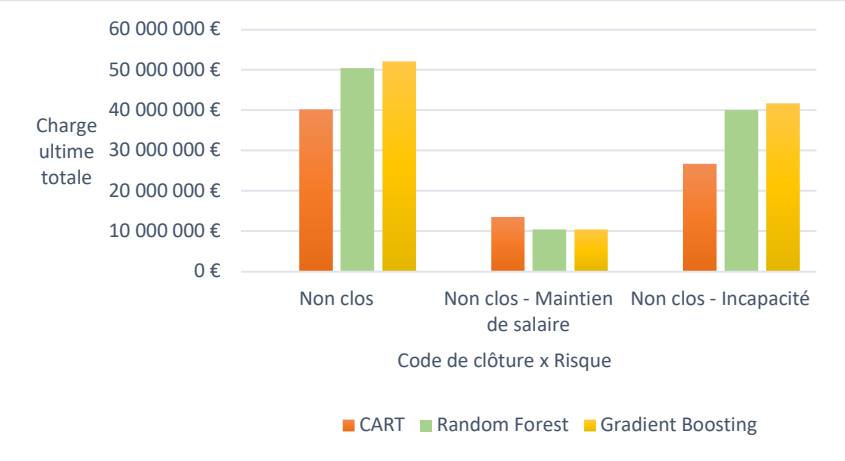


Figure 16 : Echantillon de validation - Charge ultime estimée pour les sinistres non clos

Les résultats des prédictions s'inscrivent dans la même dynamique que celle observée lors de la phase de test. Avec l'échantillon de validation, les estimations de *Gradient Boosting* pondéré sont plus élevées que celles des autres algorithmes. Il se montrerait ainsi plus prudent. En effet, les estimations s'élèvent au global à 52 131 963 € pour *Gradient Boosting* pondéré, contre 40 225 864 € pour CART pondéré et 50 458 285 € pour *Random Forest* pondéré.

Dans la partie suivante, les observations sont confrontées aux prédictions obtenues avec l'algorithme optimal *Gradient Boosting* pondéré en se raccrochant aux observations à la date d'inventaire du 31 décembre 2020.

3.3. Interprétation des résultats

3.3.1. Interprétation des résultats détaillés du modèle prédictif et explicatif

En plus des données arrêtées au 31 décembre 2019, une vision complémentaire au 31 décembre 2020 est disponible. Ces dernières données permettront d'être plus critique face aux prédictions de l'algorithme *Gradient Boosting* pondéré. A la date d'inventaire du 31 décembre 2019, les prestations versées s'élèvent à 58 003 463 €, auxquelles s'ajoutent des provisions mathématiques d'un montant de 17 041 358 €.

Avant de présenter les résultats issus de la prédiction, les prestations observées aux dates d'inventaires du 31 décembre 2019 et du 31 décembre 2020 sont restituées :

	Vision à fin 2019	**Vision à fin 2020**	
		Prestations versées (RBNS) en 2020	**TOTAL**
Prestations totales versées + PM	**75 044 821 €**	**8 472 055 €**	**83 516 876 €**
Maintien de salaire	**37 936 573 €**	**428 215 €**	**38 364 788 €**
Sinistres clos	32 598 199 €	90 684 €	32 688 883 €
Sinistres non clos	*5 338 374 €*	*337 531 €*	*5 675 905 €*
Incapacité	**37 108 248 €**	**8 043 840 €**	**45 152 088 €**
Sinistres clos	8 497 057 €	476 177 €	8 973 234 €
Sinistres non clos + PM	*28 611 191 €*	*7 567 663 €*	*36 178 854 €*

Tableau 45 : Visions à dates d'inventaire différentes pour les prestations de l'échantillon de validation

Ce tableau révèle qu'une année après la date d'arrêté comptable du 31 décembre 2019 :

- En maintien de salaire : les prestations versées en 2020 s'élèvent à 428 215 €. Nous supposons que celles observées pour les sinistres clos (90 684 €) sont liées à des opérations de régularisation ;
- En incapacité, les prestations versées en 2020 s'élèvent à plus de 8 millions d'€. Pour les sinistres clos, on observe des prestations réglées de 476 177 €, identifiées également comme étant liées à des opérations de régularisation.

Après la mise en œuvre de l'algorithme *Gradient Boosting* pondéré, les résultats sont les suivants :

	Vision à fin 2019	Vision à l'ultime avec *Gradient Boosting* pondéré	
		Provisions à l'ultime	Charge à l'ultime
Prestations totales versées + PM	**75 044 821 €**	**18 346 298 €**	**93 391 119 €**
Maintien de salaire	**37 936 573 €**	**5 171 683 €**	**43 108 256 €**
Sinistres clos	32 598 199 €	82 586 €	32 680 785 €
Sinistres non clos	*5 338 374 €*	*5 089 097 €*	*10 427 471 €*
Incapacité	**37 108 248 €**	**13 174 615 €**	**50 282 863 €**
Sinistres clos	8 497 057 €	81 314 €	8 578 371 €
Sinistres non clos + PM	*28 611 191 €*	*13 093 301 €*	*41 704 492 €*

Tableau 46 : Prédiction et provisionnement à l'ultime pour l'échantillon de validation – vision à fin 2019

On constate que les provisions à l'ultime calculées par le modèle sont supérieures de près de 10 millions d'€ par rapport aux observations au 31 décembre 2020.

- Pour le maintien de salaire, les provisions à l'ultime calculées retiennent tout particulièrement l'attention. En effet, l'écart est très important entre les provisions à horizon d'un an s'élevant à 428 215 €, contre plus de 5 millions d'€ à l'ultime. Cela suppose très clairement une surestimation des provisions à l'ultime par le modèle. A horizon d'un an, les provisions pour l'échantillon de *backtesting* représentaient 97% des provisions à l'ultime (cf. section 4.3.1 de la partie 3). Les résultats du modèle prédictif et explicatif sont près de 12 fois supérieurs à l'attendu.

- Pour l'incapacité, les provisions à l'ultime, hors provisions mathématiques, s'élèvent à plus de 13 millions d'€. En se basant sur les observations des données *backtestées*, les provisions RBNS en 2020 représentaient 80% de celles à l'ultime (cf. section 4.3.1 de la partie 3). Les résultats attendus auraient dû être de l'ordre de 10 millions d'€. Ainsi, l'algorithme *Gradient Boosting* pondéré surestime également les provisions RBNS pour l'incapacité de plus de 31**%**.

Quel que soit le risque, le résultat obtenu avec l'échantillon de validation s'éloigne des attentes. En effet, il ne correspond pas aux observations réalisées sur l'échantillon de *backtesting* qui sert de prérequis.

Lors de la phase d'apprentissage, plusieurs variables explicatives ont été mises en avant par les algorithmes. Quel est leur impact sur la prédiction des charges ultimes des arrêts de travail ? Les variables les plus discriminantes pourraient représenter des axes d'analyse afin d'expliquer ces écarts de résultats. Elles sont analysées ci-après.

3.3.2. Code de clôture et hypothèse de censure des arrêts de travail

La cause de la surévaluation des provisions vient de l'hypothèse de censure appliquée aux arrêts de travail. Les sinistres non clos sont ceux pour lesquels la date de fin d'indemnisation intervient dans les 4 mois précédant la date d'arrêté comptable. Avec les résultats de l'échantillon de validation, l'enseignement tiré de la modélisation indique un sur-provisionnement lié à cette hypothèse de

censure. Les provisions ne seraient pas *best estimate*, comme l'exige la directive de Solvabilité 2 [1]. Au contraire, les sinistres clos seraient interprétés comme étant non clos, impliquant le calcul de provisions supplémentaires.

Il est légitime de penser rechallenger cette hypothèse pour les 2 risques : une réduction de la période des 4 mois s'impose. Pour ce faire, des études approfondies sont nécessaires pour retenir une hypothèse fiable de censure.

Les variables discriminantes abordées par la suite viennent principalement expliquer la sinistralité importante du portefeuille étudié.

3.3.3. Salaire annuel de référence

Les conditions de travail jugées précaires pour le contrat étudié, peuvent effectivement justifier une forte incidence en arrêt de travail, pour le maintien de salaire et l'incapacité, comme l'introduisaient déjà les statistiques exploratoires des variables explicatives à la section 2.4.2 de la partie 3.

3.3.4. Mois de survenance

La saisonnalité des arrêts de travail constatée est spécifique au contrat. Il manque cependant des explications précises pour chaque mois de survenance. Une amélioration de l'étude serait d'explorer des effets de contexte, en lien avec les régions, les conjonctures économiques, etc.

3.3.5. Effectif des entreprises

La structure des entreprises des salariés contribue à l'explication de la sinistralité élevée du contrat étudié. Viennent s'ajouter à cette observation, les caractéristiques propres aux salariés dans leur environnement de travail (genre, catégorie socio-professionnelle, précarité, etc.).

3.3.6. Franchise contractuelle

Une franchise contractuelle peu élevée peut effectivement expliquer une incidence plus importante d'arrêt de travail, comparativement à un contrat avec une franchise plus élevée, voire sans garantie maintien de salaire. En effet, plus la franchise est élevée, plus la perte de revenus est importante.

La mise en œuvre de l'algorithme *Gradient Boosting* pondéré sur l'échantillon de validation a permis de détecter les limites du modèle et d'approcher le provisionnement des arrêts de travail sous 2 aspects différents pour le portefeuille étudié :

- D'une part, le modèle démontre que l'hypothèse de censure émise est manifestement « trop » prudente et conduit à un sur-provisionnement pour les sinistres d'arrêts de travail. Elle ne s'inscrit pas dans une démarche *best estimate* et reste à challenger.
- D'autre part, les variables discriminantes les plus impactantes ont permis d'expliquer les causes de la sinistralité importante du portefeuille étudié.

Au terme de l'application des algorithmes, le prochain chapitre recense leurs avantages et leurs inconvénients de manière générale, et plus particulièrement dans ce cas pratique.

4. Synthèse autour des modèles prédictifs et explicatifs implémentés

La prévention de la sinistralité importante d'un contrat de prévoyance en arrêt de travail a conduit à la mise en œuvre de méthodes alternatives pour le calcul des provisions. Le choix s'est porté sur des méthodes innovantes d'apprentissage automatique. Elles présentent plusieurs avantages :

- Les modèles sont non paramétriques ;
- Les axes d'analyses sont variés et ne se limitent pas à un choix restreint de paramètres : ces méthodes tiennent compte de chacune des propriétés d'un sinistre (les caractéristiques individuelles des assurés, les particularités du sinistre, les dispositions contractuelles, etc.) et de leur ordre d'influence.
- La fiabilité de la mesure de la performance est réalisée sur des données de *backtesting* ;
- La modélisation prend en compte les sinistres censurés ;
- Même si nous avons exploité les résultats dans leur globalité dans le cadre de cette étude, le provisionnement individuel est disponible pour chaque sinistre d'arrêt de travail.

En comparaison aux méthodes agrégées comme les triangles de cadences de *Chain Ladder*, faciles à implémenter, mais avec des hypothèses fortes *a priori* sur la stabilité des provisions, les méthodes de *machine learning* permettent de capter toute l'information contenue dans chacune des variables explicatives.

Malgré la pluralité de leurs avantages, ces méthodes présentent cependant des limites :

- Elles nécessitent un volume important de données. En effet, cette contrainte a conduit à écarter les sinistres en invalidité du périmètre étudié, du fait de l'insuffisance de leur volumétrie sur leur clôture ;
- Le calibrage est une étape cruciale pour la mise en œuvre des algorithmes de *machine learning*. Plusieurs itérations sont nécessaires pour trouver les paramètres optimaux, parmi ceux testés. Cependant, il n'y a pas de garantie que les paramètres retenus soient réellement les meilleurs ;
- De fait, le provisionnement individuel ne permet pas la prise en compte des sinistres tardifs. Sur ce point, les méthodes agrégées de *Chain Ladder* prennent l'avantage, car elles tiennent compte des sinistres IBNR ;
- Les années atypiques représentent une limite pour la stabilité des algorithmes de *machine learning*. En effet, nous avons été contraints d'écarter les sinistres survenus en 2020, du fait du caractère exceptionnel de cette année marquée par la crise sanitaire du COVID-19. Pour intégrer les années atypiques dans leur ensemble, notamment 2020 dans notre cas, les assureurs peuvent envisager des méthodes de provisionnement alternatives :
    - La méthode de provisionnement de Bornhuetter–Ferguson prend en charge les triangles instables. Elle présente toutefois la contrainte de l'utilisation d'un ratio de perte attendu *a priori*, déterminé à l'aide de paramètres exogènes aux triangles ;
    - Une autre démarche, moins commune, est celle de Berquist et Sherman. Elle propose 2 méthodes : la première consiste à intégrer l'augmentation des provisions, et la seconde consiste à prendre en compte l'augmentation des règlements.

Dans le cadre d'une application opérationnelle en gestion des risques, nous mesurons l'apport du modèle prédictif et explicatif pour un pilotage ciblé du contrat de l'étude.

Partie 5 – Application dans le cadre de la gestion des risques pour un pilotage ciblé

1. Cadre réglementaire

1.1. Système de gestion des risques

L'environnement au sein duquel évoluent les entreprises d'assurance est en mutation constante et se manifeste par l'augmentation des aléas au niveau mondial (climat, technologies, etc.), l'instabilité de la sphère financière et une évolution continue des normes prudentielles et comptables.

L'article 44 de la directive de Solvabilité 2 [1] rend obligatoire la création d'un système de gestion des risques par les organismes assureurs. Ce dispositif comprend les stratégies, processus et procédures d'information nécessaires pour identifier, évaluer, contrôler, maîtriser et déclarer constamment les risques, individuellement ou de manière agrégée, auxquels les organismes d'assurance sont ou pourraient être confrontés. Il est également appliqué aux interdépendances entre les risques. L'ensemble de ces éléments confèrent aux missions de la gestion des risques leur caractère primordial.

1.2. Acteurs contribuant à la gestion des risques

L'aboutissement des missions de la gestion des risques est lié à la contribution de plusieurs acteurs au sein des compagnies d'assurance. Le rôle des principaux acteurs est décrit ci-dessous :

- **Le Conseil d'Administration ou AMSB (*Administration, Management or Supervisory Body*) :** il supervise l'ensemble des activités de la société, détermine les objectifs et met en place les stratégies pour les atteindre.
- **La Direction Générale** : cette instance exécutive met en œuvre la stratégie de la société et applique les mesures d'orientation et de gestion. Elle suit les recommandations du Conseil d'Administration, réel partenaire sur la mise en œuvre des orientations stratégiques de la compagnie d'assurance.
- **Les Fonctions Clés imposées par la directive de Solvabilité 2**
    - **La fonction Gestion des Risques** : elle assure le déploiement et le suivi du système de gestion des risques. Elle prend en charge le suivi du profil de risque de l'entreprise, identifie et évalue les risques émergents. L'AMSB suit ses recommandations sur les problématiques de gestion des risques ou en lien avec la stratégie de l'entreprise, les opérations de fusion-acquisition et les projets de grande ampleur (cf. article 269 du règlement délégué [16]).
    - **La Vérification de la Conformité** : elle conseille l'AMSB sur le respect des dispositions législatives, réglementaires et administratives et évalue les conséquences de tout changement de l'environnement juridique sur les opérations de la société (cf. articles 46 de la directive Solvabilité 2 [1] et 270 du règlement délégué [16]).
    - **L'Audit Interne** : il évalue l'efficacité du système de contrôle interne et des éléments de gouvernance (cf. articles 47 de la directive Solvabilité 2 [1] et 271 du règlement délégué [16]).
    - **La Fonction Actuarielle** : elle coordonne et supervise le calcul des provisions techniques et s'assure de la qualité des données. Elle émet un avis sur les politiques de provisionnement, de souscription et de réassurance et contribue au système de gestion des risques (cf. articles 48 de la directive de Solvabilité 2 [1] et 272 du règlement délégué [16]).

Responsable de la mise en œuvre du provisionnement, la Fonction Actuarielle nous intéresse tout particulièrement dans le cadre de ce mémoire, notamment à travers ses missions présentées ci-après.

1.3. Missions de la Fonction Actuarielle

Les obligations qui sont du ressort et de la responsabilité de la Fonction Actuarielle s'articulent autour de 4 thématiques principales et sont précisées ci-dessous :

- **Politique de provisionnement et qualité de données** :
  - La Fonction Actuarielle coordonne le calcul des provisions techniques et garantit le caractère approprié des méthodologies, des modèles et des hypothèses utilisés pour le calcul des provisions techniques.
  - Elle compare les meilleures estimations aux observations empiriques et supervise le calcul des provisions techniques dans les cas visés à l'article 82 [13] de la directive Solvabilité 2 [1].
- **Politique de souscription :**
  L'avis de la Fonction Actuarielle est requis concernant :
  - La suffisance des primes compte tenu des prestations et frais futurs, la pertinence de la segmentation tarifaire retenue, les paramètres et hypothèses de tarification, le risque d'antisélection et la qualité des données ;
  - La prise en compte des facteurs environnementaux externes, l'inflation, les changements de la structure du portefeuille pouvant impacter la rentabilité du portefeuille et la souscription future ;
  - Les analyses des mesures tarifaires *a posteriori*, leur pertinence et leur suffisance.
- **Politique de réassurance** :
  L'avis de la Fonction Actuarielle est également requis concernant :
  - L'impact du programme de réassurance sur le bilan et la solvabilité ;
  - La cohérence entre les provisions techniques et le programme de réassurance ;
  - Une politique de réassurance adaptée au profil du risque et à la politique de souscription.
- **Contribution à la gestion des risques** : la Fonction Actuarielle contribue aux politiques écrites avec la fonction Gestion des Risques.

La Fonction Actuarielle produit un rapport, *a minima* annuel, pour rendre compte de l'ensemble des travaux menés et de leurs résultats (cf. article 272 du règlement délégué [16]). Les défaillances rencontrées, ainsi que les recommandations pour y remédier doivent être indiquées. Ce rapport validé par l'AMSB doit être disponible pour l'Autorité de Contrôle Prudentiel et de Réglementation (ACPR).

Par la suite, les propositions d'études pouvant être réalisées sur le contrat étudié seront émises, à travers l'analyse des 3 politiques encadrées par la Fonction Actuarielle.

1.4. Politique de provisionnement et propositions à l'égard du contrat étudié

Les méthodes de *machine learning* ont apporté un éclairage indéniable au sujet de l'hypothèse de censure : elles ont montré l'importance de la rechallenger en fonction des risques, afin de réajuster le provisionnement requis par la réglementation. C'est dans ce cadre que les analyses à mener suivantes sont proposées.

[13] L'article 82 de la directive Solvabilité 2 [1] stipule que « les États membres veillent à ce que les entreprises d'assurances et de réassurance mettent en place des processus et procédures internes de nature à garantir le caractère approprié, l'exhaustivité et l'exactitude des données utilisées dans le calcul de leurs provisions techniques ».

1.4.1. Challenger le provisionnement selon le risque

Pour le maintien de salaire

La méthode usuelle pour le calcul des provisions pour ce risque est basée sur le calcul des PSAP à l'aide des triangles de règlements de *Chain Ladder*. À la suite des résultats du modèle prédictif et explicatif sur l'échantillon de validation, la réduction de la période des 4 mois précédant la date d'inventaire pour l'hypothèse de censure s'impose. L'objectif de cette analyse est de détecter la période nécessaire pour indiquer qu'un sinistre en maintien de salaire est censuré ou non, et de la confronter aux montants des PSAP. Dans la section 4.3.2 de la partie 3, les provisions à l'ultime déterminées avec les triangles de règlements de *Chain Ladder* était de plus de 6 millions d'€ (sous réserve de prise en compte des tardifs) pour l'échantillon de validation.

Le modèle prédictif et explicatif estimait le provisionnement pour les sinistres maintien de salaire à 5 171 683 € avec l'hypothèse de censure pour les sinistres dont les règlements intervenaient pendant la période de 4 mois précédant la date d'inventaire. 12% de sinistres en maintien de salaire étaient alors non clos. Or, les sinistres non clos de l'échantillon de *backtesting* représentaient environ 1%.

Pour retrouver un taux similaire à celui observé pour l'échantillon de *backtesting*, on considère une évolution linéaire simple du taux de censure en fonction de l'hypothèse de censure comme point de départ de l'analyse. En supposant que 12% de sinistres non clos correspondent à 4 mois de censure, l'évolution linéaire nous permettrait d'aboutir à 1% de sinistres non clos avec une hypothèse de censure de **10 jours précédant la date d'inventaire**. Au sein de l'échantillon de validation, le taux réel de censure mesuré avec cette nouvelle hypothèse de 10 jours est d'environ 0,5% (12% avec l'hypothèse initiale). Nous procédons à une analyse de sensibilité et les résultats obtenus permettront de confirmer ou de réajuster l'hypothèse à retenir.

- **Hypothèse de censure : période de 10 jours précédant la date d'inventaire**

  Dans ce cas, un sinistre en maintien de salaire est considéré comme censuré lorsque sa dernière indemnisation intervient dans les 10 jours précédant la date d'inventaire. En actualisant le code de clôture des sinistres en maintien de salaire, les résultats suivants sont obtenus avec le modèle prédictif et explicatif *Gradient Boosting* pondéré :

	Vision à fin 2019	Vision à fin 2020		Vision à l'ultime avec *Gradient Boosting* pondéré	
		Prestations versées (RBNS) en 2020	TOTAL	Provisions à l'ultime	Charge à l'ultime
Maintien de salaire	**37 936 573 €**	**428 215 €**	**38 364 788 €**	**285 908 €**	**38 222 481 €**
Sinistres clos	37 719 936 €	339 909 €	38 059 845 €	75 487 €	37 795 423 €
Sinistres non clos	*216 637 €*	*88 306 €*	*304 943 €*	*210 421 €*	*427 058 €*

Tableau 47 : Hypothèses de censure de 10 jours pour le maintien de salaire

Une année après l'inventaire du 31 décembre 2019, les prestations versées en 2020 s'élèvent à 428 215 €. La variation de l'hypothèse de censure provoque toutefois un effet de « vases communicants » entre les prestations versées en fonction du code de clôture des sinistres. Nous avions supposé que les prestations versées pour les sinistres clos et postérieures à la date d'inventaire étaient des opérations de régularisation. Elles s'élèvent à 339 909 € avec cette nouvelle hypothèse et sont plus importantes que celles des sinistres non clos (88 306 €). Cette limite du modèle reste toutefois à nuancer, du fait de la faible proportion des censures (taux de censure de 0,5%).

A l'ultime, une nette baisse est observée pour les provisions des sinistres de maintien de salaire. Toutefois, elles s'élèvent à 285 908 € et sont inférieures aux prestations versées en 2020. Nous sommes face à une problématique de sous-provisionnement (-35%) pour les sinistres en maintien de salaire. Néanmoins, ce sous-provisionnement est considéré comme léger du fait des faibles montants à provisionner à l'ultime et reste à relativiser.

Même si les résultats obtenus avec la nouvelle hypothèse présentent des imperfections, ils sont nettement plus cohérents que ceux obtenus avec l'hypothèse initiale. Ils sont également plus cohérents que ceux obtenus avec les triangles de règlements de *Chain Ladder* estimés à plus de 6 millions d'€.

Pour l'incapacité

Une démarche similaire à celle des sinistres en maintien de salaire est menée pour les sinistres en incapacité. Le modèle prédictif et explicatif ayant permis de détecter un sur-provisionnement avec l'hypothèse de censure pendant les 4 mois précédant la date d'inventaire, il convient de réévaluer le provisionnement en proposant une hypothèse de censure avec une période plus courte. Pour information, dans la section 4.3.2 de la partie 3, les provisions à l'ultime déterminées avec les triangles de règlements de *Chain Ladder* était de plus de 19 millions d'€ (sous réserve de non prise en compte des tardifs).

Avec l'hypothèse initiale, le taux de censure en incapacité était de 31% pour l'échantillon de validation, contre 12% pour les données *backtestées*. Pour retrouver un taux similaire à celui observé pour l'échantillon de *backtesting*, une évolution linéaire simple du taux de censure en fonction de l'hypothèse de censure est de nouveau considérée comme point de départ de l'analyse. En supposant que 31% de sinistres non clos correspondent à 4 mois de censure, l'évolution linéaire nous permettrait d'aboutir à 12% de sinistres non clos avec une hypothèse de censure de **45 jours environ (1,5 mois) précédant la date d'inventaire**. Au sein de l'échantillon de validation, le taux réel de censure mesuré avec cette nouvelle hypothèse de 45 jours est d'environ 24% (31% avec l'hypothèse initiale). Nous procédons à une analyse de sensibilité.

- **Hypothèse de censure : période de 1,5 mois précédant la date d'inventaire**

Dans ce cas, un sinistre en incapacité est considéré comme censuré lorsque sa dernière indemnisation intervient dans les 45 jours précédant la date d'inventaire. Avec cette nouvelle hypothèse de censure, les provisions mathématiques arrêtées au 31 décembre 2019 passent de 17 041 358 € à 13 124 537 €, affichant une baisse de -23%.

En actualisant le code de clôture des sinistres en incapacité, les résultats suivants sont obtenus avec le modèle prédictif et explicatif *Gradient Boosting* pondéré :

	Vision à fin 2019	Vision à fin 2020		Vision à l'ultime avec *Gradient Boosting* pondéré	
		Prestations versées (RBNS) en 2020	TOTAL	Provisions à l'ultime	Charge à l'ultime
Incapacité	**33 191 427 €**	**8 043 840 €**	**41 235 267 €**	**10 919 328 €**	**44 110 755 €**
Sinistres clos	10 085 523 €	721 259 €	10 806 782 €	62 767 €	10 148 290 €
Sinistres non clos + PM	*23 105 904 €*	*7 322 581 €*	*30 428 485 €*	*10 856 561 €*	*33 962 465 €*

Tableau 48 : Hypothèses de censure de 1,5 mois pour l'incapacité

Comme pour le maintien de salaire, un effet de « vases communicants » est observé en 2020, entre les prestations versées en fonction du code de clôture des sinistres. Nous avions supposé que les prestations versées pour les sinistres clos étaient des opérations de régularisation. Elles s'élèvent à 721 259 €. Elles étaient de 476 177 € avec l'hypothèse de censure initiale.

A l'ultime, les provisions observées sont de près de 11 millions d'€, contre plus de 13 millions d'€ avec l'hypothèse de censure initiale. En se basant sur les observations des données *backtestées*, les prestations réglées en 2020 représentent 80% de celles à l'ultime. Les résultats attendus auraient donc dû être de l'ordre de 10 millions d'€. Au global, un léger sur-provisionnement (+9%) est constaté avec cette nouvelle hypothèse de censure.

Afin de corriger ce léger sur-provisionnement, l'analyse de sensibilité se poursuit en considérant une **hypothèse de censure moindre, fixée à 1 mois**. Au sein de l'échantillon de validation, le taux réel de censure mesuré avec cette nouvelle hypothèse est d'environ 17% (24% avec l'hypothèse des 45 jours et 31% avec l'hypothèse initiale).

- **Hypothèse de censure : période de 1 mois précédant la date d'inventaire**

Dans ce cas, un sinistre en incapacité est considéré comme censuré lorsque sa dernière indemnisation intervient au cours du mois précédant la date d'inventaire. Avec cette nouvelle hypothèse de censure, les provisions mathématiques arrêtées au 31 décembre 2019 passent de 17 041 358 € à 9 347 537 €, affichant une baisse de -45%. En actualisant le code de clôture des sinistres en incapacité, les résultats suivants sont obtenus avec le modèle prédictif et explicatif *Gradient Boosting* pondéré :

	Vision à fin 2019	Vision à fin 2020		Vision à l'ultime avec Gradient Boosting pondéré	
		Prestations versées (RBNS) en 2020	TOTAL	Provisions à l'ultime	Charge à l'ultime
Incapacité	**29 414 426 €**	**8 043 840 €**	**37 458 266 €**	**8 792 963 €**	**38 207 389 €**
Sinistres clos	11 829 978 €	1 294 203 €	13 124 181 €	58 671 €	11 888 649 €
Sinistres non clos + PM	*17 584 448 €*	*6 749 637 €*	*24 334 085 €*	*8 734 292 €*	*26 318 740 €*

Tableau 49 : Hypothèses de censure d'1 mois pour l'incapacité

L'effet de « vases communicants » s'observe également en 2020 entre les prestations versées en fonction du code de clôture des sinistres. Les opérations de régularisation s'élèvent à 1 294 203 € dans ce cas, contre 7 322 581 € avec l'hypothèse de censure de 45 jours et 476 177 € avec l'hypothèse de censure initiale.

A l'ultime, les provisions observées sont d'environ 9 millions d'€, contre près de 11 millions d'€ avec l'hypothèse de censure de 45 jours, et plus de 13 millions d'€ avec l'hypothèse de censure initiale. Or, les résultats attendus, en référence à eux de l'échantillon de *backtesting*, auraient dû être de l'ordre de 10 millions d'€. Au global, un léger sous-provisionnement (-13%) est constaté avec cette nouvelle hypothèse de censure. Toutefois, les prédictions pour les provisions des sinistres non clos sont plus élevées de 4% par rapport à l'attendu (en considérant toujours que l'observé en 2020 représente 80% des provisions à l'ultime pour les sinistres non clos).

Malgré le léger sous-provisionnement constaté, les résultats issus de cette nouvelle hypothèse restent plus cohérents comparativement à ceux obtenus avec l'hypothèse initiale. Ils sont

également plus cohérents que ceux obtenus avec les triangles de règlements de Chain Ladder estimés à plus de 19 millions d'€.

Une analyse des résultats est ensuite effectuée pour sélectionner la meilleure hypothèse de censure pour l'incapacité. L'hypothèse de censure fixée à 45 jours présente 2 avantages :

- L'écart entre les résultats attendus et ceux obtenus avec cette hypothèse est inférieur à celui observé entre les résultats attendus et ceux obtenus avec l'hypothèse de censure d'1 mois.
- Le montant des régularisations est plus modéré avec l'hypothèse de 45 jours.

Compte tenu de ces constats, l'hypothèse de censure de 45 jours (+9% sur le provisionnement à l'ultime) pourrait être privilégiée.

Ainsi, la révision des hypothèses de censure mène aux résultats suivants :

- Pour le maintien de salaire, les provisions sont légèrement sous-estimées par le modèle prédictif et explicatif avec l'hypothèse de censure à 10 jours, comme le montrait au préalable les comparaisons entre les valeurs réelles et prédites lorsque les sinistres sont non clos. Pour pallier ce léger sous-provisionnement, une solution serait d'impacter les provisions estimées d'un coefficient d'ajustement. Les provisions prédites à l'ultime s'élèvent à 285 908 €. Or, en se référant aux résultats obtenus avec l'échantillon de *backtesting*, les provisions attendues à l'ultime sont estimées à 441 458 €. Le ratio de 1,5 appliqué aux provisions prédites à l'ultime permettrait alors d'approximer 100% des provisions attendues.
- Pour l'incapacité, les provisions sont légèrement surestimées avec l'hypothèse de censure à 45 jours. Les provisions prédites à l'ultime s'élèvent à 10 919 328 €. Or, en se référant aux résultats obtenus avec l'échantillon de *backtesting*, les provisions attendues à l'ultime sont estimées à environ 10 millions d'€. Le ratio de 0,9 appliqué aux provisions prédites à l'ultime permettrait alors d'approximer 100% des provisions attendues.
  Ces coefficients d'ajustement restent toutefois à affiner quel que soit le risque.

Des effets conjoncturels propres aux survenances de 2018 et 2019 pourraient également être à l'origine des écarts de provisionnement sur l'échantillon de validation, et que le modèle prédictif et explicatif ne parvient pas à capter.

Par ailleurs, à l'exception des survenances antérieures à 2015, on observe exclusivement des boni de provisionnement pour le portefeuille de l'étude, quel que soit le risque :

- Pour le maintien de salaire, les boni réalisés au titre de la survenance 2019 représentaient plus de 3% des boni des sinistres d'arrêts de travail (2% en 2018).
- Pour l'incapacité, les boni réalisés au titre de la survenance 2019 représentaient 21% des boni des sinistres d'arrêts de travail (17% en 2018).

Malgré les limites du modèle prédictif et explicatif, l'avis de révision de l'hypothèse de censure reste pertinent. Il l'est d'autant plus avec l'observation des boni de provisionnement. En effet, l'intérêt de détenir des marges de prudence trop importantes est limité, car il existe une taxe sur les boni de provisionnement réalisés par les organismes assureurs. Revoir l'hypothèse de censure à la baisse permettrait de limiter l'impact fiscal.

1.4.2. Introduire des tables de provisionnement d'expérience

Compte tenu des spécificités de la population étudiée, il est opportun de vérifier si les tables (ou lois) de maintien utilisées sont adaptées.

Pour l'incapacité

Actuellement, la table de maintien en incapacité du BCAC est utilisée pour le calcul des provisions mathématiques en incapacité. Une méthode pour vérifier si son utilisation est adaptée à la population étudiée est de comparer à un âge moyen représentatif de la population sinistrée, le maintien en incapacité par rapport à la table de maintien en incapacité du BCAC. Pour ce faire, nous considérons l'échantillon de *backtesting* afin de disposer d'une vision à l'ultime des sinistres. Les étapes ci-après décrivent le processus de vérification :

- La première étape consiste à déterminer les tranches d'âge les plus impactées par les arrêts de travail en incapacité. Le graphique ci-dessous montre que les salariés âgés de 44 à 60 ans sont les plus représentés en incapacité.

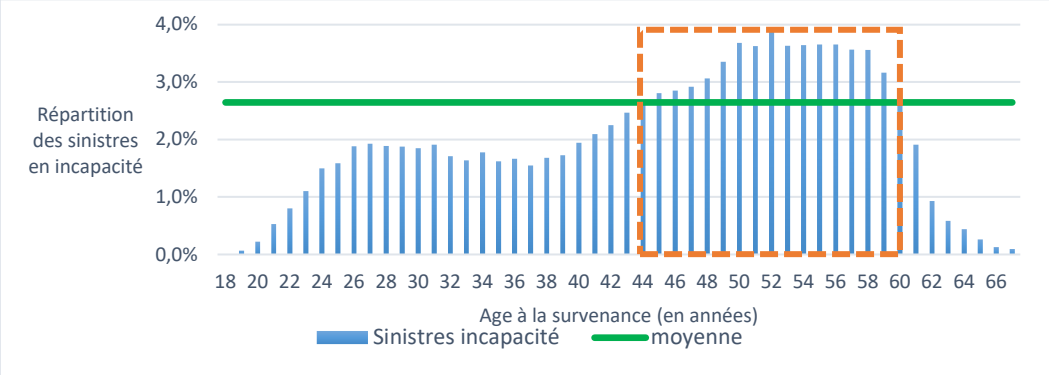


Figure 17 : Répartition des sinistres en incapacité par âge

- La seconde étape consiste à déterminer l'âge moyen des salariés de cette tranche d'âge, ainsi que la durée moyenne dans le sinistre.

	Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum
Age à la survenance (en années)	44	48	52	**52**	56	60
Durée à l'ultime (en jours)	11	117	202	**295**	411	1095

Tableau 50 : Incapacité – Analyse par quartile de l'âge et de la durée pour la tranche d'âge la plus impactée

L'analyse montre que l'âge moyen pour la tranche d'âge la plus impactée par les arrêts de travail en incapacité est de 52 ans. La durée moyenne en incapacité est de 295 jours. Ces observations respectives sont supérieures à celles déterminées pour l'ensemble de la population sinistrée en incapacité avec les statistiques exploratoires à l'inventaire 2020. En effet, l'âge moyen de la population sinistrée à cet inventaire était de 45 ans, et la durée moyenne de 203 jours.

- Enfin, la troisième étape consiste à comparer le maintien en incapacité, lorsque l'arrêt de travail survient à l'âge moyen de 52 ans au sein du portefeuille étudié, aux résultats de la table de maintien en incapacité du BCAC au même âge.
  Pour ce faire, un recensement des salariés entrant en incapacité à 52 ans, de leurs dates de survenance et fin d'indemnisation d'arrêt de travail est réalisé au sein de l'échantillon de *backtesting*. Ensuite, un suivi de l'évolution mensuelle est mis en œuvre pour actualiser le

nombre de personnes toujours en incapacité au fil des mois. Afin de permettre la comparaison avec les tables du BCAC, les compteurs ont été revus en base 10 000 : le nombre de salariés déterminé au premier recensement est ramené à 10 000, et les compteurs suivants sont mis à jour proportionnellement en conséquence.
La comparaison de la loi de maintien en incapacité à l'âge moyen de 52 ans avec la table du BCAC en vigueur est représentée ci-dessous :

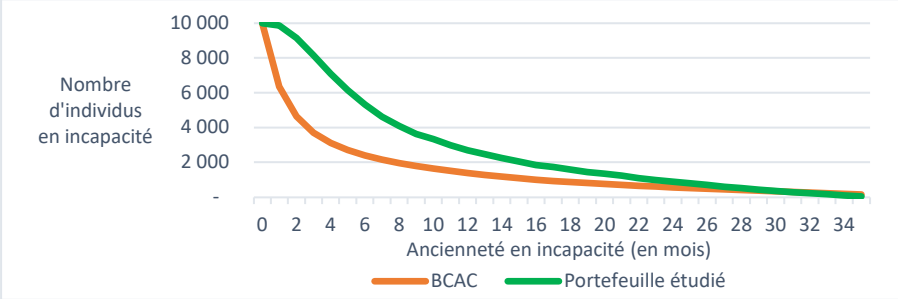


Figure 18 : Loi de maintien en incapacité à l'âge moyen de 52 ans : portefeuille étudié vs BCAC

On constate que la table du BCAC tend à sous-estimer le maintien en incapacité pour le portefeuille étudié sur l'ensemble des mois d'ancienneté. Ce n'est qu'à partir du 31ème mois que la tendance s'inverse. Toutefois, compte tenu du faible nombre d'individus concernés au-delà du 31ème mois, elle reste à relativiser.

Ce constat montre la pertinence de construire des tables d'expérience plus adaptées aux populations présentant des caractéristiques particulières. Elles permettront de s'affranchir des tables du BCAC jugées plus généralistes et d'ajuster les provisions mathématiques.

Pour le maintien de salaire

Il n'existe pas de loi de maintien en arrêt de travail pour le risque maintien de salaire. Néanmoins, il est jugé pertinent de construire une loi de maintien d'expérience dans une optique de toujours affiner la mesure des risques sous-jacents pour ce portefeuille spécifique. Une démarche identique à celle des sinistres en incapacité est mise en œuvre pour les 2 premières étapes (en considérant l'échantillon de *backtesting* afin de disposer d'une vision à l'ultime des sinistres).

- La première étape consiste à déterminer les tranches d'âge les plus impactées par les arrêts de travail en maintien de salaire avec l'échantillon de *backtesting*. Le graphique ci-dessous montre que les salariés âgés de 43 à 59 ans sont les plus représentés en maintien de salaire.

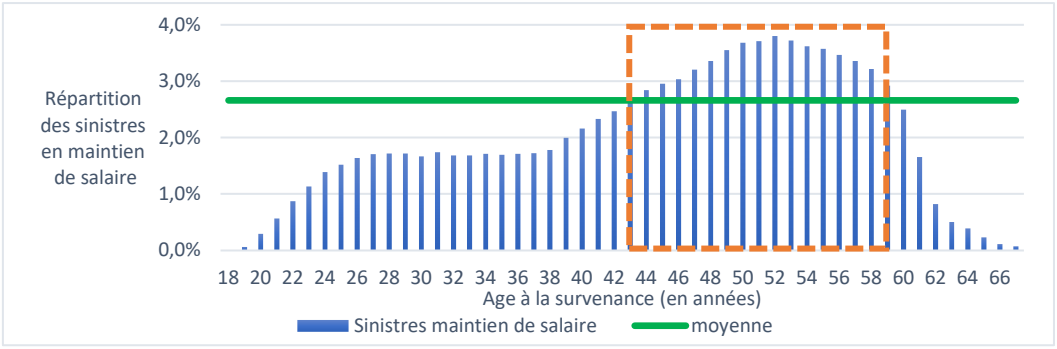


Figure 19 : Répartition des sinistres en maintien de salaire par âge

- La seconde étape consiste à déterminer l'âge moyen des salariés de cette tranche d'âge, ainsi que la durée moyenne dans le sinistre.

	Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum
Age à la survenance (en années)	43	47	51	**51**	55	59
Durée à l'ultime (en jours)	6	51	78	**97**	122	366

Tableau 51 : Maintien de salaire – Analyse par quartile de l'âge et de la durée pour la tranche d'âge la plus impactée

L'analyse montre que l'âge moyen pour la tranche d'âge la plus impactée par les arrêts de travail en maintien de salaire est de 51 ans. La durée moyenne observée est de 97 jours. Ces observations sont supérieures à celles déterminées pour l'ensemble de la population sinistrée en maintien de salaire issues des statistiques exploratoires à l'inventaire 2020. En effet, l'âge moyen de la population sinistrée à cette date d'inventaire était de 45 ans, et la durée moyenne de 25 jours.

- La troisième étape ne peut être réalisée en l'état, ne disposant pas de loi réglementaire existante pour le maintien de salaire. Nous construisons tout de même une ébauche de loi de maintien pour ce risque à l'âge moyen de 51 ans. L'ancienneté au sein du sinistre y est mesurée hebdomadairement. La comparaison est réalisée avec la loi de maintien en incapacité au même âge pour le portefeuille étudié. La table de maintien en incapacité du BCAC, construite par interpolation linéaire des valeurs mensuelles pour obtenir des valeurs hebdomadaires, est également introduite à titre comparatif.

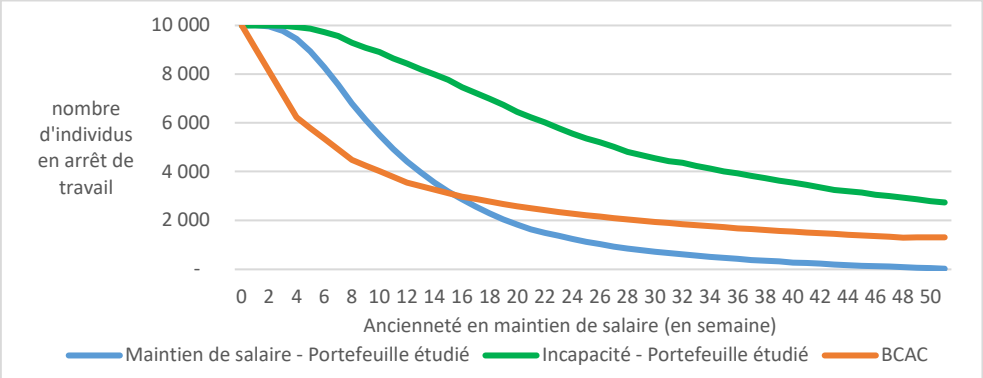


Figure 20 : Ebauche de la loi de maintien à l'âge moyen de 51 ans pour un sinistre en maintien de salaire

Les lois de maintien ci-dessus permettent de mieux situer la proposition de loi de maintien pour le risque en maintien de salaire comparativement à l'incapacité.
Quelle que soit l'ancienneté dans l'arrêt de travail, on constate que le nombre de personnes en arrêt de travail avec la loi de maintien en incapacité du portefeuille étudié est toujours supérieur au nombre de personnes en arrêt de travail dans le cadre de la loi de maintien proposée pour le maintien de salaire.
En revanche, comparativement à cette dernière loi, la table de maintien en incapacité du BCAC tend à sous-estimer le nombre de personnes en arrêt de travail sur les 15 premières semaines d'ancienneté, avant de les surestimer à partir de la 16ème semaine. Ce constat met en évidence le nombre particulièrement élevé de salariés concernés par le maintien de salaire lors des 15 premières semaines d'ancienneté dans l'arrêt de travail.

En définitive, la construction des lois de maintien d'expérience pour les risques en incapacité et maintien de salaire est pertinente, dans le but d'affiner la connaissance de la population sinistrée et d'ajuster le provisionnement.

1.4.3. Vérifier la qualité de données

Pour réaliser un provisionnement *best estimate*, il est indispensable de disposer d'une base de données robuste qui présente une stabilité des observations, c'est-à-dire :

- **Exhaustive** : une profondeur d'historique suffisante doit être disponible, pour permettre de détecter les risques sous-jacents ;
- **Exacte** : les données doivent conserver leur cohérence dans le temps, et être mises à jour régulièrement le cas échéant ;
- **Pertinente** : les données doivent être adaptées aux besoins de leurs utilisations. Elles peuvent ainsi être enrichies de nouveaux facteurs internes ou externes. Ce processus est détaillé ci-après, au sein de la politique de souscription.

1.5. Politique de souscription et propositions à l'égard du contrat étudié

Plusieurs pistes d'amélioration peuvent être explorées.

1.5.1. Enrichir les données

Introduire des facteurs externes

Les données des arrêts de travail peuvent être enrichies par des effets de contexte économiques et régionaux : elles peuvent concerner le dynamisme du secteur, les caractéristiques de la région du salarié, etc. Ces éléments permettront d'affiner les garanties et de proposer des tarifs plus adaptés.

Introduire des facteurs internes

La plupart des contrats assurés en prévoyance disposent également d'une couverture pour les frais médicaux. C'est le cas pour le contrat étudié. Les statistiques exploratoires ont montré que la cause d'un arrêt de travail était majoritairement la maladie non professionnelle. Or, il semble naturel de penser que l'état de santé est la cause principale d'une consommation médicale. Lorsqu'un individu est en mauvaise santé, il aura tendance à avoir recours aux soins (pharmacie – consultations de généralistes ou de spécialistes – soins hospitaliers, etc.). Nous pouvons faire l'hypothèse « raisonnable » que la consommation médicale est un indicateur fiable de l'état de santé d'un individu.

Avec les données du portefeuille sur le risque des frais de santé, les compagnies d'assurance disposent d'éléments pour mesurer l'état de santé des salariés. Les données du portefeuille prévoyance permettent d'identifier les arrêts de travail ainsi que leurs caractéristiques. A partir de cette information, il peut être envisageable d'analyser l'existence d'une corrélation entre la charge de sinistres d'arrêts de travail et les dépenses de santé (en montants et nombre d'actes). Ces nouveaux indicateurs pourraient se révéler utiles pour piloter les deux risques simultanément, notamment dans le cadre d'une importante sinistralité en arrêts de travail.

1.5.2. Construire une loi d'incidence d'expérience

La construction d'une loi d'incidence d'expérience permet de modéliser la fréquence d'entrée en arrêt de travail et de s'adapter à une population sous risque très spécifique. La mise en œuvre de loi d'incidence d'expérience par risque est pertinente, par analogie aux lois de maintien d'expérience proposées dans la politique de provisionnement.

Avec la Déclaration Socio-Nominative (DSN) [14], les déclarations sociales et d'événements sont remplies par tous les employeurs du secteur privé. Les organismes assureurs disposent ainsi d'informations sur l'ensemble des salariés et les arrêts de travail. L'exploitation de la DSN, strictement limitée au cadre du Règlement Général sur la Protection des Données (RGPD) [15], permettra une meilleure mesure de l'incidence.

Dans le cadre du portefeuille étudié, 2 lois d'incidence peuvent être créées : l'une pour le maintien de salaire et l'autre pour l'incapacité. Elles permettront d'approfondir la connaissance de la population sinistrée face à la population sous risque. La démarche consisterait à déterminer en moyenne, pour un âge donné, le nombre de salariés en arrêts de travail rapporté à l'ensemble de la population sous risque. Ces lois d'incidence d'expérience permettront de proposer des tarifs ajustés.

1.5.3. Etudier la rentabilité

L'enrichissement des données et une meilleure appréhension de l'incidence en arrêt de travail rendra possible l'identification d'opportunités de rentabilité sur les risques encourus. L'analyse de l'information par le biais de solutions innovantes, à l'image des méthodes d'apprentissage automatique, peut apporter un nouvel éclairage sur les leviers de rentabilité à mettre en œuvre.

Dans l'hypothèse d'un contrat assuré en accord de branches, l'étude d'opportunité des tarifs plus adéquats conduirait à un pilotage en contrat sur mesure.

En cas d'identification d'une dérive de sinistralité, un transfert de risque est à envisager. Les détails sont précisés dans la section suivante.

1.6. Politique de réassurance et propositions émises à l'égard du contrat étudié

Pour garantir une politique de réassurance efficace, il conviendra de s'assurer de :

- L'analyse des traités significatifs ;
- La prise en compte de l'appétence aux risques ;
- La prise en compte du risque de crédit (défaut du réassureur).

Le contrat étudié est réassuré en quote-part à 70%. En cas de dérive de la sinistralité, il peut être pertinent de revoir la quote-part, ou d'étudier l'opportunité de mise en œuvre d'un traité non proportionnel. Les révisions potentielles des traités de réassurance ne peuvent se faire sans mesure du coût de la réassurance et des impacts sur les exigences en capital.

Les perspectives proposées dans le cadre des missions de la Fonction Actuarielle permettraient de mieux appréhender le risque d'arrêt de travail. Elles restent cependant à nuancer. Le suivi des arrêts de travail et des lois d'expérience demeure une réelle contrainte, dans un environnement en perpétuel mouvement où les facteurs socio-économiques sont difficiles à maîtriser.

En complément des propositions émises, il est opportun de mesurer la valeur ajoutée des études statistiques et du modèle prédictif et explicatif pour un pilotage plus adéquat du contrat étudié.

[14] La Déclaration Socio-Nominative est mise en place dans le cadre de la loi Warsmann, n° 2012-387 du 22 mars 2012. Elle est relative à la simplification du droit et à l'allègement des démarches administratives. Avec les informations concernant chacun des salariés, elle sert à régler les cotisations sociales des entreprises et à transmettre les données aux organismes sociaux. Elle est obligatoire depuis le 1er janvier 2017.

[15] Le Règlement Général sur la Protection des Données (RGPD) encadre le traitement des données personnelles sur le territoire de l'Union Européenne. Il s'inscrit dans la continuité de la Loi Française Informatiques et Libertés de 1978. Il renforce le contrôle par les citoyens de l'utilisation qui peut être faite de leurs données personnelles. Les règles en Europe sont ainsi harmonisées, dans un cadre juridique unique aux professionnels. Il est entré en application le 25 mai 2018.

2. Absentéisme : apport du modèle prédictif et explicatif

2.1. Etat des lieux pour le contrat étudié

L'étude statistique menée sur le portefeuille du mémoire a permis une meilleure connaissance des assurés en termes de typologie : une population caractérisée par une majorité de femmes, principalement non-cadres avec des salaires annuels peu élevés, et vivant en couple.

La mise en œuvre du modèle prédictif et explicatif du provisionnement des sinistres d'arrêts de travail a, pour sa part, mis en lumière des variables discriminantes telles que la rémunération des salariés, la saisonnalité de l'arrêt de travail, la taille des entreprises et la franchise contractuelle qui expliqueraient en grande partie la sinistralité importante.

Ce premier état des lieux des absences liées aux arrêts de travail doit inciter non seulement les employeurs, mais aussi les assureurs à une vigilance face à l'absentéisme. Afin de mettre en œuvre une réelle politique de pilotage de l'absentéisme, il est capital de réaliser un diagnostic juste de la situation. La prochaine section présente la démarche que pourraient entreprendre les organismes assureurs afin d'identifier les leviers d'actions prioritaires et agir sur leurs enjeux spécifiques.

2.2. Nouveaux outils de gestion des arrêts de travail et prévention

Les compagnies d'assurance disposent d'un historique de données chiffrées sur l'absentéisme des populations assurées. Bien que l'analyse des chiffres puisse se révéler complexe, elle est une étape indispensable pour permettre de poser le bon diagnostic de la situation. Plusieurs leviers d'action pour les compagnies d'assurance ont été identifiés.

- Avec un enrichissement des données et la mise en œuvre de méthodes innovantes, à l'image des méthodes d'apprentissage automatique, des causes moins évidentes peuvent être détectées pour apporter des solutions rapidement ;
- Ainsi, les compagnies d'assurance peuvent fournir des services différenciants aux employeurs pour les sensibiliser au sujet de l'absentéisme. Les chiffres, tableaux, graphiques établis constituent de véritables supports qui pourront inciter au débat entre les différents acteurs et être à l'origine de plans d'actions efficaces ;
- Un partenariat avec des organismes de prévention et les fondations d'entreprises peut être mis en place et proposé aux entreprises ou aux partenaires sociaux pour une meilleure gestion de l'absentéisme.

Les résultats d'une telle démarche conduiraient pour l'assureur à une baisse du coût en capital généré par les sinistres d'arrêts de travail.

Par ailleurs, l'année 2020 marquée par la pandémie du COVID-19 a vu les chiffres de l'absentéisme évoluer à la hausse, comme le précise notamment le 13ème baromètre de l'absentéisme du cabinet Ayming [17]. Par rapport à l'année de survenance 2019, on note en 2020 la hausse du nombre d'arrêts de travail (+12%), la hausse du nombre de salariés absents (+13%) ainsi qu'un accroissement important des arrêts de 8 à 30 jours (+21%) pour le portefeuille étudié. Les indicateurs à la hausse impactent également toutes les classes d'âge.

Malgré les hausses importantes observées au cours de cette crise sans précédent, une constante est à retenir pour l'absentéisme : les chiffres continuent de varier à la hausse pour les entreprises. Il est donc important de ne pas considérer les hausses liées aux circonstances exceptionnelles de l'année 2020 comme ponctuelles, mais d'envisager des leviers d'action à plus long terme.

3. Synthèse autour de l'apport du modèle prédictif et explicatif pour un pilotage ciblé

La mise en œuvre du provisionnement des sinistres d'arrêts de travail par des méthodes d'apprentissage automatique est le point de départ pour déterminer des leviers de pilotage ciblé du contrat étudié à travers deux volets :

Le premier volet concerne le cadre réglementaire imposé par la directive de Solvabilité 2 [1]. Le modèle prédictif et explicatif sert d'outil à la fonction Gestion des Risques, et plus précisément à la Fonction Actuarielle, pour challenger les méthodes utilisées, afin d'émettre un avis éclairé sur la politique de provisionnement :

- Tout d'abord, la pertinence de la réduction de l'hypothèse de censure selon le risque a été vérifiée à travers une étude de sensibilité. Bien que présentant des limites, les résultats en termes de provisionnement permettent de corriger le sur-provisionnement observé avec l'hypothèse initiale de censure. De plus, la réduction des marges de prudence permet de limiter l'impact fiscal de la taxation des boni de provisionnement ;
- Ensuite, la construction d'une loi de maintien d'expérience en maintien de salaire s'est avérée pertinente, même avec l'absence de comparaison avec une table dédiée du BCAC ;
- Enfin, la construction d'une loi de maintien d'expérience en incapacité s'est également avérée plus adaptée que l'utilisation de la table de maintien en incapacité du BCAC.

Les apports se mesurent également au niveau des propositions émises pour les politiques de souscription (études de rentabilité et d'opportunités à la suite de possibles enrichissements de données, introduction de lois d'incidence d'expérience, choix du pilotage le plus adapté pour le contrat) et de réassurance (pour pallier une dérive de la sinistralité).

L'absentéisme est le second volet que met en lumière l'apport des statistiques et des enseignements du modèle prédictif et explicatif, pouvant conduire à la mise en œuvre d'une politique efficace de pilotage de l'absentéisme.

- Pour les organismes assureurs : à l'aide de méthodes innovantes de *machine learning* entre autres, les indicateurs ciblés permettraient une meilleure compréhension des comportements au sein du portefeuille assuré en termes d'arrêts de travail ainsi qu'un ajustement du coût en capital de ce risque. De plus, avec leurs données chiffrées disponibles, les organismes assureurs pourraient proposer des services à forte valeur ajoutée aux employeurs afin de les sensibiliser sur les problématiques rencontrées au sein de leurs entreprises ;
- Pour les employeurs : une fois sensibilisés, des actions à court, moyen et long termes pourront être mises en œuvre pour une meilleure gestion de l'absentéisme.

Malgré les nombreuses opportunités que présente cette étude pour le pilotage ciblé du contrat en termes de gestion des risques et d'absentéisme, l'anticipation des risques sous-jacents auxquels les assureurs sont exposés demeure d'une complexité considérable dans un monde en perpétuel changement. L'actualisation et le suivi permanents d'indicateurs clés du système de gestion des risques sont indispensables pour permettre aux organismes assureurs d'une part de garantir leur engagement auprès des assurés et d'autre part, de maîtriser les arrêts de travail, notamment en termes de coût. En particulier, ce dernier point conduirait à une réduction de l'exigence en capital requise au titre de la formule standard.

Conclusion

L'objectif du mémoire était de proposer une méthode alternative de provisionnement des arrêts de travail pour un contrat spécifique présentant une sinistralité importante.

L'intérêt de l'étude a été de réussir à prendre en compte de nombreuses variables (non prises en compte dans les méthodes classiques de provisionnement) dans le cadre d'une modélisation non paramétrique grâce aux algorithmes de *machine learning*. Les méthodes mises en œuvre intègrent la censure des sinistres en pondérant les données dans l'objectif d'obtenir une modélisation au plus près de la réalité. La pertinence de la modélisation du provisionnement a été mesurée en s'appuyant sur un large ensemble de données historiques réelles d'arrêts de travail pour lesquels la charge ultime était disponible (*backtesting*). Une mise en situation concrète de l'estimation du provisionnement avec des données nouvelles a permis d'identifier un sur-provisionnement ainsi qu'un certain nombre de facteurs explicatifs de la sinistralité élevée du portefeuille.

L'application opérationnelle du modèle prédictif et explicatif a permis d'identifier des pistes concrètes pour un pilotage ciblé du contrat.

- D'une part, la fonction Gestion des Risques, et plus particulièrement la Fonction Actuarielle, pourront se servir du modèle présenté pour émettre un avis pertinent sur la politique de provisionnement. La démarche pour un provisionnement *best estimate* passe par la réadaptation des hypothèses de censure, ce qui sous-tend, dans l'environnement social, la réduction de l'impact fiscal des boni de provisionnement réalisés. De plus, l'introduction de lois de maintien d'expérience plus adaptées permet une meilleure appréhension du risque d'arrêt de travail. Les préconisations s'étendent également à la politique de souscription notamment par l'enrichissement des données et les opportunités qui en découlent, et à la politique de réassurance pour un meilleur transfert de risque en cas de dérive de la sinistralité ;

- D'autre part, une première évaluation a été réalisée à travers le modèle prédictif et explicatif pour permettre un pilotage ciblé de l'absentéisme. Une analyse des absences liées aux arrêts de travail offrirait entre autres la réduction de l'exigence en capital des organismes assureurs intervenant sur l'arrêt de travail. Des services différenciants pourront également être proposés aux employeurs pour relever les enjeux stratégiques liés à l'absentéisme et bâtir des plans d'action dans le cadre de la prévention.

Toutefois, le modèle prédictif et explicatif pourrait être amélioré :

- En réussissant à formuler des hypothèses d'inflation à l'ultime, le provisionnement serait plus précis ;
- En intégrant le risque invalidité (sous réserve d'élargissement de la période d'observation), les années atypiques et les sinistres tardifs, le modèle s'améliorerait en termes d'exhaustivité.

Les réseaux de neurones, algorithmes de *machine learning* plus élaborés, pourraient être implémentés dans l'objectif d'obtenir des résultats toujours plus affinés. Une étude des sinistres d'arrêts de travail par année de survenance permettrait de dégager des tendances d'évolution plus spécifiques à intégrer aux estimations.

En définitive, bien que le modèle prédictif et explicatif en l'état ne se substitue pas aux méthodes actuarielles usuelles, sa réelle valeur ajoutée est d'offrir une meilleure visibilité du coût du risque et une meilleure interprétation du provisionnement.

Références

[1] Directive 2009/138/CE du Parlement européen et du Conseil du 25 novembre 2009 sur l'accès aux activités de l'assurance et de la réassurance et leur exercice (solvabilité 2)

[2] Lopez O., Milhaud X., Thérond P.E. (2015) « *Tree-based censored regression with applications to insurance* ». Article de recherche

[3] DRESS (2020). Les dépenses de santé en 2019 – Résultats des comptes de la santé – Edition 2020

[4] https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000027813861/2021-08-15/ (Codes des Assurances, Article R. 331-6 – Provisions techniques des autres opérations d'assurance)

[5] Ressources actuarielles. http://www.ressources-actuarielles.net/bcac

[6] Breiman, L., Friedman, J., Olshen, R. Stone, C. (1984) « *Classification And Regression Trees* »

[7] Vock D., Wolfson J., Bandyopadhyay S., Adomavicius F., Johnson P.E., Vasquez-Benitez G. (2016) « *Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting* ». *Journal of biomedical informatics*

[8] Kaplan, E.L. Meier, P. (1958) « *Nonparametric estimation from incomplete observations* » - *Journal of the American Statistical*

[9] Breiman, L., (2001) « *Random Forests* ». *Machine Learning*

[10] Genuer, R., Poggi, J-M., Tuleau-Malot, C., (2012) « *Variable selection using Random Forests* »

[11] Gregorutti, B., Michel, B., Saint-Pierre, O., (2014) – « Corrélation et importance des variables dans les forêts aléatoires »

[12] Schapire, R. (1990) « *The strength of weak learnability* »

[13] Schapire, R., Freund, Y. (1996) « *Experiments with a new boosting algorithm* »

[14] Friedman, J-H., (2002) « *Stochastic gradient boosting, Computational Statistics and Data Analysis* »

[15] Ayming (2019). 12ème baromètre de l'Absentéisme et de l'Engagement – Edition 2020

[16] Règlement délégué (UE) 2015/35 de la Commission du 10 octobre 2014 complétant la directive 2009/138/CE du Parlement européen et du Conseil sur l'accès aux activités de l'assurance et de la réassurance et leur exercice (solvabilité 2)

[17] Ayming (2020). 13ème baromètre de l'Absentéisme et de l'Engagement – Edition 2021

Cours d'université

[18] Saint Pierre, P. (2015) « Introduction à l'analyse des durées de survie » – Université Pierre et Marie Curie

[19] Rakotomalala, R. (2016) « *Gradient Boosting* » – Université Lumière Lyon 2

Mémoires d'actuariat

[20] Yvroud, H. « Construction de lois de maintien en arrêt de travail en Australie par segmentation non-paramétrique de la population » – ENSAE ParisTech, 2018

[21] Barbaste, M. « Une méthode de provisionnement individuel par apprentissage automatique » – ISFA, 2017

[22] Gibaud, G. « Revue des provisions dossier/dossier avec des méthodes de *machine learning* » – ISFA, 2017

[23] Ottou, P. « Méthodes d'apprentissage automatique appliquées au provisionnement ligne à ligne en assurance non-vie » – Université Paris Dauphine, 2017

[24] Kiema, F. « Méthodes de *machine learning* en provisionnement non-vie : Construction de provisions dossier/dossier pour des sinistres de protection juridique » – Université de Strasbourg, 2018

Thèses

[25] Olympio, A. (2019). « Contribution au provisionnement en assurance de personnes et à la gestion des risques » – ISFA

[26] Baudry, M. (2020) « Quelques problèmes d'apprentissage statistique en présence de données incomplètes » – Université de Lyon

Note de synthèse

Introduction

Les organismes assureurs sont tenus réglementairement de provisionner les prestations futures d'un sinistre d'arrêt de travail en cours d'indemnisation jusqu'à la fin de l'arrêt. La thématique du provisionnement est d'autant plus importante qu'elle représente une exigence de la directive de Solvabilité 2 [1], qui impose que les provisions calculées soient *best estimate*. L'approche classique utilisée est basée sur les triangles de *Chain Ladder* pour les provisions pour sinistres à payer (PSAP) et le calcul paramétrique pour les provisions mathématiques. Toutefois, lorsqu'un portefeuille présente une sinistralité importante (augmentation continue de la sinistralité, population sous risque avec des caractéristiques particulières, etc.), il est opportun d'envisager des méthodes alternatives de calcul des provisions.

Ces dernières décennies, les algorithmes de *machine learning* se sont révélés très utiles et adaptés à l'analyse des données du monde de l'assurance. Dans leur article « *Tree-based censored regression with applications to insurance* », Lopez et al. (2015) [2] propose la prédiction de la charge ultime d'un sinistre, clos ou toujours ouvert, en utilisant des arbres de régression. L'originalité de cette méthode est l'application des algorithmes à des observations censurées, en introduisant une pondération pour corriger le biais induit par la censure.

L'objectif du mémoire est d'évaluer la pertinence des méthodes d'apprentissage automatique pour challenger les méthodes de provisionnement usuelles.

Contexte

Depuis plusieurs années, les dépenses liées aux arrêts de travail sont en constante hausse. La Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (DREES) estime le montant des indemnités journalières prises en charge par la Sécurité sociale à 15,7 milliards d'euros en 2019 [3]. Une progression continue des arrêts maladie est observée depuis 2014. Les facteurs explicatifs sont une population active qui vieillit, des arrêts longs pour les actifs les plus âgés ou fragilisés, etc.

Pour les organismes assureurs, l'enjeu représenté par les provisions mathématiques des sinistres d'arrêts de travail est d'autant plus important qu'il est fonction des taux non-vie en vigueur au sein de l'environnement social. Avec la problématique des taux bas voire négatifs, les provisions techniques à constituer augmentent considérablement.

Parmi les contrats des branches du secteur interprofessionnel et les contrats sur mesure assurés par AG2R La Mondiale, l'un d'entre eux attire l'attention en termes d'arrêts de travail. Tout d'abord, le taux d'absentéisme de ce contrat atteint 15% contre 5,11% en moyenne nationale en 2019 [15]. De plus, il présente un chiffre d'affaires parmi les plus élevés. Enfin, ses garanties d'arrêts de travail – maintien de salaire, incapacité et invalidité – n'ont pas évolué depuis 2012. Nous souhaitons mesurer l'apport du provisionnement issu des méthodes d'apprentissage automatique pour ce contrat à la sinistralité importante.

Cadre méthodologique

Lorsqu'un sinistre est clos, les algorithmes disposent de la valeur exacte de la charge ultime à prédire. Dans ce cas, aucune provision n'est à constituer au titre du sinistre. En revanche, pour un sinistre toujours ouvert à la date d'arrêté comptable (non clos), la charge de sinistre disponible représente une charge moindre par rapport à l'ensemble des prestations attendues au titre de ce sinistre.

Lopez et al. (2015) [2] proposent une pondération des observations pour une meilleure qualité de prédiction dans le cadre du provisionnement des sinistres en assurance non-vie, en adaptant l'arbre de régression CART aux observations censurées. Le poids $w_{i,n}$ issu de la méthode *Inverse Probability of Censoring Weighting* (IPCW) est associé à chaque sinistre i au sein d'un échantillon de n sinistres. Ce poids correspond à l'inverse de l'estimation de la probabilité d'être non censuré, sachant que le sinistre i a une durée de vie de t. Lopez et al. (2015) [2] proposent d'estimer la probabilité d'être non censuré avec l'estimateur de survie de Kaplan-Meier.

En complément de l'application de l'algorithme CART pondéré, les versions pondérées des algorithmes plus élaborés *Random Forest* et *Gradient Boosting* seront mises en œuvre à des fins de comparaison.

La qualité de prédiction de la charge ultime de sinistres est mesurée avec la forme pondérée du coefficient de détermination de Pearson et de l'erreur carrée relative (R_w^2 et $WRSE$).

Afin d'être au plus près de la réalité, les algorithmes s'appuient sur des données historiques réelles d'arrêts de travail pour lesquels la charge ultime est disponible (*backtesting*). L'adaptabilité des modèles sera vérifiée à travers l'application à des données nouvelles n'ayant servi ni à la phase d'apprentissage ni à la phase de test des algorithmes mis en œuvre.

Les méthodes proposées ne permettent pas de prendre en compte les sinistres tardifs qui pourront faire l'objet d'une future étude.

Présentation des données

Les sinistres d'arrêts de travail en maintien de salaire, incapacité et invalidité survenus entre le 1er janvier 2012 et le 31 décembre 2020 sont disponibles pour le contrat de l'étude, avec les caractéristiques des sinistres, des assurés, des contrats de prévoyance et des provisions mathématiques. De nouvelles variables calculées, jugées utiles pour la modélisation, sont intégrées aux données : l'âge à la survenance de l'arrêt de travail, les durées totale et indemnisée du sinistre, l'ancienneté de l'assuré au sein du contrat de prévoyance et le salaire de référence annuel. L'effectif des entreprises, provenant de données publiques, a également permis d'enrichir les données. Puis, une fiabilisation des données est réalisée.

Les statistiques exploratoires ont permis d'arrêter des choix sur les données finales retenues. D'une part, la répartition des arrêts de travail par mois de survenance révèle une année de survenance 2020 atypique, liée à la pandémie du COVID-19. D'autre part, les sinistres en invalidité représentent moins de 2% de l'ensemble des arrêts de travail et leur taux de censure est très élevé. Nous avons donc peu de recul sur les sinistres en invalidité. A partir de ces constats, et afin d'assurer la qualité des prédictions, les sinistres survenus en 2020 et les sinistres en invalidité sont exclus du périmètre de l'étude.

L'hypothèse de censure est ensuite précisée. En effet, pour le portefeuille étudié, les sinistres en incapacité sont considérés, à dire d'expert, comme étant en cours de paiement lorsque le dernier règlement de prestations a été effectué dans les 4 mois précédant la date d'inventaire. Cependant, il n'y a pas de règle précise pour qualifier la censure d'un sinistre en maintien de salaire. Pour conserver une cohérence pour l'ensemble des arrêts de travail, la même règle est appliquée aux sinistres en maintien de salaire et incapacité dans le cadre de cette étude.

L'analyse statistique du portefeuille révèle une population très typée avec majoritairement des femmes, principalement non-cadres et vivant en couple, et percevant des salaires annuels peu élevés.

Préparation à la modélisation

Afin d'entrainer et calibrer les algorithmes de *machine learning* dans une démarche de *backtesting*, les arrêts de travail survenus entre 2012 et 2017 et observés à la date d'inventaire du 31 décembre 2017 sont sélectionnés. Les durées maximales contractuelles des sinistres maintien de salaire et incapacité étant respectivement de 12 mois glissants et de 1 095 jours, nous pouvons considérer que les sinistres seront clos à l'arrêté comptable du 31 décembre 2020, soit 1 095 jours après leur survenance. Le taux réel de censure est d'environ 1% pour le maintien de salaire et 12% pour l'incapacité.

Pour vérifier l'adaptabilité des modèles, l'échantillon de validation est constitué de sinistres survenus entre 2018 et 2019 et observés à la date d'inventaire du 31 décembre 2019. En appliquant l'hypothèse de censure retenue (4 mois), le taux de censure est de 12% pour le maintien de salaire et 31% pour l'incapacité.

Nous souhaitons conserver une cohérence entre les montants pour la prédiction de la charge ultime des arrêts de travail pour lesquels la durée peut s'étaler sur 3 ans. Pour cela, l'inflation à la date maximale d'observation du 31 décembre 2020 est prise en compte. La revalorisation des indemnités journalières est fonction de l'évolution du Salaire Minimum de Croissance (SMIC). Ainsi, les règlements des sinistres sont inflatés en fonction de l'évolution du SMIC. Les provisions mathématiques en incapacité sont également impactées par l'évolution du SMIC, et leur calcul intègre le taux technique en vigueur au 31 décembre 2020. Cependant, les salaires de ce portefeuille n'ayant pas été revalorisés depuis une dizaine d'année ne sont pas inflatés. Il aurait fallu émettre des hypothèses d'inflation pour les sinistres qui seront clos au-delà de l'année 2020 pour l'échantillon de validation. Ce point est noté comme étant un axe d'amélioration de l'étude.

Enfin, pour estimer les règlements futurs de l'échantillon de validation, nous avons recours aux coefficients de passage déterminés à partir de l'échantillon de *backtesting* avec la méthode de *Chain Ladder*, en cohérence avec le choix méthodologique réalisé à date sur ce portefeuille.

Application des algorithmes de *machine learning* retenus et interprétation des résultats

Les algorithmes CART, *Random Forest* et *Gradient Boosting* pondérés sont appliqués à l'échantillon de *backtesting*. Les données de la phase d'apprentissage correspondent à 70% de l'échantillon de *backtesting* et les données de la phase de test correspondent aux 30% restants.

Les algorithmes montrent que le montant des prestations réglées, le code de clôture (clos ou non clos) du sinistre et le montant de salaire annuel de référence semblent contenir de l'information pertinente pour la prédiction de la charge ultime des sinistres d'arrêts de travail du portefeuille. *Gradient Boosting* pondéré détecte également le mois de survenance et l'effectif des entreprises comme variables d'importance. La franchise contractuelle est sélectionnée par *Random Forest* pondéré.

Les indicateurs de performance attribuent la meilleure prédiction à *Gradient Boosting* pondéré.

	Tous les sinistres			Sinistres clos			Sinistres non clos		
	CART	***Random Forest***	***Gradient Boosting***	**CART**	***Random Forest***	***Gradient Boosting***	**CART**	***Random Forest***	***Gradient Boosting***
WRSE	18,2%	7,8%	5,4%	6,8%	1,3%	0,5%	37,4%	18,7%	13,7%
R^2w	81,8%	92,2%	94,6%	93,2%	98,7%	99,5%	62,6%	81,3%	86,3%

Tableau A : Indicateurs statistiques WRSE et R^2w issus de l'échantillon de test

Les résultats observés lors de la phase de test montrent également que *Gradient Boosting* pondéré réalise la meilleur prédiction, compte tenu des écarts les plus faibles entre les valeurs réelles et prédites.

	Valeur réelle	CART pondéré		*Random Forest* pondéré		*Gradient Boosting* pondéré	
		Prédiction	Ecarts observés	Prédiction	Ecarts observés	Prédiction	Ecarts observés
Tous les sinistres	78 423 690 €	77 244 845 €	-1,5%	78 502 709 €	0,1%	78 472 499 €	0,1%
Sinistres clos	65 146 608 €	67 559 356 €	3,7%	65 657 022 €	0,8%	65 248 791 €	0,2%
Sinistres non clos	13 277 082 €	9 685 489 €	-27,1%	12 845 687 €	-3,2%	13 223 709 €	-0,4%

Tableau B : Echantillon de test – Charge ultime globale ventilée par code de clôture

Pour s'assurer de l'adaptabilité de l'algorithme aux données nouvelles, les observations sont confrontées aux prédictions obtenues avec *Gradient Boosting* pondéré sur l'échantillon de validation, en se raccrochant aux observations à la date d'inventaire du 31 décembre 2020. Les résultats obtenus ne suivent pas la tendance des résultats issus de l'échantillon de *backtesting* qui sert de prérequis. On constate un sur-provisionnement :

- En maintien de salaire, les règlements futurs estimés sont environ 12 fois supérieurs à l'attendu.
- En incapacité, l'algorithme surestime les règlements futurs de plus de 31**%**.

Les variables d'importance détectées par les algorithmes permettent d'expliquer ces écarts :

- La surévaluation des provisions vient essentiellement de l'hypothèse de censure (4 mois) qu'il faudra rechallenger afin d'être *best estimate*.
- Une saisonnalité spécifique au portefeuille étudié, une précarité liée aux salaires faibles, une franchise contractuelle peu élevée et une répercussion des absences plus importante dans les petites et moyennes entreprises sont autant de facteurs qui pourraient expliquer la sinistralité élevée du portefeuille.

Application opérationnelle : impacts en termes de gestion des risques

Dans le cadre réglementaire imposé par la directive de Solvabilité 2 [1], le modèle *Gradient Boosting* pondéré sert d'outil à la fonction Gestion des Risques, et plus précisément à la Fonction Actuarielle, pour challenger les méthodes utilisées, afin d'émettre un avis éclairé sur la politique de provisionnement :

- Tout d'abord, la pertinence de la réduction de l'hypothèse de censure selon le risque a été vérifiée à travers une étude de sensibilité. Bien que présentant des limites, les résultats conduisent à corriger le sur-provisionnement observé avec l'hypothèse initiale de censure. De plus, la réduction des marges de prudence permet de limiter l'impact fiscal de la taxation des boni de provisionnement au sein de l'environnement social.
- Ensuite, la construction d'une loi de maintien d'expérience en maintien de salaire s'est avérée pertinente, même avec l'absence de comparaison avec une table dédiée du BCAC.
- Enfin, la construction d'une loi de maintien d'expérience en incapacité s'est également avérée plus adaptée que l'utilisation de la table de maintien en incapacité du BCAC. La loi de maintien d'expérience en incapacité permettra d'ajuster les provisions mathématiques pour le contrat étudié.

Les apports se mesurent également au niveau des propositions émises pour les politiques de souscription (études de rentabilité et d'opportunités à la suite de possibles enrichissements de données, introduction de lois d'incidence d'expérience, choix du pilotage le plus adapté pour le contrat) et de réassurance (pour pallier une dérive de la sinistralité).

A partir des statistiques et des enseignements du modèle prédictif et explicatif, un premier diagnostic des absences d'arrêts de travail est établi. Il pourra conduire à la mise en œuvre d'une politique efficace de pilotage de l'absentéisme. Les résultats d'une telle démarche pour l'assureur conduiront à la réduction de l'exigence en capital pour les sinistres d'arrêts de travail en environnement Solvabilité 2.

Conclusion

L'objectif du mémoire était de proposer une méthode alternative de provisionnement des arrêts de travail pour un contrat spécifique présentant une sinistralité importante.

Les méthodes de *machine learning* ont permis de prendre en compte de nouvelles variables dans le cadre d'une modélisation non paramétrique du provisionnement. Cette modélisation, en présence de données censurées, s'appuie sur des données historiques réelles d'arrêts de travail pour lesquels la charge ultime était disponible (*backtesting*).

Une application opérationnelle illustre l'apport du modèle prédictif et explicatif en termes de gestion des risques pour un pilotage ciblé du contrat. D'une part, la fonction Gestion des Risques et plus particulièrement la Fonction Actuarielle induites par la directive de Solvabilité 2 pourront se faire leur propre opinion sur les méthodes actuelles de provisionnement et proposer des mesures concrètes d'améliorations pour le suivi prospectif des arrêts de travail. Les apports se mesurent également au niveau des propositions émises pour les politiques de souscription et de réassurance. D'autre part, face à un premier état des lieux des absences liées aux arrêts de travail, de nouveaux outils de gestion des arrêts de travail pourraient être envisagés et ouvrir la voie à un pilotage approprié de l'absentéisme.

Malgré les imperfections du modèle prédictif et explicatif, son apport indéniable à la gestion de risques est d'offrir une meilleure visibilité du coût du risque et une meilleure interprétation du provisionnement.

Executive summary

Introduction

Insurance organizations are required by law to fund future benefits from a work interruption claim in the process of being compensated until the end of the interruption. The theme of reserving is even more important as it represents a requirement of the Solvency 2 directive [1], which demands a best estimate reserving. The traditional approach used is based on the Chain Ladder method for claims in course of settlement reserving and the parametric calculation of mathematical actuarial reserves. However, faced with a high loss experience of a contract (continuous increase in claims, population at risk with specificities, etc.), it is advisable to consider alternative methods of reserving calculation.

In recent decades, machine learning algorithms have proven to be very useful and suitable for analyzing data in the insurance world. In their article "Censored tree regression with insurance applications", Lopez and al. (2015) [2] propose the prediction of the ultimate cost of a claim, closed or still open, using regression trees. The originality of this method is the application of algorithms to censored observations, by the introduction of a weighting to correct the bias induced by the censorship.

The objective of the master thesis is to assess the relevance of machine learning methods to challenge the usual ones.

Background

For several years, expenses related to work interruptions have been rising steadily. The Department of Research, Studies, Evaluation and Statistics estimates the amount of daily allowances paid by Social Security at 15.7 billion euros in 2019 [3]. A continuous increase in sick leaves has been observed since 2014. The explanatory factors are an aging working population, long sick leaves for the oldest or most vulnerable workers, etc.

For insurers, the issue represented by the mathematical provisions for work interruption claims is even more important as it depends on the non-life rate in use within the social environment. With the low or even negative rates problem, the technical reserving to be constituted increase considerably.

Among the contracts of the inter-professional sector and the tailor-made contracts insured by AG2R La Mondiale, one of them draws attention in terms of work interruptions. First, the absenteeism rate for this contract reached 15% against 5.11% on the national average in 2019 [15]. In addition, it has one of the highest turnovers. Finally, its work interruption guarantees – salary continuance, incapacity, and invalidity – have not changed since 2012. We would be measuring the contribution of machine learning reserving methods for this contract with a significant loss experience.

Methodology framework

When a claim is closed, the algorithms have the exact value of the ultimate cost to predict. In this case, there are no reserving calculation for the claim. On the other hand, for a claim still open on the inventory date, the available payments represent a lower cost compared to all the expected payments for this claim.

Lopez and al. (2015) [2] propose a weighting of the observations for a better quality of prediction within the non-life insurance reserving framework, by adapting the CART regression tree to the censored observations. The weight $w_{i,n}$ from the Inverse Probability of Censoring Weighting (IPCW) method is associated with each claim i in a sample of n claims. This weight is the inverse of the estimate of the probability of being uncensored, knowing that the claim i has a lifetime of t. Lopez and al. (2015) [2] propose to estimate the probability of being uncensored with the Kaplan-Meier survival estimator.

In addition to the application of the weighted CART algorithm, weighted versions of the more elaborate Random Forest and Gradient Boosting algorithms will be implemented for comparison purposes.

The predictive quality of the ultimate claim cost is measured with the weighted form of the Pearson coefficient of determination and the relative square error (R_w^2 and $WRSE$).

To be as close to reality as possible, the algorithms are based on real historical data from work interruptions for which the ultimate cost is available (backtesting). The adaptability of the models will be verified through application to new data that has not been used for either the learning phase or the testing phase of the implemented algorithms.

The suggested methods do not allow late claims to be considered, which may be the subject of a future study.

Data presentation

Work interruption claims in salary continuance, incapacity and invalidity occurring between January 1, 2012 and December 31, 2020 are available for the study contract, with the characteristics of claims, policyholders, insurance contracts and mathematical actuarial reserves. New calculated features, considered useful for modeling, are integrated into the data: age at work interruption occurrence, total and compensation duration of the claim, seniority of the insured workers within the insurance contract and the annual reference salary. The companies' headcount, from public data, also enriched the database. Then, data reliability is achieved.

The exploratory statistics made it possible to make choices on the final data retained. On the one hand, the breakdown of work interruptions by month of occurrence reveals an atypical year of occurrence in 2020, linked to the COVID-19 pandemic. On the other hand, invalidity claims represent less than 2% of all work interruptions and their censorship rate is very high. We therefore have little perspective on invalidity claims. Based on these findings, and to ensure the predictions' quality, claims occurring in 2020 and invalidity claims are excluded from the scope of the study.

The censorship hypothesis is then clarified. In fact, for the studied portfolio, incapacity claims are considered, according to expert judgment, as being in payment when the last payment has been made in the 4 months preceding the inventory date. However, there is no precise rule for qualifying censorship for a claim in salary continuance. To maintain consistency for all work interruptions of the study, the same rule of 4 months is applied to salary continuance and incapacity claims.

The statistical analysis of the portfolio reveals a very typical population with predominantly women, mainly non-executive, living as a couple, and receiving low annual salaries.

Modeling preparation

To train and calibrate the machine learning algorithms in a backtesting process, work interruptions which occurred between 2012 and 2017 and observed on the inventory date of December 31, 2017 are selected. Since the maximum contractual durations for salary continuance and incapacity claims are respectively 12 rolling months and 1,095 days, we can consider that the claims will be closed at the inventory date of December 31, 2020, i.e., 1,095 days after their occurrence. The real censorship rate is around 1% for salary continuance and 12% for incapacity.

To check the models' adaptability, the validation sample is made up of claims occurring between 2018 and 2019, on the inventory date of December 31, 2019. By applying the censorship hypothesis adopted (4 months), the censorship rate is 12% for salary continuance and 31% for incapacity.

It is important to maintain consistency between the amounts for the predicted ultimate cost of work interruptions for which the duration can be spread over 3 years. For this, inflation at the maximum observation date of December 31, 2020 is considered. The revaluation of daily allowances depends on the evolution of the Minimum Growth Wage. Thus, claims settlements are inflated according to the evolution of the Minimum Growth Wage. Mathematical actuarial provisions for incapacity are also concerned, and their calculation incorporates the technical rate retained on December 31, 2020. However, salaries not having been revalued for ten years for this portfolio are not inflated. Inflation assumptions should have been made for claims that will close beyond 2020 for the validation sample. This point is noted as an area for improvement in the study.

Finally, to estimate the future settlements of the validation sample, we use the coefficients determined from the backtesting sample with the Chain Ladder method, in line with the methodological choice made to date on this portfolio.

Application of the selected machine learning algorithms and results' interpretation

The weighted CART, Random Forest and Gradient Boosting algorithms are applied to the backtesting sample. The data from the learning phase corresponds to 70% of the backtesting sample and the data from the test phase corresponds to the remaining 30%.

The algorithms show that the amount of benefits paid, the closing code (closed or not closed) and the reference annual salary amount seem to contain relevant information for the ultimate cost prediction of work interruption claims. Weighted Gradient Boosting also detects month of occurrence and company sizes as important features. The contractual franchise is also selected by weighted Random Forest.

The performance indicators give the best prediction to weighted Gradient Boosting.

	All claims			**Closed claims**			**Unclosed claims**		
	CART	***Random Forest***	***Gradient Boosting***	**CART**	***Random Forest***	***Gradient Boosting***	**CART**	***Random Forest***	***Gradient Boosting***
WRSE	18,2%	7,8%	5,4%	6,8%	1,3%	0,5%	37,4%	18,7%	13,7%
R^2w	81,8%	92,2%	94,6%	93,2%	98,7%	99,5%	62,6%	81,3%	86,3%

Table A: WRSE and R^2w statistical indicators from the test sample

The results observed during the test phase also show that weighted Gradient Boosting achieves the best prediction, with the smallest differences between the real and predicted values.

	Real value	Weighted CART		Weighted Random Forest		Weighted Gradient Boosting	
		Prediction	Observed difference	Prediction	Observed difference	Prediction	Observed difference
All claims	78 423 690 €	77 244 845 €	-1,5%	78 502 709 €	0,1%	78 472 499 €	0,1%
Closed claims	65 146 608 €	67 559 356 €	3,7%	65 657 022 €	0,8%	65 248 791 €	0,2%
Unclosed claims	13 277 082 €	9 685 489 €	-27,1%	12 845 687 €	-3,2%	13 223 709 €	-0,4%

Table B: Test sample – Ultimate cost by closing code

To ensure the algorithm's adaptability to new data, the observations are compared to the predictions obtained with weighted Gradient Boosting on the validation sample, by hanging on to the observations on the inventory date of December 31, 2020. The obtained results do not follow the trend of the results from the backtesting sample which serves as a prerequisite. There is over-reserving:

- In salary continuance, the estimated future settlements are about 12 times higher than expected.
- In incapacity, the algorithm overestimates future settlements by more than 31%.

The important features detected by the algorithms explain these differences:

- The reserving overvaluation comes mainly from the censorship assumption (4 months) that will have to be re-challenged to be best estimate.
- Specific seasonality to the studied portfolio, precariousness linked to low wages, a low contractual franchise, and a greater impact of absences in small and medium-sized enterprises are all factors which could explain the high loss experience of the portfolio.

Operational application: impacts in terms of risk management

In the regulatory framework imposed by the Solvency 2 directive [1], the weighted Gradient Boosting model serves as a tool for the Risk Management function, and more precisely for the Actuarial Function, to challenge the usual methods and to have an informed opinion on the reserving policy:

- First, the relevance of reducing the censorship hypothesis according to risk was verified through a sensitivity study. Although presenting limits, the results lead to correcting the over-evaluation of reserving observed with the initial censorship hypothesis. In addition, the reduction in the margins of prudence makes it possible to limit the fiscal impact of the taxation of reserving bonuses in the social environment.
- Next, the construction of an experience table in salary continuance proved to be relevant, even with the absence of a comparison with a dedicated BCAC table.
- Finally, the construction of an incapacity experience table has also proven to be more suitable than the use of the BCAC's incapacity table. The incapacity experience table will make it possible to adjust the mathematical actuarial reserves for the studied contract.

The contributions are also measured at the level of the proposals made for the underwriting policies (profitability and opportunity studies following possible data enrichment, introduction of experience incidence tables, most suitable management choice for the contract) and reinsurance (to compensate for a drift in claims).

Based on statistics and lessons learned from the predictive and explanatory model, an initial diagnosis of work interruption absences is established. It could lead to the implementation of an effective policy for absenteeism management. The results of such an approach for the insurer will lead to a reduction in the capital requirement for work interruption claims in the Solvency 2 environment.

Conclusion

The objective of the master thesis was to propose an alternative reserving method for work interruptions for a specific contract with a significant loss experience.

Machine learning methods have made it possible to consider new variables in the context of non-parametric modeling of reserving. This modeling, in presence of censored data, is based on real historical data of work interruptions for which the ultimate cost was available (backtesting).

An operational application illustrates the contribution of the predictive and explanatory model in terms of risk management for targeted management of the contract. On the one hand, the Risk Management function and more particularly the Actuarial Function induced by the Solvency 2 directive will be able to form their own opinion on the current methods of reserving and propose concrete improvement measures for the prospective monitoring of work interruptions. Contributions are also measured at the level of the proposals made for underwriting and reinsurance policies. On the other hand, faced with an initial inventory of absences linked to work interruptions, new work interruptions management tools could be considered and pave the way for an appropriate management of absenteeism.

Despite the imperfections of the predictive and explanatory model, its undeniable contribution to risk management is to provide a better visibility of the cost of risk and a better interpretation of reserving.

Lexique

Argmin : argument minimum d'une fonction. Il représente la valeur de la variable qui minimise la fonction concernée.

Biais : procédé qui engendre des erreurs dans les résultats d'une étude.

Boni : excédent des recettes sur les dépenses.

Bruit : signal parasite lors d'une mesure.

Big data : ensemble très volumineux de données, difficile à traiter par des outils classiques de gestion de bases de données ou de gestion de l'information.

Compte de résultats : document comptable présentant l'ensemble des produits et des charges d'une société durant un exercice comptable. Comme le bilan et les annexes, il fait partie des états financiers des entreprises.

Convergence d'un algorithme : au fur et à mesure des itérations, la sortie d'un algorithme se rapproche de plus en plus d'une valeur spécifique.

Convexité : une fonction est convexe lorsque sa courbe représentative se trouve au-dessus de ses tangentes.

Descente de gradient : lorsqu'une résolution analytique n'est pas possible (coût, complexité), elle est approximée avec une approche itérative, jusqu'à la convergence.

Différentiabilité : existence d'un développement limité à l'ordre 1 en un point.

Ecart-type : mesure de la dispersion d'un ensemble de valeurs autour de leur moyenne. Plus il est faible, plus la population est homogène.

Espérance mathématique : moyenne des valeurs prises par la réalisation d'une variable aléatoire, pondérées par leur probabilités respectives.

Estimation empirique : estimation s'appuyant uniquement sur l'observation, et non une théorie ou le raisonnement.

Exigence en capital : montant de fonds propres suffisant pour couvrir le profil de risque d'une entreprise. Le Capital de Solvabilité Requis (ou SCR pour *Solvency Capital Requirement*) et le Minimum de Capital Requis (MCR) sont des indicateurs requis par la directive de Solvabilité 2 pour mesurer le besoin de fonds propres.

Fonction coût-complexité : fonction qui estime le nombre d'opérations élémentaires effectuées par un algorithme (coût) et mesure sa performance en termes d'utilisation de la mémoire et de la vitesse d'exécution (complexité).

Fonction de répartition : fonction en escalier. Elle représente la probabilité que les valeurs prises par une variable aléatoire soient strictement inférieures à une valeur donnée.

Fonction de perte : fonction qui évalue l'écart entre les prédictions réalisées par un algorithme et les valeurs réelles des observations utilisées pendant l'apprentissage.

Gouvernance : ensemble de décisions, règles et pratiques visant à assurer le fonctionnement optimal d'une organisation, ainsi que les organes structurels chargés de les formuler, de les mettre en œuvre et d'en assurer le contrôle.

Gradient : vecteur qui caractérise la variabilité d'une fonction au voisinage d'un point.

Hétérogénéité : caractéristique d'une population pour laquelle l'écart-type est élevé.

Homogénéité : caractéristique d'une population pour laquelle l'écart-type est faible.

Identiquement distribués : (variables aléatoires) qui suivent la même loi de probabilité.

Insuffisances d'un modèle : ensemble des résidus issus d'un modèle.

Loi de distribution : description du comportement aléatoire d'un phénomène dépendant du hasard.

Loi d'incidence : loi qui mesure le nombre d'individus concernés par un événement pendant une période donnée (population incidente). Ce nombre est rapporté à la population dont ces individus sont issus (population cible) selon des critères prédéfinis.

Loi de maintien : loi qui mesure les effectifs de personnes pour lesquelles un événement est avéré selon des critères spécifiques et qui mesure l'évolution des effectifs pendant une période donnée.

Mali : déficit de recettes sur les prévisions.

Médiane : valeur qui sépare la moitié inférieure de la moitié supérieure d'un échantillon.

Percentile : chacune des 99 valeurs qui divisent un échantillon de données triées en 100 parties égales.

Pondération : attribution à chacune des valeurs qui composent une statistique, d'un poids différent et relatif à leur importance dans la considération de la statistique.

Quartile : chacune des 3 valeurs qui divisent un échantillon de données triées en 4 parties égales.

Rentabilité : capacité d'un investissement à procurer un bénéfice.

Résidu : erreur observée définie comme étant la différence entre une valeur observée et une valeur estimée par un modèle de régression.

Troncature : on parle de troncature lorsqu'une donnée n'est observable qu'à partir d'une date ultérieure à la date de survenance.

Variable aléatoire : variable dont la valeur est déterminée après un tirage aléatoire, en théorie des probabilités.

Variance : moyenne des carrés des écarts entre les valeurs observées et les valeurs estimées ou prédites.

Vecteur aléatoire : valeur aléatoire multidimensionnelle.

Liste des figures

Liste des tableaux

Annexes

Annexe A - Méthode de rééchantillonnage : validation croisée

Il est essentiel de vérifier qu'un modèle construit avec des méthodes d'apprentissage automatique n'est pas sujet au sur-apprentissage, pour s'assurer de son adaptabilité à de nouvelles données. On mesure la performance d'un modèle en classification par la proportion de points mal étiquetés, et en régression par la moyenne des erreurs quadratiques. Pour procéder à l'échantillonnage d'un jeu de données, la première idée consiste à scinder ce dernier en deux échantillons distincts : l'échantillon d'apprentissage et l'échantillon de test. La limite de cette démarche est de ne pas utiliser l'entièreté des données pour entrainer le modèle. Le même constat est établi pour la phase de test. Or, plus il y a de données disponibles pour l'apprentissage, meilleurs sont les résultats de prédiction. L'estimation de la performance d'un modèle à partir d'un tel jeu de test pourrait être biaisée.

Pour pallier cette limite, la validation croisée permet l'utilisation de l'intégralité du jeu de données non seulement pour l'apprentissage, mais aussi pour le test. Pour la mettre en œuvre :

- Le jeu de données est découpé en k parties (*folds* en anglais), à peu près égales ;
- Chacune des parties, tour à tour, sera utilisée comme jeu de test ;
- Les $(k-1)$ autres parties seront utilisées ensemble pour la phase d'apprentissage.

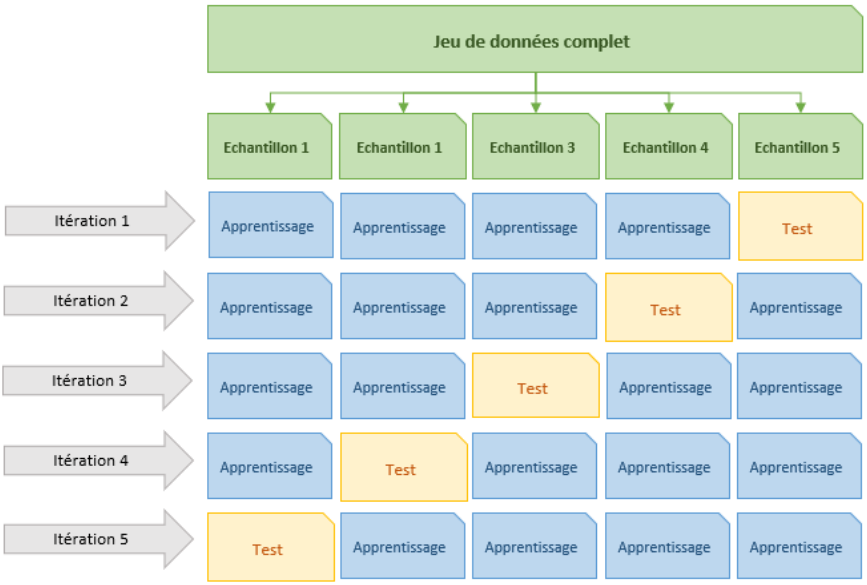


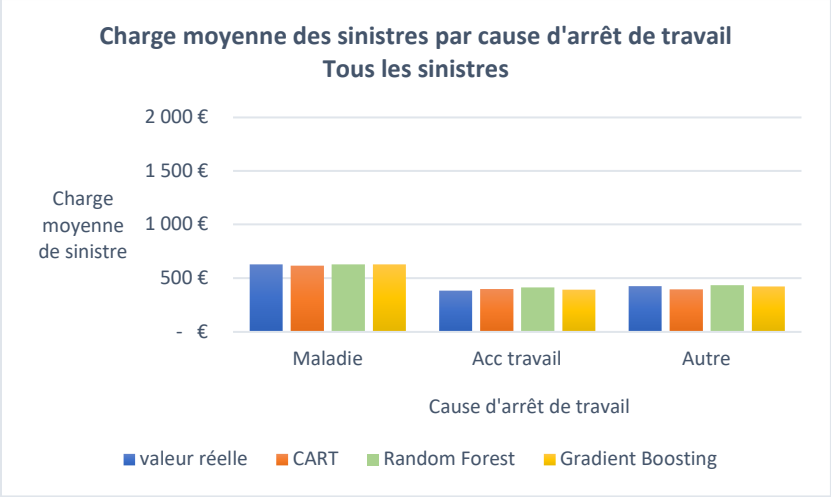
Illustration d'une validation croisée, avec *k=5*

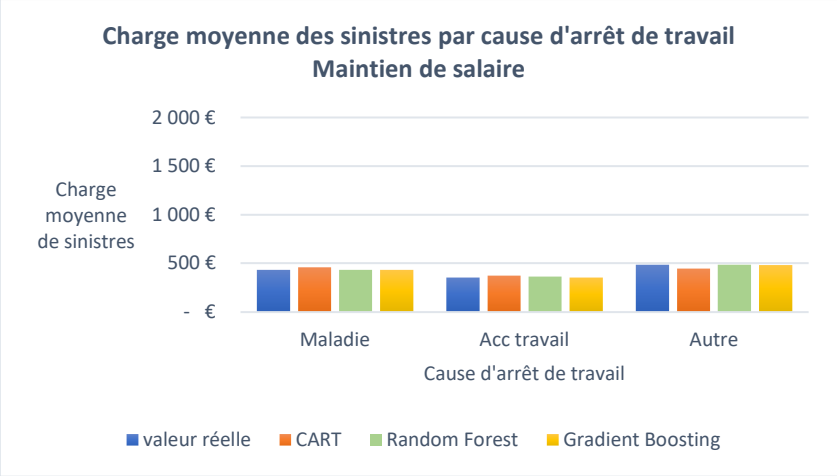
A la fin de la validation croisée, chaque point aura servi une fois dans un jeu de test et une fois dans un jeu d'apprentissage. A chaque tour, le principe de ne pas valider le modèle sur des données de l'apprentissage aura été respecté.
Il existe deux approches pour analyser la performance d'un modèle dans le cas d'un rééchantillonnage par validation croisée. La première approche consiste à évaluer les prédictions faites sur l'ensemble des données, et la seconde détermine la moyenne des performances obtenues sur les k-*folds*.

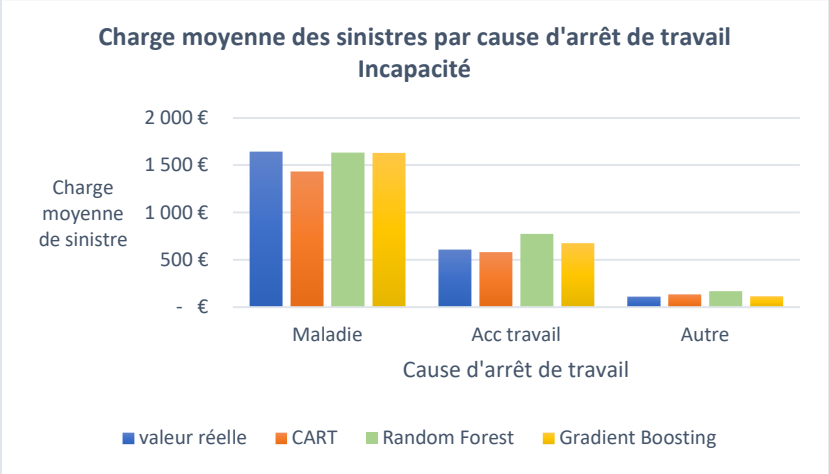
Le cas particulier d'une validation croisée où $k = n$ (n est le nombre de points dans l'échantillon) s'appelle la validation croisée *leave-one-out*. Le jeu d'entrainement est presqu'aussi important en volume que le jeu complet. Lorsqu'elle est utilisée, cette technique permet d'avoir une performance avec le plus faible biais. En revanche, sa mise en œuvre augmente fortement le temps de calcul. En pratique, on utilise $k = 5$ ou $k = 10$ pour la validation croisée.

Annexe B – Echantillon de test : charge moyenne des sinistres

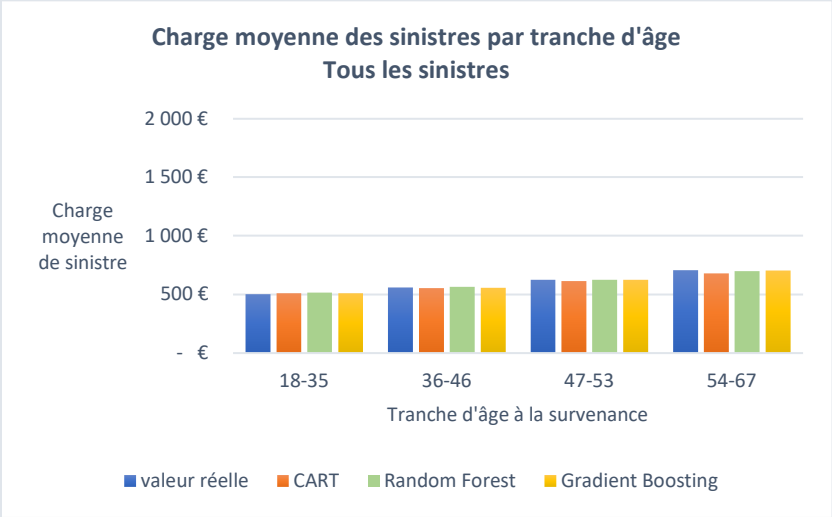
- Par cause d'arrêt de travail

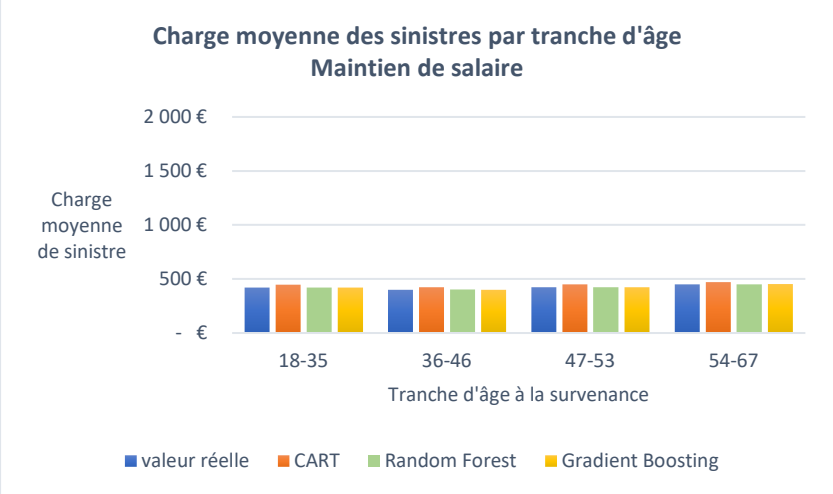


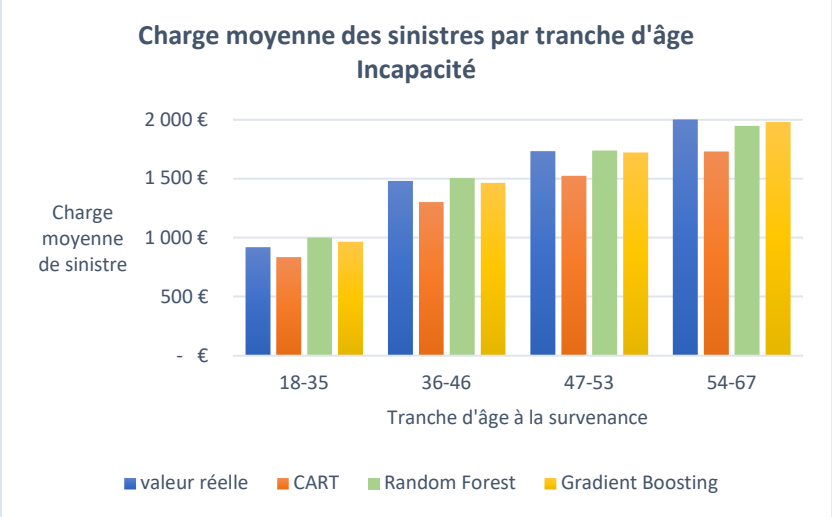




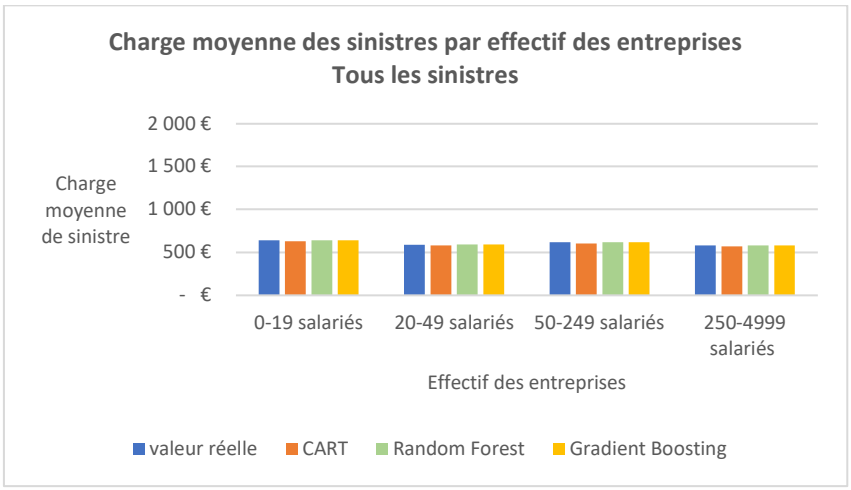
- Par tranche d'âge à la survenance

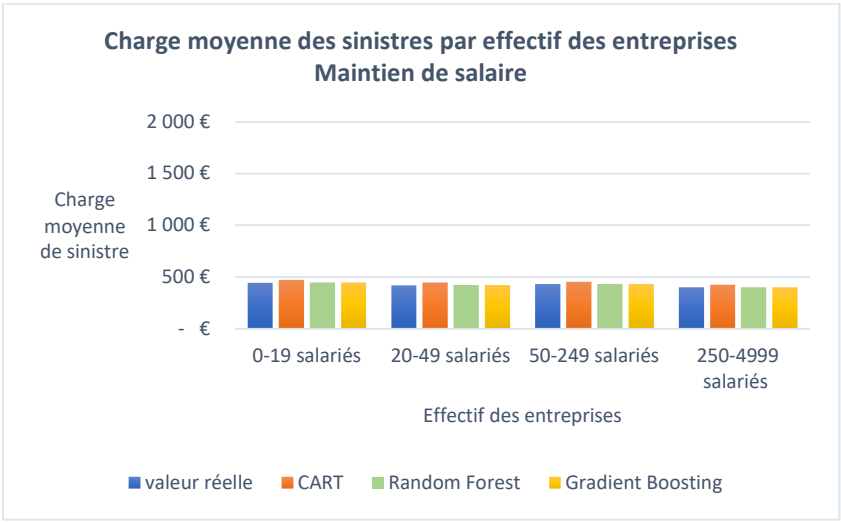


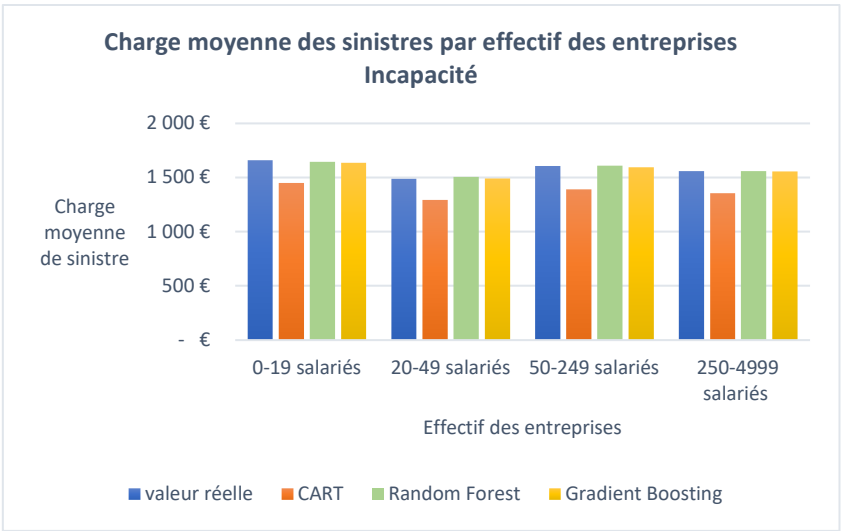




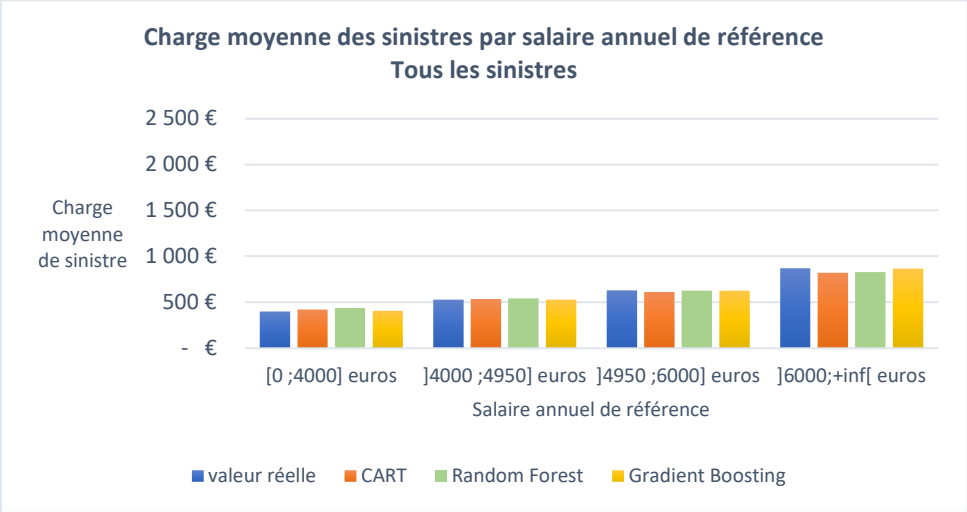
- Par effectif des entreprises des salariés

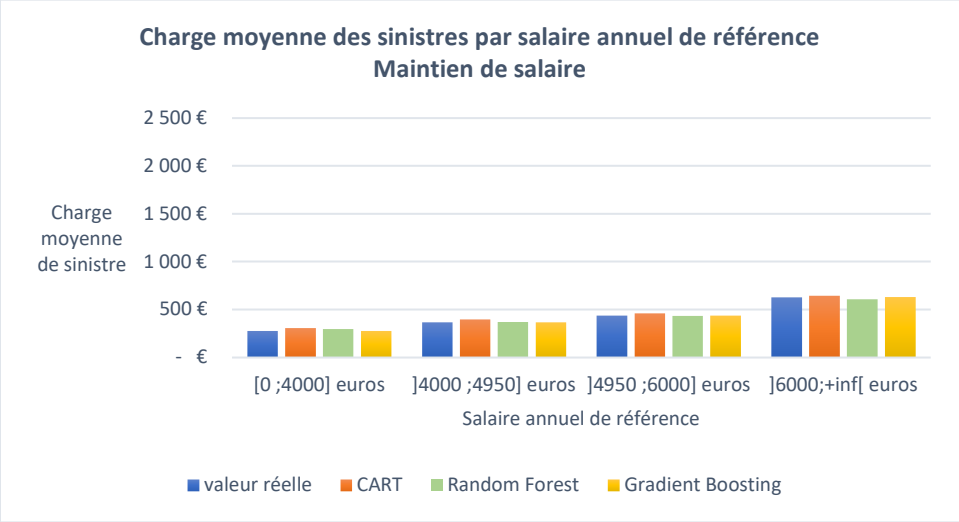


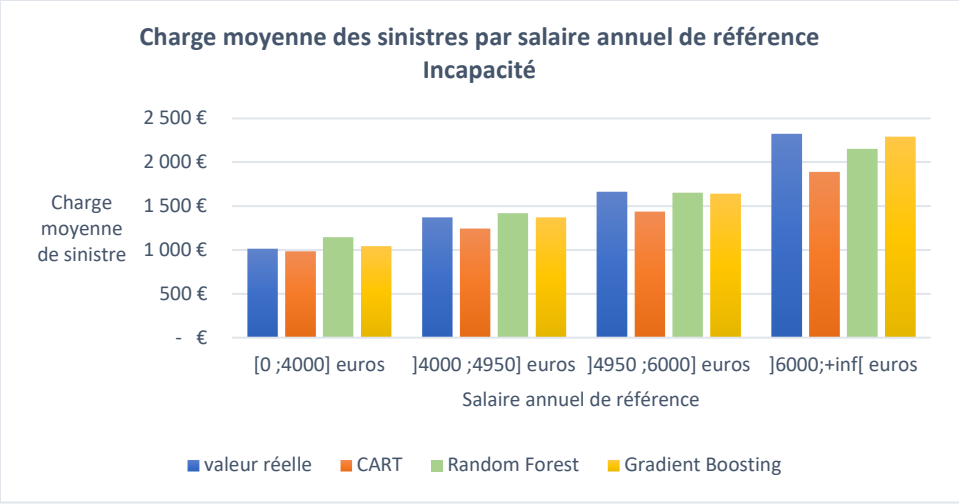




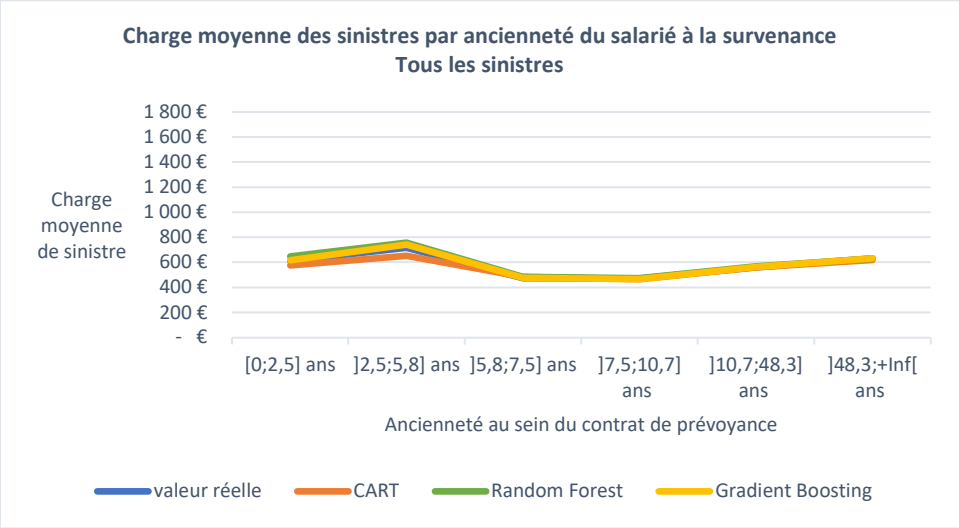
- Par salaire annuel de référence

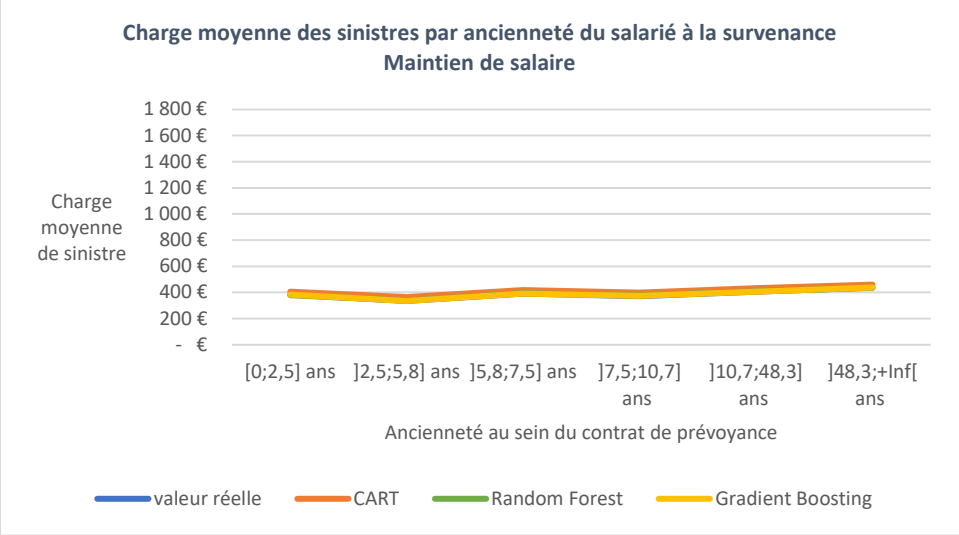


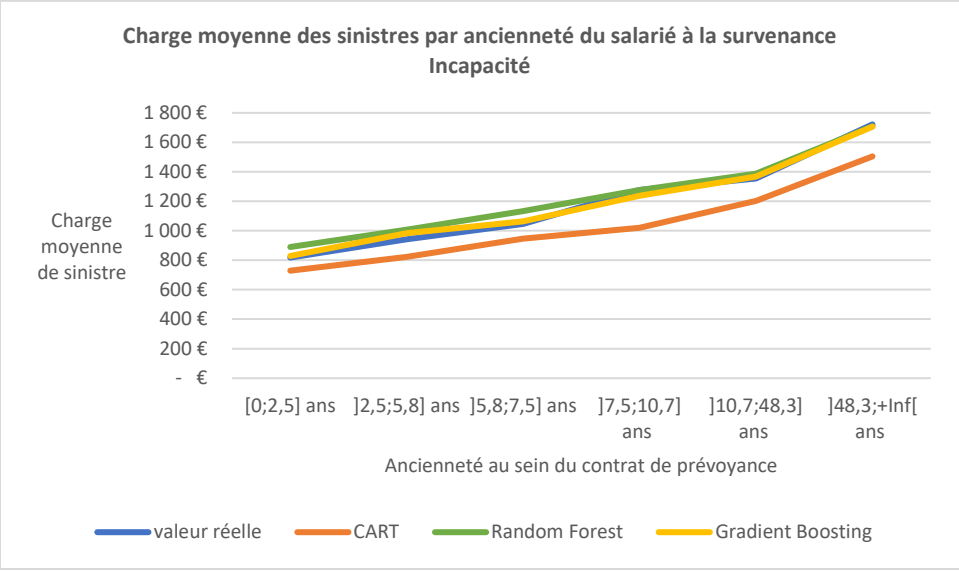




- Par ancienneté du salarié au sein du contrat de prévoyance, à la survenance de l'arrêt de travail







Annexe C – Résultats de CART et *Random Forest* pondérés sur l'échantillon de validation

	Vision à fin 2019	Vision à l'ultime avec CART pondéré		Vision à l'ultime avec *Random Forest* pondéré	
		Provisions à l'ultime	Charge à l'ultime	Provisions à l'ultime	Charge à l'ultime
Prestations totales versées + PM	**75 044 821 €**	9 240 810 €	**84 285 631 €**	17 515 593 €	**92 560 414 €**
Maintien de salaire	**37 936 573 €**	10 673 388 €	**48 609 961 €**	5 467 382 €	**43 403 955 €**
Sinistres clos	32 598 199 €	2 485 250 €	35 083 449 €	367 173 €	32 965 372 €
Sinistres non clos	*5 338 374 €*	*8 188 138 €*	*13 526 512 €*	*5 100 209 €*	*10 438 583 €*
Incapacité	**37 108 248 €**	-1 432 578 €	**35 675 670 €**	12 048 211 €	**49 156 459 €**
Sinistres clos	8 497 057 €	479 261 €	8 976 318 €	639 700 €	9 136 757 €
Sinistres non clos + PM	*28 611 191 €*	*-1 911 839 €*	*26 699 352 €*	*11 408 511 €*	*40 019 702 €*