Mémoire présenté le :

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

Par : GERBOUIN Raphaël

Titre : Conception d'un tarificateur prévoyance destiné au marché des TNS : Approche GLM et apprentissage supervisé.

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

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

Membre présents du jury de l'Institut des Actuaires — signature

Membres présents du jury de l'ISFA

Entreprise :

Nom : ACTELIOR

Signature :

Directeur de mémoire en entreprise :

Nom : M. Emmanuel GINESTE

Signature :

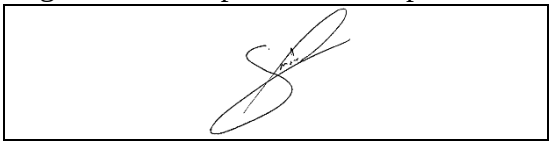
Invité :

Nom :

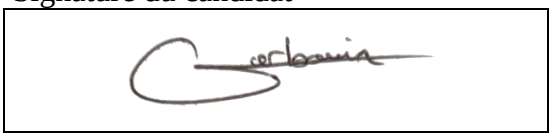
Signature :

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

Signature du responsable entreprise



Signature du candidat



Mémoire présenté devant l'Institut des Actuaires

Année 2021 - 2022

Conception d'un tarificateur prévoyance destiné au marché des TNS : Approche GLM et apprentissage supervisé

Raphaël GERBOUIN

Responsable ACTELIOR : Monsieur Emmanuel GINESTE

Responsable ISFA : Madame Anne EYRAUD-LOISEL

Résumé

Mots clés : Travailleur non salarié, prévoyance individuelle, arrêt de travail, tarificateur, modèle linéaire généralisé, arbre CART, forêt aléatoire, *gradient boosting*, apprentissage supervisé, fréquence, coût moyen

La population des TNS (Travailleurs Non Salariés) est une population spécifique. Elle recouvre des secteurs d'activité multiples et regroupe des catégories professionnelles diverses. Elle se compose alors de profils disparates rendant sa tarification particulière par rapport à celle d'une autre population. Cette population est en perpétuelle augmentation et reste peu couverte en cas d'arrêt de travail. En effet, il a été constaté que les non-salariés présentent un niveau de couverture en prévoyance inférieur à celui des salariés. Une prévoyance individuelle est alors nécessaire afin d'obtenir une couverture optimale mais elle peut présenter un coût conséquent pour le TNS. La loi dite « Madelin » réduit ce coût en permettant de déduire les cotisations du revenu imposable. Tous ces éléments contribuent à faire du marché de la prévoyance des TNS, un marché à fort potentiel de développement.

Le choix de la tarification s'est porté sur la décomposition de la charge ultime des sinistres en fréquence et en coût moyen d'un sinistre. Le modèle de fréquence est alors construit sur l'ensemble du portefeuille tandis que le modèle de coût moyen est établi sur la base des données sinistrées. Le modèle de coût moyen a nécessité d'être adapté au risque arrêt de travail. Le modèle développé s'avère finalement être un modèle de durée moyenne en arrêt à laquelle est multiplié le montant des indemnités journalières.

Pour effectuer cette tarification, différents algorithmes ont été implémentés. Tout d'abord, la méthode traditionnelle du modèle linéaire généralisé a été développée. Cette méthode se révèle relativement simple à implémenter ainsi qu'à interpréter. Toutefois, elle repose sur des hypothèses relativement fortes qui n'ont pas été satisfaites avec le portefeuille étudié, notamment sur la normalité des résidus. Cette observation a nourri l'intérêt de développer un modèle reposant sur d'autres algorithmes moins restrictifs en termes d'hypothèses et de paramètres. Les algorithmes d'apprentissage supervisé, méthodes non paramétriques, ont alors été présentés puis appliqués. Tout d'abord, l'arbre CART a été implémenté. Il s'agit de l'algorithme qui a fourni les moins bonnes métriques d'erreur. Des méthodes ensemblistes ont ensuite été présentées, elles reposent sur la combinaison de plusieurs arbres CART que ce soit par agrégation pour la forêt aléatoire ou de manière adaptative pour le *gradient boosting*. Finalement, le meilleur modèle au sens des métriques d'erreur est celui obtenu par *gradient boosting*. Cet algorithme a donc été sélectionné pour construire les bases techniques de l'outil de tarification.

Abstract

Key words : Self-employed person, individual soical protection, work stoppage, pricing tool, generalised linear model, classication and regression tree, random forest, gradient boosting, machine learning, frequency, average cost

The self-employed persons are a specific population. It covers multiple business lines and regroups various occupational categories. This population is then composed of disparate profils which make the pricing particular compared to the one considering another population. This population is constantly growing and remains poorly covered in case of work stoppage. In fact, self-employed persons have a lower level of pension cover than employees. An individual social protection is then necessary in order to obtain an optimal coverage nonetheless the cost of this additional protection can discourage some self-employed persons from getting one. The so-called "Madelin" law reduces this cost by allowing the premiums to be deducted from taxable income. All these elements contribute to making the self-employed persons's social protection market, a market with high development potential.

The choice of pricing focused on the decomposition of the total cost of claims into the frequency and the average cost of a claim. The frequency model is then built on the entire portfolio whereas the average cost model is established on the data with claims. The average cost model needed to be adapted to the risk of work stoppage. The developed model ultimately turns out to be a model of average duration in work stoppage at which is multiplied the amount of daily allowances.

To perform this pricing, different algorithms have been implemented. First of all, the traditional method of generalised linear model was developped. This method turns out to be relatively simple to compute as well as to interpret. However, it rests on relatively strong hypothesis that have not been satisfied when applied to this portfolio, in particular on the normality of the residuals. This observation has fueled the interest of developing a model based on other algorithms that are less restrictive in terms of hypothesis and parameters. Machine learning algorithms, non-parametric methods, were then presented and computed. First, the CART algorithm was implemented. This is the algorithm that provided the worst error metrics. Set methods were then presented, they are based on the combinaison of several decision trees either by agregation for the random forest or adaptatively for the gradient boosting. Eventually, the best model in the sense of the error metrics is the one computed by gradient boosting. Therefore, this algorithm was selected to build the technical bases of the pricing tool.

Remerciements

Je souhaite tout d'abord remercier M. Emmanuel Gineste, mon directeur de mémoire à Actélior, pour sa disponibilité et son aide dans la direction de mes recherches, l'organisation générale de mon travail ainsi que pour ses différentes relectures.

Je tiens à remercier M. David Echevin, directeur général d'Actélior, pour m'avoir offert l'opportunité de réaliser mon alternance au sein du cabinet. Également et surtout pour ses efforts entrepris afin que mon alternance se déroule dans les meilleures conditions.

Je souhaite également remercier Mme. Anne Eyraud-Loisel, ma tutrice à l'ISFA, pour son suivi, sa disponibilité, ses conseils précieux et le temps accordé à la relecture de mon mémoire.

Un grand merci à l'ensemble de mes collaborateurs d'Actélior pour leur accueil et bienveillance tout au long de mon alternance et jusqu'à ce jour.

Enfin, je tiens à remercier ma famille et mes proches pour leur soutien moral indéfectible.

Table des matières

Introduction

La population des TNS (Travailleurs Non Salariés) est une population possédant des caractéristiques spécifiques et disparates. Elle couvre un large éventail de catégories professionnelles et de secteurs d'activité. Les TNS regroupent ainsi des profils hétéroclites tels que les métiers de l'artisanat et du commerce, les professions libérales médicales et paramédicales, réglementées et non réglementées. Cette population est en constante augmentation, avec une croissance de 25% depuis le début des années 2000. Cette population spécifique reste jusqu'à aujourd'hui peu couverte en matière de protection sociale et notamment de prévoyance. Elle représente ainsi un marché à fort potentiel de développement pour les organismes d'assurance.

Ce mémoire a pour objectif de concevoir un outil de tarification à destination du marché de la prévoyance pour les TNS. Le portefeuille étudié est une base de données composée de plus de 100 000 années d'observation du risque d'arrêt de travail de deux contrats de prévoyance destinés aux TNS. La tarification sera décomposée en un modèle de fréquence et un modèle de coût moyen.

Les modélisations actuarielles, un temps dominé par des modèles de statistiques économétriques classiques tel que le modèle linéaire généralisé (GLM), tendent à se tourner toujours d'avantage vers des modèles statistiques d'apprentissage également appelés *Machine Learning*. Ces méthodes plus récentes que les GLM, associées à des bases de données d'avantage qualitatives et structurées, permettent d'entrevoir de nouvelles alternatives concernant la modélisation du risque. Il convient cependant de garder à l'esprit qu'il est primordial de connaître au mieux leurs mécanismes afin d'exploiter de manière optimale les prévisions de ces méthodes. La compréhension et la comparaison des modèles statistiques classiques avec les modèles statistiques d'apprentissage sera donc un élément majeur de ce mémoire.

De multiples problématiques apparaissent dans l'élaboration de ce mémoire. Comment adapter le modèle de coût moyen et de fréquence au risque arrêt de travail ? Comment prendre en compte les caractéristiques spécifiques de la population des TNS dans la modélisation ? La méthode GLM est-elle la plus adaptée aux données ? Un algorithme d'apprentissage supervisé peut-il produire un modèle plus performant ?

Une première partie abordera la définition et les caractéristiques des TNS. Un état des lieux de la couverture prévoyance des TNS en France sera dressé. En somme, il s'agira dans cette section de présenter le contexte de la prévoyance des TNS.

Une deuxième partie se portera sur la description du portefeuille et des variables qui le composent. Les statistiques descriptives se feront en deux temps. Dans un premier temps, une description des variables sur l'ensemble du portefeuille sera effectuée. Cette description confèrera des éléments de compréhension pour la mise en place du modèle de fréquence. Dans un second temps, seront réalisées les statistiques descriptives des variables explicatives sur la base des données sinistrées exclusivement. Cette seconde analyse permettra de mieux appréhender le modèle de coût moyen des arrêts de travail.

Après ces éléments, une présentation des différents algorithmes considérés dans ce mémoire sera dressée. Une comparaison des résultats obtenus à la suite de l'application des différents algorithmes sera ensuite effectuée. La définition de métriques d'erreur permettra de comparer la performance des modèles obtenus entre eux. Un dernier point d'attention sera porté sur la nécessité d'un lissage du tarif brut dans l'objectif de considérer les contraintes d'une politique de souscription.

I - Les TNS : une population spécifique et peu couverte en prévoyance

1 - La présentation de la population des TNS et de ses caractéristiques

1.1 - La définition du travailleur non salarié

Les TNS font partie d'une catégorie plus large dite des travailleurs indépendants. Les indépendants sont définis par le code de la sécurité sociale comme des travailleurs n'ayant « pas de lien de subordination juridique permanente à l'égard d'un donneur d'ordre et ne disposent pas de contrat de travail ». Ils ne peuvent donc pas bénéficier de la protection du droit du travail. Une grande majorité des travailleurs indépendants sont des travailleurs non-salariés et une minorité regroupe les assimilés salariés :

- Les travailleurs non salariés (TNS) sont affiliés à un régime de protection sociale des TNS. Ils regroupent les entrepreneurs individuels classiques, les micro-entrepreneurs ainsi que les gérants majoritaires de sociétés à responsabilité limitée (SARL, SELARL, EARL, etc.).
- Les assimilés salariés quant à eux cotisent au régime général. Ils correspondent à des dirigeants salariés, à savoir des directeurs généraux ou présidents de société anonyme, des gérants minoritaires de SARL ou encore des présidents de sociétés par actions simplifiées.

Un TNS ne perçoit donc pas de salaire, mais plutôt un revenu. Cette rémunération correspond aux bénéfices calculés à partir du chiffre d'affaires du TNS. Les travailleurs non salariés sont affiliés soit au régime de protection de la Sécurité sociale des indépendants (avant 2020 RSI ; Régime Social des Indépendants), soit à l'Urssaf, soit à la mutualité sociale agricole (MSA).

1.2 - La distinction non-salarié « classique » et micro-entrepreneur

Les non-salariés peuvent être séparés en 2 catégories, les micro-entrepreneurs et les non-salariés dits classiques :

- Les micro-entrepreneurs, désignés auto-entrepreneurs avant le 19 décembre 2014, correspondent à des entrepreneurs individuels qui font la demande d'affiliation au régime du micro-entrepreneur. Ce régime permet des formalités de création d'entreprise plus souples ainsi qu'un mode de calcul de l'impôt sur le revenu et des cotisations sociales simplifié. Le régime du micro-entrepreneur s'applique à titre principal ou complémentaire aux activités libérales du commerce ou encore de l'artisanat. Il permet notamment de bénéficier du régime micro-social ainsi qu'à une franchise en base de TVA à la condition que le chiffre d'affaires ne dépasse pas un certain seuil.
- Les non-salariés classiques correspondent quant à eux aux non-salariés hors micro-entrepreneurs.

Ce mémoire se concentre uniquement sur la population des non-salariés classiques. Les micro-entrepreneurs ont été sortis du cadre de l'étude.

1.3 - Les trois grandes catégories professionnelles des TNS

Les TNS peuvent être répartis entre trois grandes catégories professionnelles :

- Les entrepreneurs individuels
- Les gérants majoritaires
- Les professionnels libéraux

Concernant les entrepreneurs individuels, il s'agit de la forme d'activité non salariée la plus répandue. L'entreprise est dans ce cas dirigée par une unique personne et ne dispose pas de personnalité morale. De ce fait, sur le plan juridique, l'entreprise et l'entrepreneur constituent une seule entité. Les entrepreneurs individuels peuvent faire le choix d'adopter le régime fiscal de la micro-entreprise (régime micro-fiscal) ou également le statut de micro-entrepreneur (régime micro-social). Le régime micro-fiscal concerne les règles d'imposition. Il a pour objectif d'alléger les déclarations fiscales ainsi que la comptabilité des activités des TNS. En comparaison, le régime micro-social concerne les contributions sociales et les cotisations. L'assiette des cotisations est définie comme le chiffre d'affaires du TNS. Le taux de cotisation évolue en fonction de l'activité exercée.

Concernant les gérants majoritaires, il s'agit de la forme d'activité dans laquelle les entrepreneurs créent une société à responsabilité limitée (SARL). L'avantage de la SARL est qu'elle limite la responsabilité aux apports de l'associé par le biais d'une structure juridique souple. Cette forme de société est adaptée aux « petits » projets car elle ne nécessite pas d'apport important en capital. Il existe également les entreprises agricoles à responsabilité limitée (EARL) qui peuvent être créées par un exploitant agricole. Une société d'exercice libéral à responsabilité limitée (SELARL) peut quant à elle être constituée par des professionnels libéraux. Enfin dans le cas particulier où la société est créée par un seul associé, la SARL est alors dite unipersonnelle ou bien nommée entreprise unipersonnelle à responsabilité limitée (EURL).

L'article 29 de la loi 2012-387 définit les professionnels libéraux comme : « Les professions libérales groupent les personnes exerçant à titre habituel, de manière indépendante et sous leur responsabilité, une activité de nature généralement civile ayant pour objet d'assurer, dans l'intérêt du client ou du public, des prestations principalement intellectuelles, techniques ou de soins mises en œuvre au moyen de qualifications professionnelles appropriées et dans le respect de principes éthiques ou d'une déontologie professionnelle, sans préjudice des dispositions législatives applicables aux autres formes de travail indépendant. ». Les professions libérales peuvent être :

- Réglementées : Il s'agit de professions qui sont soumises à une réglementation spécifique. Elles portent essentiellement sur les conditions d'exercice et d'accès ainsi que sur les obligations déontologiques. Ces réglementations sont régies par des instances professionnelles (chambre ou ordre). Les domaines juridiques et de la santé sont particulièrement concernés. Il est possible de distinguer, les professions organisées en ordres professionnels (architecte, avocat, médecin, …), les officiers publics titulaires d'un office conféré par l'Etat (commissaire-priseur, huissier de justice, notaire, …) et les auxiliaires médicaux dont l'activité est réglementée par le code de la santé (diététicien, orthophoniste, pédicure-podologue, …).
- Non réglementées : Ces professions sont définies par ce qu'elles ne sont pas. Il s'agit de toutes les professions qui ne relèvent pas d'une activité agricole, artisanale, commerciale ou industrielle et qui ne sont pas des professions libérales réglementées. Elles peuvent être soit soumises à déclaration d'activité (moniteur d'auto-école), soit totalement libres (actuaire consultant indépendant).

La répartition des effectifs TNS en France est distribuée de la manière suivante :

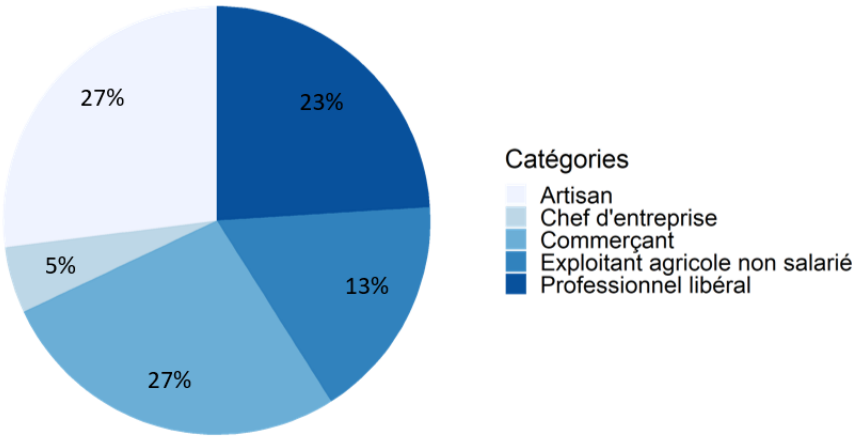


Figure 1 : Diagramme circulaire de la répartition des catégories professionnelles

Source : Urssaf, 2021

1.4 - La multiplicité des secteurs d'activité des travailleurs non salariés

Les TNS couvrent une grande diversité de secteurs d'activité de l'économie. Le graphique ci-dessous représente la répartition des TNS en fonction des secteurs d'activité regroupés de la manière suivante :

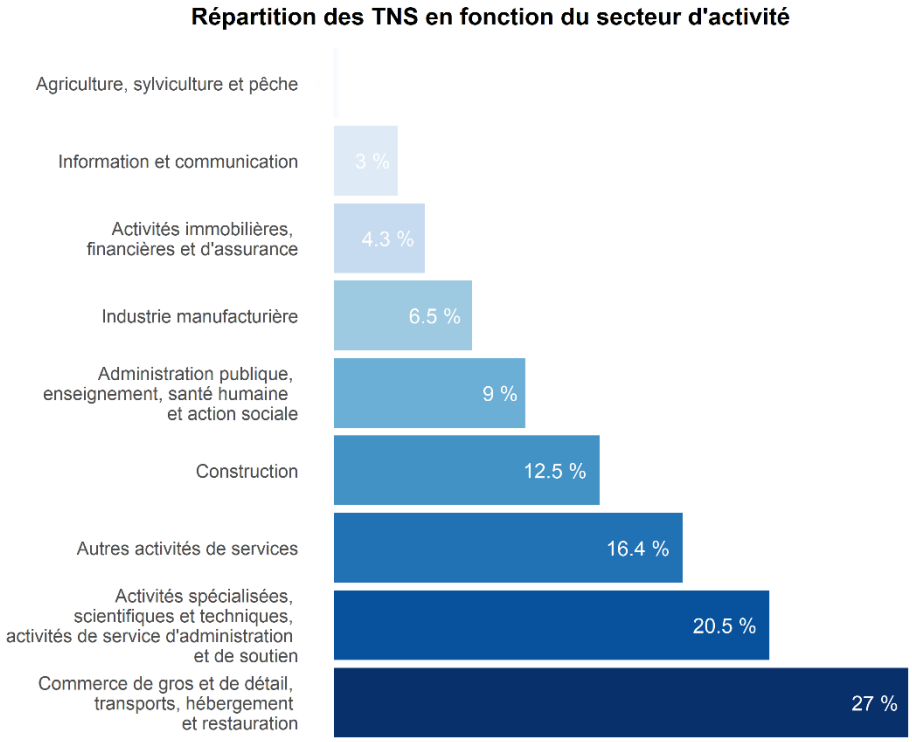


Figure 2 : Répartition des TNS en fonction du secteur d'activité

Source : Urssaf, 2021

Cette répartition est la répartition observée sur l'année 2020 des TNS classiques et micro-entrepreneurs confondus. Les trois secteurs d'activité les plus représentés sont les commerces de gros et de détail,

hébergements, transports et restauration (27% des cotisants), les activités spécialisées, scientifiques et techniques, activités de service d'administration et de soutien (20,5% des cotisants) ainsi que les autres activités de services (16,4% des cotisants). Le secteur d'activité le moins représenté est le secteur de l'agriculture, la sylviculture et la pêche (0,4% des cotisants).

Les secteurs d'activité qui présentent les plus fortes évolutions par rapport à l'année 2019 sont :

- Information et communication, $+ 3{,}8\%$.
- Commerce de gros et de détail, transports, hébergement et restauration, $+ 3{,}6\%$.
- Agriculture, sylviculture et pêche, $- 7{,}9\%$.
- Autres activités de services, $- 3{,}7\%$.

1.5 - Les différentes caisses des travailleurs non salariés

Depuis le 1er janvier 2020, la protection sociale des TNS, auparavant gérée par le Régime social des indépendants (RSI), est intégrée au régime général de la Sécurité sociale via l'organisme appelé Sécurité Sociale des Indépendants (SSI). Ainsi, les activités auparavant gérées par le RSI sont désormais prises en charge par les trois branches du régime général de la Sécurité sociale que sont l'assurance maladie, la CNAV (Caisse Nationale d'Assurance Vieillesse) et le réseau URSSAF.

Tous les TNS sont concernés par la SSI. Concernant les artisans et les commerçants, la SSI est chargée de collecter l'ensemble des cotisations vieillesse, maladie et complémentaire. La spécificité des professions libérales est que seul le corps de métier maladie (santé) dépend de la SSI. En effet, concernant la vieillesse (retraite) et la complémentaire (prévoyance), les professions libérales ne relèvent pas de la SSI mais de la CNAVPL (Caisse Nationale d'Assurance Vieillesse des Professions Libérales). La CNAVPL est constituée des dix sections suivantes :

- Chirurgien-dentiste, CARCDSF
- Médecin, CARMF
- Auxiliaire médical, CARPIMKO
- Pharmacien, CAVP
- Vétérinaire, CARPV
- Agent d'assurance, CAVAMAC
- Expert-comptable, CAVEC
- Officier ministériel, CAVOM
- Architecte, CIPAV
- Notaire, CPRN

Il semble intéressant de noter que les avocats ne dépendent pas de la CNAVPL mais possèdent un dispositif spécifique.

Le poids de chaque section dans la CNAVPL peut être représenté par le diagramme circulaire suivant :

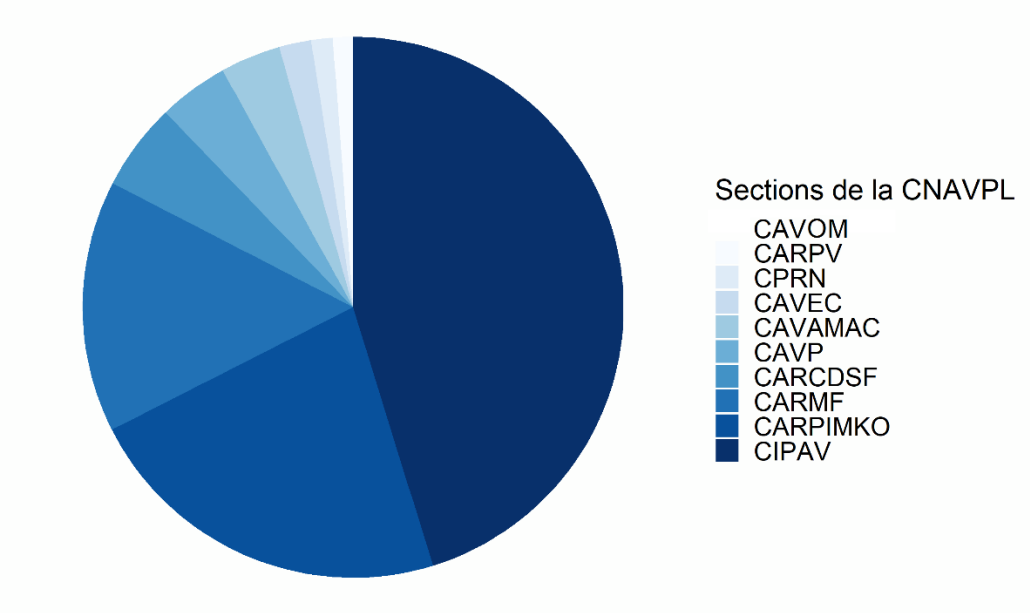


Figure 3 : Diagramme circulaire des effectifs des différentes caisses

Source : CNAVPL, 2022

Comme observé sur ce graphique, la CARPIMKO, la CARMF et la CIPAV représentent à elles seules plus de 85% des effectifs de la CNAVPL.

En prévoyance, les modalités de calcul des différentes cotisations et prestations sont spécifiques à chacune de ces caisses.

1.6 - Quelques chiffres clés sur les TNS

La source des données nationales utilisées dans ce paragraphe est une étude panoramique détaillée du travail indépendant en France effectuée par l'Insee. Cette étude porte sur les départements de la France métropolitaine ainsi que la Guadeloupe, la Martinique, la Réunion et la Guyane. Ces données datent de décembre 2017, la situation a donc pu évoluer depuis, cependant ces données permettent une première approche afin de dépeindre les caractéristiques des TNS.

Le nombre de travailleurs indépendants était de 3,5 millions répartis de cette manière :

- TNS : 3,2 millions
- Assimilés-salariés : 0,3 million

D'après l'INSEE, la proportion de TNS tend à croître dans le secteur tertiaire et à diminuer dans le secteur industriel et agricole. Les TNS sont plus âgés et présentent une proportion d'hommes plus importante que la population des salariés. En effet, l'âge moyen des TNS est de 46 ans contre 40 ans pour les salariés. Cette moyenne d'âge plus élevée chez les TNS s'explique en partie par le fait qu'une portion des TNS a d'abord été salariée avant de devenir non-salarié dans une seconde vie professionnelle. Le pourcentage de femmes est quant à lui de 35% chez les TNS contre 47% pour les salariés.

Les disparités de revenus au sein des TNS sont très marquées. En effet, le revenu mensuel moyen des non-salariés classiques est de 3580 euros avec une rémunération moyenne des médecins qui s'élève à

8 800 euros tandis que celle des micro-entrepreneurs ne s'élève qu'à hauteur de 470 euros. Il convient également de préciser que 16% des TNS sont pluriactifs dont 75% pratiquent une activité salariée en parallèle. Ainsi, les TNS peuvent exercer une activité à titre principal mais aussi en complément d'une activité salariée.

La population des TNS en France représente 10% de la population en emploi. Il est entendu par population en emploi, la population active n'étant pas au chômage à savoir la population active occupée. La répartition des TNS par département français est représentée ci-dessous :

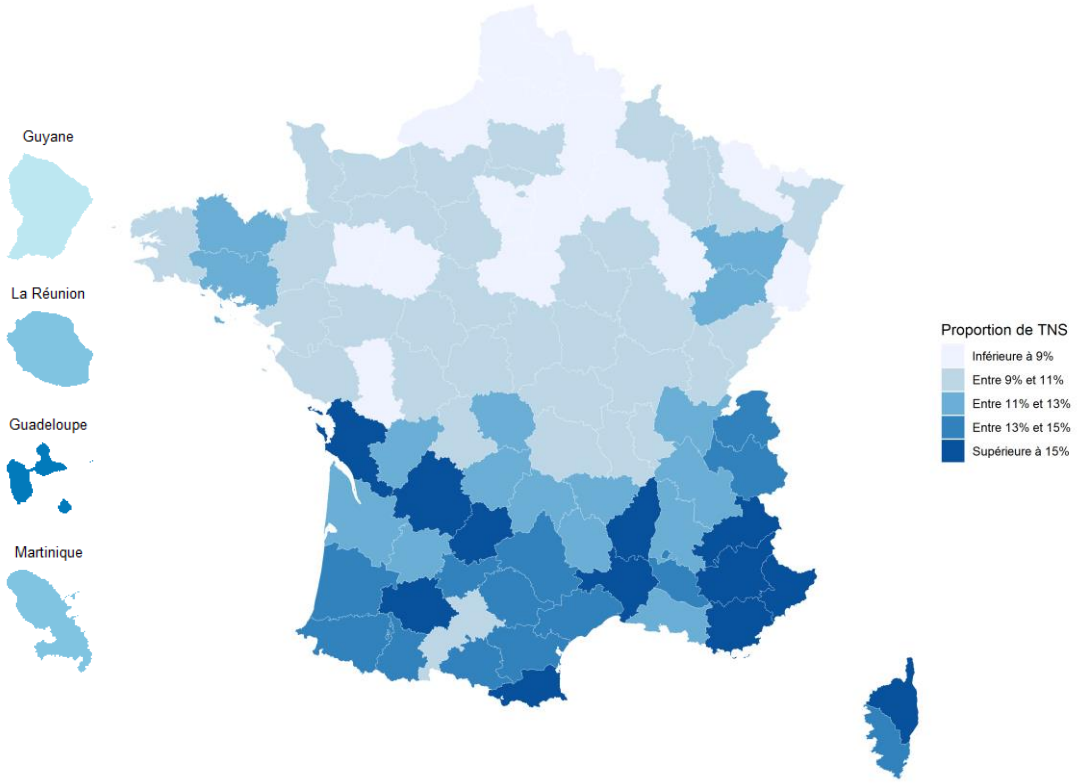


Figure 4 : Proportion des TNS parmi les personnes en emploi par département en décembre 2017

Source : Insee, Estimations d'emploi localisées 2017

La part des TNS dans la population active occupée varie de manière considérable d'un département à l'autre avec des extrêmes en Hautes-Alpes avec 19% et dans les Hauts-de-Seine avec 6%. Il se dégage de cette représentation une différence entre les départements de la moitié nord de la France et ceux de la moitié sud. En effet, dans la moitié nord uniquement les départements des Côtes-d'Armor, du Morbihan, des Vosges et de la Haute-Saône dépassent le seuil des 11% de TNS dans la population en emploi. Au contraire, dans la moitié sud de la France métropolitaine, le seuil des 13% est très souvent dépassé notamment dans les régions Occitanie, Provence-Alpes-Côte-d'Azur et Corse. Deux justifications interdépendantes peuvent être apportées pour expliquer ces disparités :

- Le type des activités exercées : Les départements où le poids des services aux entreprises et de l'industrie est fort et où, à l'inverse, l'économie est peu portée par les services aux particuliers regroupent comparativement moins de TNS que dans les départements où ces dynamiques sont inversées.
- Les caractéristiques de la population : Les départements possédant une population plus âgée nécessitent une plus forte présence des services aux particuliers ce qui est le cas des départements de la moitié sud de la France. Cela peut ainsi se traduire par une forte représentation des TNS dans ces régions.

Concernant les départements et régions d'outre-mer, seule Mayotte n'est pas intégrée dans le périmètre de l'étude. Tout comme la moitié nord de la France métropolitaine trois des quatre départements d'outre-mer sont en-dessous du seuil des 11%. Seule la Guadeloupe présente une proportion plus élevée que 11% mais reste tout de même en dessous des 13%.

Ces différentes informations permettent de saisir la multitude de profils ainsi que la disparité des revenus représentant une grande hétérogénéité au sein de la catégorie des TNS. Il s'agira dans notre étude de bien capter les spécificités et la diversité des profils composant la catégorie des TNS dans la modélisation de la sinistralité.

2 - La prévoyance des TNS

D'après la loi Evin, loi n°89-1009 du 31 décembre 1989, la prévoyance désigne « les opérations ayant pour objet la prévention et la couverture du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail ou d'invalidité ou du risque chômage ».

Dans le référentiel assurantiel, la prévoyance qualifie de manière générique l'ensemble des garanties et contrats couvrant les risques sociaux liés à la personne en cas de cessation d'activité professionnelle. Cette dernière peut être de différente nature, temporaire ou définitive. Elle peut être causée par un accident ou une maladie de la vie quotidienne mais aussi professionnelle engendrant une incapacité, une invalidité ou un décès.

L'intérêt de la souscription de contrats de prévoyance est d'assurer une compensation d'une potentielle perte de revenu liée à ces risques sous forme de capital, rente ou indemnités journalières.

La prévoyance peut couvrir les risques suivants :

- L'incapacité
- L'invalidité
- Le décès
- La maternité
- La perte d'emploi

Concernant les risques incapacité et invalidité, après un délai de carence variable en fonction de la nature de l'arrêt, une prestation est versée pour compenser la baisse de revenu provoquée par l'insolubilité de travailler. Ce mémoire a pour objectif de modéliser la durée en arrêt de travail. Ce sont donc uniquement l'incapacité et l'invalidité qui seront développées dans la section suivante.

2.1 - Focus sur l'arrêt de travail

L'arrêt de travail désigne l'incapacité physique ou psychique d'accomplir les tâches liées à son travail. Plusieurs types d'incapacité existent. Il existe l'incapacité temporaire de travail qui peut être partielle ou totale. Il existe également l'incapacité permanente de travail qui peut être partielle ou totale.

L'arrêt de travail peut avoir deux causes différentes :

- Maladie ordinaire et accident de la vie privée
- Accident du travail et maladie professionnelle

L'arrêt de travail pour maladie ordinaire ou pour accident de la vie privée est plus communément appelé arrêt maladie. En comparaison, par accident du travail, il est entendu tous les accidents pouvant survenir durant l'activité professionnelle. Il semble alors intéressant de notifier qu'un accident survenu sur le trajet entre le domicile et le lieu de travail du travailleur est considéré comme un accident de travail. La maladie d'origine professionnelle est reconnue si elle est répertoriée dans la liste des tableaux des maladies professionnelles sur le site de l'INRS (Institut National de Recherche et de Sécurité). Cependant, une maladie peut être reconnue comme maladie professionnelle même si elle ne figure pas dans la liste citée précédemment. Dans ce cas-ci, elle est reconnue maladie professionnelle par un Comité Régional de Reconnaissance des Maladies Professionnelles.

L'invalidité fait le plus généralement suite à l'incapacité de travail. Le travailleur passe en invalidité lorsque l'état de ce dernier en arrêt de travail demeure permanent. L'incapacité ne pouvant pas durer plus de 3 ans, le passage en invalidité intervient au maximum à la suite de ces 3 années d'incapacité.

Il existe trois catégories d'invalidité :

- Catégorie 1 : Elle désigne la possibilité d'exercer une activité rémunérée.
- Catégorie 2 : Elle désigne l'impossibilité pour le salarié d'exercer une quelconque profession.
- Catégorie 3 : Tout comme la catégorie 2, elle désigne l'impossibilité pour le salarié d'exercer une quelconque profession. A cela s'ajoute un besoin de l'assistance d'une tierce personne pour effectuer les actes de la vie quotidienne.

Les différents types de l'arrêt de travail ont ainsi été définis. Ce mémoire se concentre sur la tarification de ces risques dans le cadre du travail non salarié. Il semble alors intéressant d'étudier le niveau de couverture proposé aux salariés en comparaison de celui proposé au TNS. Ce sera donc l'objet de la prochaine section.

2.2 - Un écart important entre la couverture prévoyance des salariés et celle des TNS

Dans le cadre du salariat, le système de couverture prévoyance français repose sur plusieurs niveaux :

- Le régime de base de la Sécurité sociale : Une prestation indemnitaire est versée, après un certain délai de carence, afin de palier la diminution de salaire due à l'arrêt de travail. Cette prestation compensatoire s'élève à 50 % du salaire journalier de base et est plafonnée à 1,8 SMIC.
- Le régime complémentaire obligatoire : Ce régime vient en complément de celui de la Sécurité sociale. Les entreprises sont tenues de respecter un niveau minimum de couverture légal. La loi de mensualisation du 19 janvier 1978 oblige les entreprises à assurer, un niveau minimum de salaire en cas d'arrêt de travail aux salariés justifiant au moins un an d'ancienneté. La couverture s'élève à 90 % du salaire brut de référence sur les 30 premiers jours et 66,66 % du salaire brut de référence sur les 30 jours suivants. Ces périodes sont augmentées de 10 jours supplémentaires par tranche de cinq ans d'ancienneté.
- La prévoyance collective : Il s'agit d'un troisième niveau de protection qui peut venir compléter les prestations versées par les deux premiers régimes. Elle peut être facultative ou obligatoire.

Il peut être représenté graphiquement de la manière suivante :

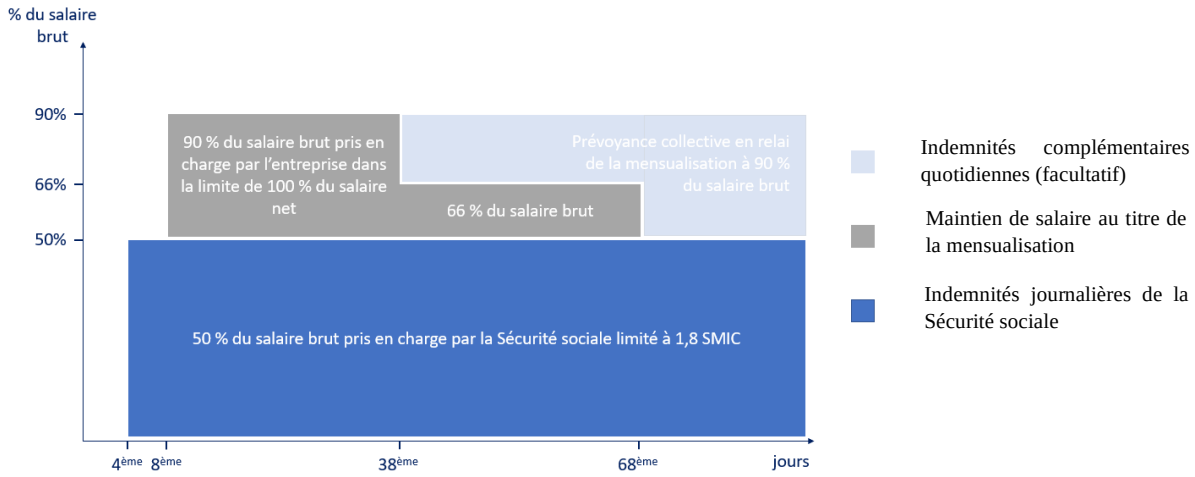


Figure 5 : Exemple de couverture prévoyance d'un salarié

Ce système à multiples niveaux permet une relative bonne couverture en termes de prévoyance des travailleurs salariés. Les TNS, quant à eux, ne disposent pas de ces différents niveaux de couverture. Ils ne possèdent qu'une couverture du régime obligatoire qui est cependant moins couvrante que celle de la Sécurité sociale des salariés. Cela s'accompagne d'un montant de cotisations au régime obligatoire inférieur pour les travailleurs non salariés. Il est donc nécessaire pour les TNS de souscrire une assurance individuelle afin de disposer d'une protection plus couvrante.

A titre d'exemple, un événement traduisant ce manque de couverture par le régime de base est celui de l'introduction d'un dispositif d'indemnités journalières pour certains TNS dans la loi de financement de la Sécurité sociale de 2021. En effet, avant cette date, seule la moitié des caisses des libéraux accordaient des indemnités journalières à leurs affiliés, et cela au terme d'un délai de 90 jours. La crise sanitaire de la covid-19 a mis en lumière cette carence de couverture. L'État a ainsi accoré des indemnités journalières à titre dérogatoire et exceptionnel. Celles-ci ont été financées par l'Assurance Maladie (269 millions d'euros d'indemnités journalières dérogatoires ont été versées aux professionnels libéraux par la CNAM entre le 12 mars et le 24 juillet 2020). La loi de Financement de la Sécurité sociale de 2021 définit un dispositif d'indemnités journalières en cas d'arrêt maladie, commun et obligatoire pour l'ensemble des professionnels libéraux affiliés à la Caisse Nationale d'Assurance Vieillesse des Professions Libérales (CNAVPL). Depuis le 1er juillet 2021, les travailleurs non salariés relevant de cette caisse bénéficient d'indemnités journalières jusqu'au 91ème jour. Ils sont en contrepartie redevables d'une cotisation supplémentaire.

Malgré ces nouvelles réglementations, les régimes de base n'assurent toujours qu'une prise en charge partielle en cas d'incapacité, de décès ou même de cessation d'activité. D'après une étude réalisée par Metlife, le taux d'équipement était seulement de 46% sur l'arrêt de travail en 2021. Malgré la faible couverture proposée par le régime obligatoire, les TNS semblent donc ne pas souscrire systématiquement une prévoyance individuelle.

Il existe essentiellement deux éléments qui peuvent expliquer ce sous-équipement en termes de prévoyance :

- Une première est que les TNS considèrent, de manière injustifiée, posséder une protection élevée et stable. Cette observation a été étudiée dans le cadre du baromètre annuel sur les indépendants effectué par MetLife en collaboration avec CSA. D'après l'étude 2021, 46% et 47% des TNS se sentent bien protégés par leur régime obligatoire respectivement en incapacité et en invalidité. Cette surestimation de leur couverture prévoyance est influencée par une

relative méconnaissance des couvertures proposées par le régime obligatoire. En effet, il a été estimé que seuls 28% des TNS interrogés connaissaient le montant des prestations garanties dans le cas de l'incapacité. En invalidité, ce nombre atteint 21%. Ces chiffres sont équivalents à ceux publiés en 2020.

- Une seconde explication au faible taux de souscription de prévoyance individuelle est celle du coût que cela occasionne pour le TNS. Les garanties prévoyance ont un coût plus élevé pour le TNS que pour un salarié.

Tous ces éléments font de la couverture en protection sociale des TNS un marché à fort potentiel de développement, une couverture complémentaire restant nécessaire pour une couverture optimale. Afin d'encourager à souscrire une couverture prévoyance en réduisant le coût de l'assurance, un dispositif d'incitation fiscal appelé Loi Madelin a donc été mis en place. Il permet de déduire le montant des cotisations des revenus imposables.

2.4 - La loi Madelin

La loi Madelin qui a été promulguée le 11 février 1994 confère la possibilité aux TNS de déduire du revenu imposable les cotisations versées dans le cadre d'un contrat d'assurance dit « Madelin ». L'objectif est pour le TNS de s'assurer en prévoyance que ce soit en incapacité, invalidité ou décès, de financer une retraite complémentaire ou bien de se couvrir contre le risque chômage.

En ce sens, la loi Madelin a pour objectif d'inciter les TNS à se constituer par eux-mêmes une protection sociale, dans l'objectif de combler la relative faiblesse de couverture de leurs régimes obligatoires. Il convient ici de rappeler qu'un TNS est moins bien couvert qu'un salarié concernant les risques liés à la retraite et à la prévoyance. Les TNS, contrairement aux salariés, sont également dépourvus de protection sur le risque chômage. La dimension déductible d'un contrat Madelin permet de contrebalancer l'effort de cotisation à la charge du TNS en finançant une partie par la baisse d'impôt engendrée.

Les cotisations d'un contrat Madelin ne sont cependant pas déductibles sur le plan social : il n'est pas possible de les déduire du calcul des charges sociales dans le cas des régimes industriels et commerciaux (BIC) et du bénéfice non commercial (BNC). Concernant les gérants TNS, les cotisations d'un contrat Madelin se réintègrent au revenu de gérance.

Concernant les contrats Madelin couvrant la prévoyance spécifiquement, ils permettent d'assurer, en cas d'incapacité, d'invalidité, un maintien de revenu au TNS et un capital et/ou les rentes à destination des ayants droit.

Tout comme les contrats de prévoyance « classiques » des salariés, les contrats Madelin ne peuvent engendrer un enrichissement de l'assuré. Cependant, la multitude de régimes obligatoires rend complexe la détermination du niveau de garantie proposée. Dans le but de limiter le risque d'enrichissement dû à la non-considération de la couverture du régime de base, certains contrats sont proposés sous forme indemnitaires. Ces indemnités sont versées sous déduction de celles attribuées par le régime de base. Ce type de contrat s'identifie en général comme étant des contrats haut de gamme puisqu'ils requièrent plus d'informations pour leur mise en œuvre et leur gestion est plus complexe.

Il existe deux types de tarification pour les contrats Madelin concernant la prévoyance :

- Les contrats tarifés à l'âge atteint. Le coût de ces contrats évolue au cours de la vie du contrat en fonction de l'âge atteint par l'assuré. Ce type de tarification est généralement associé à des contrats bas de gamme, mais permet d'être compétitif pour les jeunes assurés sur le court terme.

- Les contrats tarifés à l'âge à l'adhésion. Le coût d'un tel contrat est fixe dans le temps et est déterminé au moment de sa mise en œuvre en fonction de l'âge de l'adhérent. Ce type de contrat est avantageux sur une vision long terme.

Ainsi les différents dispositifs qu'offrent la loi Madelin ont pour objectif de faciliter l'accès des TNS à la couverture prévoyance en réduisant le coût de l'assurance relativement élevé pour un TNS.

Les différentes caractéristiques du TNS et leurs spécificités viennent d'être présentées. Les enjeux de la couverture prévoyance des TNS ont également été abordées. Il convient désormais de présenter le portefeuille d'étude ainsi que ses différentes statistiques descriptives.

II - Statistiques descriptives

L'objectif est de modéliser la fréquence des sinistres ainsi que le coût moyen d'un sinistre à l'aide de méthodes de régression et d'apprentissage de la base de TNS.

1 - Présentation du portefeuille

Le portefeuille dont dispose Actélior est une base comptant plus de 100 000 lignes de données collectées sur une période de neuf années, chaque ligne correspondant à un adhérent pour une année de survenance. Elle recense en parallèle la fréquence des arrêts de travail par adhérent et par année ainsi que la durée de chaque arrêt de travail. Ainsi, à titre d'exemple, si un adhérent a souscrit un contrat pendant quatre ans, il est retranscrit dans la base de données pour ce même adhérent quatre lignes correspondant à ces quatre années de couverture. Pour chaque ligne et donc chaque année de survenance, la fréquence des arrêts de travail, la durée d'exposition au risque et la durée d'indemnisation totale sur l'année concernée sont également renseignées (vide si pas de sinistre).

La base de données est composée de données de sinistralité de deux types de contrats TNS distincts. Ces deux contrats sont des contrats loi Madelin qui ont été présentés précédemment :

- Le contrat A est un contrat à l'âge à l'adhésion.
- Le contrat B est un contrat à l'âge atteint.

Ces deux contrats présentent également une différence sur la nature de la prestation journalière servie en cas d'arrêt de travail :

- Le contrat A sert des prestations indemnitaires.
- Le contrat B sert des prestations forfaitaires.

La prestation indemnitaire est une prestation qui, lorsqu'un arrêt de travail survient, indemnise en fonction du préjudice subi dans les limites des garanties souscrites sans contribuer à l'enrichissement de l'assuré. Dans le cadre de l'arrêt de travail, le principe indemnitaire implique la prise en compte du remboursement du régime obligatoire afin d'en déduire les indemnités journalières à verser et ainsi respecter le principe de non-enrichissement de l'assuré. En comparaison, la prestation forfaitaire est une prestation déterministe qui est indépendante de la nature du préjudice subi. Dans le cadre de la prévoyance, cela correspond à garantir un certain montant fixe de prestations journalières qui sera servi en cas de sinistre. Les contrats versant des prestations indemnitaires sont plus complexes du fait de la nécessité d'une prise en compte du remboursement du régime obligatoire qui varie avec la caisse à laquelle le TNS est rattaché. La tarification doit également tenir compte de ce paramètre.

Ces deux éléments réunis font du contrat A, un contrat plus complexe et donc plus haut de gamme que le contrat B. Le contrat A s'adresse aux gérants majoritaires et aux professions libérales (à l'exclusion des entrepreneurs individuels).

Il a été fait le choix de porter cette étude sur une population TNS adhérente à ces deux offres de prévoyance. Cela présente l'intérêt d'obtenir une taille plus conséquente de la base de données. Il en résulte un modèle plus stable présentant une plus faible variance des estimateurs. Cette combinaison des deux portefeuilles peut entraîner un biais dans la modélisation. Cependant, le gain en stabilité des estimateurs obtenus nous a orienté vers ce choix. Par la suite, la distinction entre les entrepreneurs individuels, adhérents du produit à l'âge atteint, et les gérants majoritaires, adhérents du produit à l'âge à l'adhésion, a été conservée afin de prendre en compte en partie cette caractéristique dans la modélisation, l'objectif étant d'atténuer le biais engendré par ce choix.

2 - Introduction de la variable à modéliser

Dans le cadre d'une tarification en prévoyance, pour calculer la prime pure, une méthode consiste à estimer deux variables. Il s'agit de la variable relative à la fréquence d'un sinistre ainsi que la variable relative au coût moyen d'un sinistre. Le modèle de fréquence est construit sur l'ensemble du portefeuille tandis que le coût moyen est uniquement construit sur la base des lignes sinistrées. Les statistiques descriptives relatives à la base des contrats sinistrés et donc à la variable représentant le coût moyen seront présentées dans des parties ultérieures.

La variable à expliquer sur l'ensemble du portefeuille est la fréquence des sinistres. Cette variable correspond au nombre d'arrêts de travail pondéré par l'exposition. La fréquence s'écrit comme suit :

$$Fréquence = \frac{Nombre\ de\ sinistres}{Exposition}$$

L'exposition sur l'année N est calculée de la manière suivante :

$$Exposition_N = \frac{date_{radiation_N} - date_{adhésion_N} + 1}{365{,}25}$$

Avec :

- $date_{radiation_N} = \min\left(date_{radiation}, date_{finPériode_N}\right)$
- $date_{adhésion_N} = \max\left(date_{adhésion}, date_{débutPériode_N}\right)$

La fréquence moyenne des sinistres du portefeuille peut se calculer selon deux méthodes :

- Méthode 1 : La fréquence est considérée comme le nombre total de sinistres sur l'exposition totale du portefeuille :

$$Fréquence_{Portefeuille} = \frac{\sum_{i=1}^{n} Nombre\ de\ sinistres_i}{\sum_{i=1}^{n} Exposition_i}$$

- Méthode 2 : La fréquence du portefeuille est considérée comme la moyenne des fréquences individuelles obtenues pour chaque ligne du portefeuille.

$$Fréquence_{Portefeuille} = \frac{\sum_{i=1}^{n} \frac{Nombre\ de\ sinistres_i}{Exposition_i}}{n}$$

Ces deux méthodes conduisent à des fréquences moyennes différentes. La fréquence obtenue avec la deuxième méthode est 7% plus faible que celle obtenue avec la première méthode. Cela est principalement dû aux valeurs extrêmes. En effet, il sera constaté lors de la construction du modèle qu'après suppression des valeurs extrêmes les deux méthodes de calculs donnent des fréquences similaires. La méthode de calcul retenue dans la suite du mémoire est la méthode 2.

Une analyse de l'évolution des effectifs du portefeuille en fonction des années a été réalisée. La taille du portefeuille est en constante augmentation entre la première année et la neuvième année de commercialisation. Cette évolution concorde avec l'expansion du marché des TNS observée au niveau national et traduit d'une certaine manière le potentiel de croissance de ce marché sur les équipements de prévoyance.

Une étude sur l'évolution de la sinistralité en fonction des années a été menée. Il n'est pas ressorti de tendance particulière, la sinistralité s'est révélée être globalement stable sur la période. La variable *Année* n'a donc pas été retenue comme explicative dans la section suivante.

Après avoir présenté le portefeuille ainsi que la variable d'intérêt, il convient désormais d'effectuer l'étude des statistiques descriptives afin d'obtenir une première vision du portefeuille et de ses différentes caractéristiques.

3 - Présentation des variables discriminantes

L'étude des statistiques descriptives est une étape essentielle dans la construction d'un modèle. Elle permet de visualiser les estimateurs empiriques de chaque variable explicative. Il convient dans un premier temps de représenter les effectifs composant les différentes variables explicatives du modèle ainsi que leurs fréquences associées.

Les variables disponibles dans la base de données sont nombreuses. Elles ne peuvent donc pas toutes être détaillées. Cependant, il va être explicité celles qui, au regard des tarifications effectuées dans les parties suivantes, sont considérées comme discriminantes.

Deux types de variables sont présentées, à savoir les variables liées aux caractéristiques :

- De l'assuré :
  - Région
  - Catégorie professionnelle
  - Age
  - Genre
- Du contrat :
  - Durée d'exposition au risque
  - Franchise
  - Equipement en santé

La variable correspondant à la durée d'exposition ne possède pas le même statut que les autres variables explicatives considérées ci-dessus. Il conviendra de préciser son utilité dans cette partie.

Les variables peuvent être de nature :

- Quantitative : Il s'agit des variables numériques. Elles peuvent être discrètes (âge) ou continues (montant des indemnités journalières).
- Qualitative : Il s'agit des variables dont les modalités sont des qualités. Ces variables peuvent être ordinales, c'est-à-dire qu'il existe une relation d'ordre entre les différentes modalités (les régions d'un zonier). A l'inverse, ces variables sont nominales s'il n'existe pas de relation d'ordre entre les modalités de la variable (les catégories professionnelles).

Les parties suivantes présenteront alors les variables explicatives retenues dans la construction du modèle de fréquence.

3.1 - Les régions

La variable région est une variable qualitative nominale composée des 22 anciennes régions de la France métropolitaine ainsi que d'une région nommée DROM rassemblant les Départements et Régions d'Outre-Mer.

La répartition des adhérents varie fortement parmi ces catégories. Les régions prépondérantes sont la région Rhône Alpes, la région parisienne ainsi que la région PACA. Elles représentent à elles-seules la moitié des effectifs. Les régions Corse et Limousin s'avèrent être les régions les moins exposées. Plusieurs régions présentent des effectifs trop faibles pour permettre de conserver cette segmentation dans la modélisation. En effet dans ce cas, le risque de faible exposition serait trop important. Pour chaque catégorie présentant trop peu d'effectifs, une seule valeur extrême pourrait fausser la fréquence moyenne. Ainsi, lors du regroupement, il est nécessaire de porter une attention particulière à ce biais.

Le regroupement des classes de la variable *Région* s'assimile à la construction d'un zonier.

Deux critères nous ont permis de classer les différentes régions :

- Un premier critère est celui des coefficients obtenus pour chaque région avec le GLM (*Generalized Linear Model*). Il s'agit ici de regrouper les régions ayant obtenu des coefficients GLM proches.
- Un second critère est celui de la distance entre chaque région. Il est attendu qu'entre deux régions voisines la fréquence ne varie pas fortement afin de conserver une certaine cohérence.

L'arbitrage entre ces deux critères a permis de créer une nouvelle variable *zoneGéographique* ne comprenant plus que cinq niveaux :

- *Région_1*
- *Région_2*
- *Région_3*
- *Région_4*
- *Région_5*

En considérant ces cinq régions, la modalité la moins représentée est *Région_4*. L'effectif de cette modalité reste cependant suffisamment important et neutralise ainsi le risque de faible exposition.

Les fréquences empiriques de la *Région_5* à la *Région_1* peuvent être classées de manière décroissante. Elles indiquent ainsi une certaine cohérence dans le choix du regroupement des régions. Il est cependant à noter un faible écart entre les fréquences empiriques des régions 3 et 4 et celles des régions 1 et 2. La fréquence empirique de la région 5 indique une importante sinistralité.

3.2 - Les catégories professionnelles

Les TNS couvrent une grande diversité de catégories professionnelles. La base de données comprend cette multitude de catégories. Par souci de présentation et de compréhension, il a été fait le choix de regrouper ces différentes catégories en 4 grandes classes :

- Artisans, commerçants et chefs d'entreprise
- Professions libérales (PL) médicales et paramédicales
- Autres professions libérales
- BTP et Agricole

Les statistiques nationales présentent la répartition suivante :

- Artisans, commerçants et chefs d'entreprise : 59%
- Professions libérales : 24 %
- BTP et Agricole : 17%

Le portefeuille tend à avoir une surreprésentation des professions libérales et une sous-représentation des artisans et commerçants par rapport aux statistiques nationales.

Lors de la modélisation, un maillage plus fin sera pris en compte afin de mieux appréhender la spécificité de chaque profession tout en conservant une exposition satisfaisante pour chaque classe. Il sera également fait la distinction entre les gérants majoritaires et les entrepreneurs individuels.

3.3 - Les franchises

La variable *Franchise* est une variable qualitative composée initialement de 9 classes. La nomenclature pour chaque classe est la suivante :

$$X/Y/Z$$

Avec :

- X, la franchise appliquée dans le cas d'un arrêt de travail de type vie privée.
- Y, la franchise appliquée dans le cas d'un arrêt de travail faisant suite à un accident du travail ou à une maladie professionnelle.
- Z, la franchise appliquée dans le cas d'une hospitalisation.

Une forte disparité des effectifs existe dans chaque classe de la variable *Franchise*. En effet, la franchise *30/3/3* représente plus de la moitié des effectifs tandis que les 6 classes les moins représentées ne comprennent à elles-seules moins de 10% des effectifs. Il convient ici de regrouper les différentes franchises afin de réduire l'impact des classes ayant de faibles expositions.

Nous avons considéré comme hypothèse pour le regroupement des franchises que le passage de 0 à 3 jours de la durée de carence en hospitalisation ne présente pas un impact important sur la survenance de l'arrêt de travail pour deux raisons. Tout d'abord, l'écart de 3 jours entre ces deux franchises reste faible. De plus, la partie Hospitalisation de la franchise s'avère moins sujette à de l'antisélection ou de l'aléa moral que la partie vie privée par exemple.

Finalement, les regroupements de franchise effectués sont les suivants :

- Franchise 7/3 : Comprend les franchises 7/3/0 et 7/3/3.
- Franchise 15/3 : Comprend les franchises 15/3/0 et 15/3/3.
- Franchise 30/3 : Comprend les franchises 30/3/0 et 30/3/3.
- Franchise 30/30 : Correspond à la franchise 30/30/30.
- Franchise 60/60 : Correspond à la franchise 60/60/60.
- Franchise 90/90 : Correspond à la franchise 90/90/90.

Avec cette répartition, la plus faible exposition est la franchise *60/60*. Les effectifs de cette modalité restent relativement faibles. Cette particularité sera prise en compte dans l'étape de modélisation. La modalité *90/90* présente également une faible exposition. Cependant, cette modalité présente des effectifs suffisants permettant ainsi de réduire le biais de faible exposition. Grâce à ces regroupements, le problème des faibles expositions a été contourné concernant la variable *Franchise*.

De manière relativement logique, la fréquence empirique diminue avec la durée de franchise du contrat. En effet, la variable *Franchise* implique une troncature à gauche dans l'observation des sinistres. De cette manière, les sinistres de durée inférieure à la franchise ne sont pas déclarés auprès de l'organisme assureur puisque dans ce cas l'indemnisation reste nulle. Une partie de la baisse observée de la fréquence des sinistres avec l'augmentation de la franchise est ainsi expliquée par l'existence de cette troncature.

Cette tendance peut également être expliquée par d'autres facteurs comme l'antisélection qui peut exister au moment de la souscription du contrat ou encore l'aléa moral au moment de la survenance de l'arrêt de travail. L'antisélection résulte d'une asymétrie d'information qui rend possible un comportement opportuniste précontractuel de l'assuré. A titre d'exemple, dans le portefeuille étudié, il est possible qu'un individu détienne l'information qu'il n'est pas en bonne santé et qu'il risque donc de tomber en arrêt de travail. Or, en l'absence d'un questionnaire médical, l'assureur ne détient pas cette information. En comparaison, l'aléa moral résulte d'une asymétrie d'information qui rend possible un comportement opportuniste post-contractuel de l'assuré. Il s'agit d'un biais où l'individu se comporte différemment que s'il était totalement lui-même exposé au risque. Un exemple en arrêt de travail est l'absence de franchise. Dans le cas où l'assuré possède une couverture optimale, il pourrait décider de se déclarer plus facilement en arrêt de travail et d'y rester plus longtemps que s'il n'était que partiellement couvert. Une dérive de la sinistralité est alors susceptible d'apparaître dans chacune de ces deux situations. L'antisélection et l'aléa moral représentent alors des biais comportementaux difficilement quantifiables mais qu'il est nécessaire de prendre en compte dans la tarification.

La troncature introduite par la variable *Franchise* réduit la fréquence observée des sinistres dans le portefeuille pour les grandes franchises mais n'introduit cependant pas de biais dans la modélisation. En effet, l'objectif final est la construction d'un modèle de fréquence et d'un modèle de coût moyen. Ainsi, la non-observation de sinistres, dont le coût est nul, n'influe pas sur les résultats.

3.4 - L'âge

L'âge de l'assuré est une variable quantitative comprise entre 18 et 67 ans. Il convient pour la modélisation et notamment pour le GLM de discrétiser la variable *Age*. Une première approche consiste à tracer les fréquences empiriques et de repérer une discrétisation possible. Une seconde approche consiste à réaliser un arbre CART (*Classification And Regression Tree*) et de repérer les différentes classes formées. En arbitrant entre ces deux méthodes, la discrétisation qui est apparue comme la plus adaptée est la suivante :

- *29 ans et moins*
- *30-34 ans*
- *35-39 ans*
- *40-49 ans*
- *50-54 ans*
- *55-59 ans*
- *60 ans et plus*

Ce sont ces différentes classes d'âge qui seront retenues dans la modélisation. Ces modalités ont été intégrées dans une nouvelle variable *classeAge*.

Cette discrétisation conduit à des classes de tailles hétérogènes. La classe possédant la plus faible exposition est celle des *60 ans et plus*. Cependant, les effectifs de cette classe d'âge sont suffisants car ils empêchent l'apparition d'un biais de faible exposition. Ainsi, en considérant cette segmentation spécifique, chacune des modalités de cette variable explicative possède une exposition au risque suffisante.

3.5 - Le genre

Le genre est une variable qualitative composée de deux classes :

- Femme
- Homme

L'exposition de ces deux classes est relativement équilibrée avec une légère surreprésentation d'hommes par rapport aux femmes. A titre de comparaison, les statistiques nationales qui présentent quant à elles un écart plus marqué de 35% de femmes chez les TNS. Concernant la fréquence empirique des sinistres, elle s'avère plus élevée chez les femmes que chez les hommes. Il n'est effectué aucun retraitement sur cette variable.

Dans le cadre d'une tarification de produit individuel, le tarif ne peut dépendre du genre néanmoins il est nécessaire de prendre en compte cette variable dans la modélisation avec par exemple, l'utilisation d'un coefficient de majoration appliqué à tous les adhérents quel que soit leur sexe. Un portefeuille composé principalement d'hommes ne possédera alors pas le même risque qu'un portefeuille composé principalement de femmes. Il convient donc de vérifier que le portefeuille conserve la même proportion de femmes et d'hommes au cours du temps que celle utilisée pour construire le tarificateur. Autrement, un biais dans la sinistralité est susceptible d'apparaître. Cette variable reste importante pour la modélisation car elle a un impact non négligeable sur la fréquence.

3.6 - L'équipement en santé

La variable *équipementSanté* est une variable binaire qui prend la valeur 1 si l'adhérent a également souscrit un produit Santé chez ce même organisme assurance et 0 dans le cas contraire.

Malgré la faible exposition relative de la classe 1, à savoir les adhérents équipés en santé, elle contient tout de même des effectifs suffisants. Ainsi il n'apparaît pas de risque de biais lié à la faible exposition pour la variable *équipementSanté*. Cette variable n'est donc pas retraitée.

J'ai fait le choix d'intégrer cette variable dans la modélisation afin de prendre en compte le fait que si un TNS se couvre également en santé cela peut induire essentiellement deux effets positifs :

- S'équiper en santé et en prévoyance montre la volonté du TNS de se couvrir contre les différents risques, et pas seulement contre le risque AT. De cette manière, la souscription à un équipement santé en parallèle d'un équipement prévoyance pourrait réduire le risque d'antisélection sur la garantie prévoyance.
- Si elle est utilisée, la couverture santé permet de mieux prévenir les problèmes de santé. Cette prévention peut avoir un impact positif sur la santé du TNS et ainsi sur le risque arrêt de travail.

Ces arguments sont à nuancer car le TNS peut avoir souscrit un contrat santé chez un organisme assureur autre que celui du portefeuille étudié. Dans ce cas, et malgré sa couverture en santé, le TNS est classé dans la modalité 0. Ce biais peut atténuer les effets positifs attendus sur la fréquence des sinistres de la modalité 1 par rapport à la modalité 0.

Toutefois, les effets positifs attendus sur la sinistralité semblent concorder avec les fréquences empiriques constatées sur le portefeuille entre ces deux modalités. En effet, ces dernières montrent une plus faible fréquence des sinistres sur les TNS qui se couvrent également en santé chez le même organisme assureur.

3.7 - La durée d'exposition au risque

La variable *duréeExposition* est une variable continue qui représente la durée sur laquelle l'adhérent est couvert sur une année. La durée d'exposition varie alors de 0 à 1 et vaut 1 dans le cas où l'adhérent est couvert sur l'année entière. Dans ce portefeuille, la majorité des contrats court sur une année entière.

Cette variable se distingue des autres variables explicatives du modèle. En effet, elle peut être intégrée dans la modélisation de deux manières :

- La durée d'exposition peut être prise en compte en *offset.*
- Cette variable peut aussi être intégrée à la variable à expliquer.

En GLM, une variable *offset* est traitée comme une variable explicative avec un coefficient fixé à 1,0. Elle est généralement utilisée afin de mettre à l'échelle la modélisation des observations individuelles. Elle peut représenter une taille, une durée d'exposition ou toute autre mesure temporelle. A titre d'exemple, dans le portefeuille d'étude, l'objectif est de prendre en compte le fait que si un adhérent présente un sinistre pour une faible durée d'exposition, il présenterait probablement un plus grand nombre de sinistres dans le cas où sa durée d'exposition était égale à 1.

La durée d'exposition sera alors considérée en *offset* dans l'application des algorithmes GLM et *gradient boosting* tandis qu'elle sera directement intégrée dans la variable à expliquer pour les algorithmes arbre CART et forêt aléatoire. Ces différentes méthodes seront expliquées dans les parties suivantes de ce mémoire.

4 - Etude de la corrélation entre les variables

L'étude de la corrélation des variables entre elles doit être réalisée. Pour ce faire, une matrice de V Cramer a été utilisée. La matrice de V Cramer permet de mesurer le degré de dépendance entre les lignes et les colonnes d'un tableau. Ici, il s'agit de mesurer l'intensité des relations entre les variables explicatives. Le test du khi-deux permet seulement de connaître si les variables explicatives entretiennent une relation avec un certain degré de certitude. Le test de Cramer, quant à lui, prend en compte la taille de l'échantillon ainsi que le nombre de degrés de liberté. Cela permet de quantifier l'intensité des relations.

Le test V Cramer s'écrit :

$$V\ Cramer = \sqrt{\frac{\chi^2}{N * DDL}}$$

Avec :

χ^2 ; le khi2

N ; l'effectif total (la taille de l'échantillon)

DdL ; le degré de liberté = min(nb de ligne;nb de colonne) -1

En appliquant ce test pour chaque couple de variable, la matrice V Cramer obtenue est la suivante :

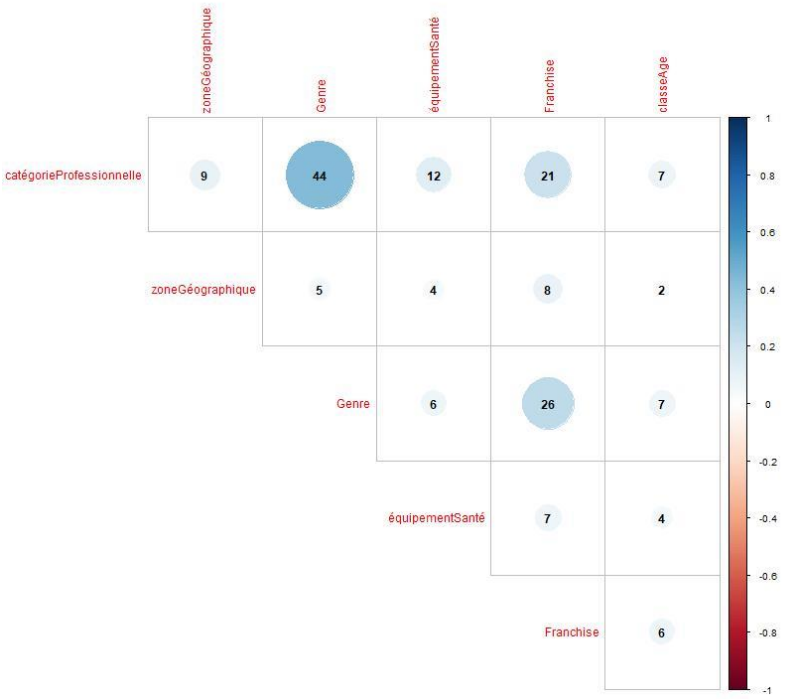


Figure 6 : Matrice V Cramer des variables explicatives

Plus le V Cramer est proche de 0, plus les variables sont décorrélées. A l'inverse plus le V Cramer est proche de 1, plus elles sont liées. D'après la matrice V Cramer représentée ci-dessus, les variables sont globalement peu corrélées entre elles à l'exception des variables *catégoriePro* et *Genre*. Cette corrélation vient du fait que certaines catégories socio-professionnelles sont plus susceptibles d'attirer un genre plutôt que l'autre. Cette corrélation est forte, cependant, il semble nécessaire de conserver ces deux variables dans l'étape de modélisation. Ainsi toutes les variables sont conservées pour la construction du modèle.

La présentation du portefeuille ainsi que les statistiques descriptives permettent d'établir une première approche des données. Naturellement ces estimateurs empiriques ne possèdent qu'une fonction descriptive et ne peuvent être utilisés dans le cadre d'une modélisation. Il s'agit désormais de présenter les différents algorithmes qui vont être utilisés lors de l'étape de modélisation de la fréquence. Dans un premier temps, l'algorithme traditionnellement utilisé en tarification à savoir le GLM va être étudié et présenté. Dans un second temps, des méthodes d'apprentissage supervisé aussi nommées, *machine learning*, seront également étudiées et présentées.

III - Présentation des différents algorithmes

Il existe deux grandes techniques de régression :

- La première méthode dite « classique » regroupe :
  - La régression linéaire
  - Le modèle linéaire généralisé
- La deuxième méthode dite « d'apprentissage supervisé » rassemble :
  - L'arbre CART
  - La forêt aléatoire
  - Le gradient boosting

Les différentes techniques de régression de ces deux méthodes, à commencer par le modèle linéaire généralisé, vont être détaillées dans la section suivante.

1 - Le GLM

Avant d'introduire le GLM, la méthode de régression linéaire sera présentée. Il s'agit de la méthode sur laquelle repose le GLM.

1.1 - La régression linéaire

La régression linéaire est née de l'envie d'exprimer de manière quantitative les liens entre certaines variables. L'objectif est donc d'exprimer une variable dite « expliquée » en fonction de variables dites « explicatives » par le biais d'une combinaison linéaire. Le modèle de régression linéairepour un individu $i \in [\![1; n]\!]$ s'exprime sous la forme suivante :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

Cette expression peut également s'exprimer sous forme matricielle, avec $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ le vecteur à expliquer :

$$\boldsymbol{Y} = \boldsymbol{\beta}'\boldsymbol{X} + \boldsymbol{\varepsilon}$$

Avec :

- $\boldsymbol{X} = \begin{pmatrix} X_{11} \cdots X_{1i} \cdots X_{1n} \\ \vdots \quad \vdots \quad \vdots \\ X_{j1} \cdots X_{ji} \cdots X_{jn} \\ \vdots \quad \vdots \quad \vdots \\ X_{p1} \cdots X_{pi} \cdots X_{pn} \end{pmatrix}$; correspondant à la matrice des variables explicatives.

- $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$; correspondant à la matrice des coefficients des variables explicatives.
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$; correspondant à la composante aléatoire couramment appelée résidu.

Les hypothèses suivantes doivent être vérifiées pour appliquer ce modèle :

- Les variables explicatives X_j sont indépendantes deux à deux.
- Les individus Y_i sont indépendants entre eux.
- $\mathbb{E}(\varepsilon_i) = 0$; les résidus ont une espérance nulle.
- $\mathbb{V}(\varepsilon_i) = \sigma^2$; traduit l'hypothèse d'homoscédasticité.
- $\varepsilon_i \sim \boldsymbol{\mathcal{N}}(0, \sigma^2)$; correspond à l'hypothèse de normalité.

L'hypothèse de normalité des résidus apparaît rapidement contraignante puisqu'elle ne semble pas pertinente pour la majorité des événements aléatoires.

Le modèle linéaire généralisé vient compléter la régression linéaire car il propose une solution à la rigidité de l'hypothèse de normalité en introduisant une fonction lien. Cette hypothèse est remplacée par le fait que la variable expliquée *Y* appartient à une distribution de la famille exponentielle. La méthode des GLM est présentée dans la partie suivante.

1.2 - Introduction au GLM

L'intérêt du GLM réside dans le fait qu'il permet de :

- Modéliser des réponses diverses (réelles, réelles positives, entières…)
- Intégrer tout type d'information exogène susceptible d'influer sur la variable dépendante (réponse $\boldsymbol{Y}$).
- Quantifier l'impact des facteurs de risque X_j (sens et intensité)

Le GLM nécessite d'introduire deux hypothèses fondamentales :

- Les individus Y_i sont indépendants entre eux
- Les variables explicatives X_j sont indépendantes deux à deux

McCullagh et Nelder dans *Generalised linear models* (1989) introduisent les 3 éléments qui composent le modèle linéaire généralisé pour un individu $i \in [\![1; n]\!]$:

- La loi de la réponse aléatoire Y_i
- Le prédicteur $\eta_i = \sum_{j=1}^{p} \beta_j X_{ij}$
- La fonction lien g

1.2.1 - La loi de réponse aléatoire

Comme indiqué précédemment, Y_i appartient par hypothèse à une distribution de la famille exponentielle.

La densité d'une variable aléatoire Y_i appartenant à la famille exponentielle peut s'écrire sous la forme :

$$f(y_i, \theta_i, \phi, \omega_i) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)}\omega_i + c_i(y_i, \phi) \right)$$

Avec :

- $\theta_i \in \mathbb{R}$; correspond au paramètre canonique
- $\phi \in \mathbb{R}$; correspond au paramètre de dispersion
- a ; correspond à une fonction définie sur $\mathbb{R}$, non nulle
- b ; correspond à une fonction définie sur $\mathbb{R}$, dérivable deux fois.
- c_i ; correspond à une fonction définie sur $\mathbb{R}^*$
- ω_i ; correspond à un poids

En considérant maintenant le vecteur à expliquer $Y = (Y_1, \dots, Y_n)$, la densité devient :

$$f_Y(y) = exp\left(\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)}\omega_i + \sum_{i=1}^{n} c_i(y_i, \phi) \right)$$

La famille exponentielle comprend notamment les lois suivantes, à savoir :

- La loi binomiale
- La loi de poisson
- La loi géométrique
- La loi exponentielle
- La loi normale
- La loi gamma

1.2.2 - Le paramètre de dispersion

Le paramètre de dispersion aussi appelé paramètre de nuisance ϕ est supposé connu. Si ce n'est pas le cas, ce paramètre est estimé préalablement et est considéré ensuite comme connu. Il est noté ϕ_0 et dans une grande majorité de cas, la relation suivante est vérifiée : $a(\phi_0) = \phi_0$.

Le tableau ci-dessous représente les paramètres de différentes lois exponentielles :

Distribution		$\boldsymbol{\theta(\mu)}$	$\boldsymbol{b(\theta)}$	$\boldsymbol{a(\phi_0)}$
Normale	$\mathcal{N}(\mu,\sigma^2)$	μ	$\frac{\theta^2}{2}$	σ
Bernoulli	$\mathcal{B}(1,\mu)$	$\log(\frac{\mu}{1-\mu})$	$\log(1+e^{\theta})$	1
Poisson	$\mathcal{P}(\mu)$	$\log(\mu)$	e^{θ}	1
Gamma	$\mathcal{G}(\mu,\nu)$	$-\frac{1}{\mu}$	$-\log(-\theta)$	$\frac{1}{\nu}$
Gauss Inverse	$\mathcal{IG}(\mu,\sigma^2)$	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{\frac{1}{2}}$	σ^2

Tableau 1 : Paramètres de quelques lois appartenant à la famille exponentielle

1.2.3 - La fonction score

Les propriétés de la fonction score, permettent d'établir que :

$$\mu_i = \mathbb{E}(Y_i) = b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}$$

Et

$$\mathbb{V}(Y_i) = a(\phi_0)b''(\theta_i) = \frac{\partial^2 b(\theta_i)}{\partial {\theta_i}^2}$$

Le tableau ci-dessous récapitule l'espérance et la variance pour quelques lois de la famille exponentielle :

Distribution		$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{b'(\theta)}$	$\mathbb{V}(\boldsymbol{Y}) = \boldsymbol{a(\phi_0)b''(\theta)}$
Normale	$\mathcal{N}(\mu,\sigma^2)$	$\mu = \theta$	σ^2
Bernoulli	$\mathcal{B}(1,\mu)$	$\mu = \frac{e^\theta}{1+e^\theta}$	$\mu(1-\mu)$
Poisson	$\mathcal{P}(\mu)$	$\mu = e^\theta$	e^θ
Gamma	$\mathcal{G}(\mu,\nu)$	$\mu = -\frac{1}{\theta}$	$\frac{\mu^2}{\nu}$
Gauss Inverse	$\mathcal{IG}(\mu,\sigma^2)$	$\mu = (-2\theta)^{-\frac{1}{2}}$	$\mu^3\sigma^2$

Tableau 2 : Espérance et variance de quelques lois appartenant à la famille exponentielle

1.2.4 - Le prédicteur

Le prédicteur noté $\eta_i = \sum_{j=1}^{p} \beta_j X_{ij}$ est linéaire et déterministe. Les facteurs de risque explicatifs le constituent.

1.2.5 - La fonction lien

Cette fonction fait le lien entre la composante aléatoire et la composante déterministe. La fonction de lien g doit être monotone, dérivable et inversible.

Elle est alors définie telle que :

$$E(\boldsymbol{Y}|\boldsymbol{x}) = \mu = g^{-1}(\boldsymbol{\beta X})$$

Ou de manière équivalente pour $i \in [\![1; n]\!]$:

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} = \eta_i$$

1.2.6 - Le paramètre canonique

A toute loi de probabilité de la composante aléatoire est associée une fonction spécifique de l'espérance appelée paramètre canonique.

La fonction de lien qui utilise le paramètre canonique dans la famille des modèles linéaires généralisés, est appelée la fonction de lien canonique. La fonction de lien canonique est souvent utilisée pour construire les GLM.

Le tableau ci-dessous récapitule les fonctions de loi canonique de quelques lois appartenant à la famille exponentielle

Loi de probabilité	Fonction de lien canonique
Normale	$\eta = \mu$
Poisson	$\eta = ln(\mu)$
Gamma	$\eta = -\frac{1}{\mu}$
Inverse gaussienne	$\eta = -\frac{1}{\mu^2}$
Binomiale	$\eta = ln(\mu) - \ln(1-\mu)$

Tableau 3 : Fonctions de lien canonique de quelques lois appartenant à la famille exponentielle

Le modèle de régression linéaire devient alors un cas spécifique du GLM puisqu'il est possible de le retrouver en prenant :

- $g = Id$
- $\eta_i = \sum_{j=1}^{p} \beta_j X_{ij}$
- $Y_i \sim \mathcal{N}(\eta_i, \sigma^2)$

A titre d'exemple, la fonction de lien $g(\mu) = \ln(\mu)$ est souvent utilisée et permet de modéliser le logarithme de l'espérance. Les modèles utilisant cette fonction de lien sont des modèles log-linéaires.

En complément, le principe de l'estimation des paramètres et les différentes étapes de construction du GLM sont présentés en annexe.

L'approche par méthode linéaire généralisée est ainsi introduite. Les méthodes d'apprentissage supervisé vont être présentées dans la partie suivante. Il sera tout d'abord présenté l'algorithme CART. Cet algorithme sert de fondement aux méthodes ensemblistes que sont les forêts aléatoires et le *gradient boosting*.

2 - L’Arbre CART

Dans cette partie, une présentation de l’arbre CART sera effectuée. Les différentes étapes de mise en œuvre de l’algorithme seront ensuite explicitées. Enfin, une description des avantages et limites de l’algorithme sera décrite.

2.1 - La présentation de l’algorithme

L’acronyme anglais CART (*Classification And Regression Trees)* désigne une technique d’apprentissage supervisé, introduite par Breiman et al. (1984). Cette technique a pour objectif de construire des prédicteurs sous forme d’arbre aussi bien en classification qu’en régression. On parle d’arbre de classification lorsque la variable à expliquer est qualitative et d’arbre de régression lorsque cette même variable est quantitative. Il sera expliqué par la suite que le principe et la construction d’un arbre de régression sont les mêmes que pour les arbres de classification. Pour s’adapter à la nature quantitative ou qualitative de la variable à expliquer, seuls les outils mathématiques définissant la condition de coupure sont modifiés.

En principe, il existe plusieurs façons d’établir des arbres de décision CART en changeant la fonction de coût, la famille de coupures autorisée ou bien la règle d’arrêt. Cependant, par la suite une seule méthode sera abordée. Il s’agit de celle qui est la plus souvent utilisée, introduite dans l’œuvre de Breiman et al. (1984).

La méthode des arbres de décision se fonde sur la classification d’un individu par une suite de tests effectués sur un ensemble de variables explicatives qui le décrivent. Le terme d’arbre concernant l’algorithme CART se traduit par la notion de récursivité qui se retrouve au niveau des tests réalisés de manière hiérarchique. En effet la réponse d’un test possède une influence sur les tests suivants qui en découlent.

Le graphique ci-dessous représente la structure d’un arbre de décision CART :

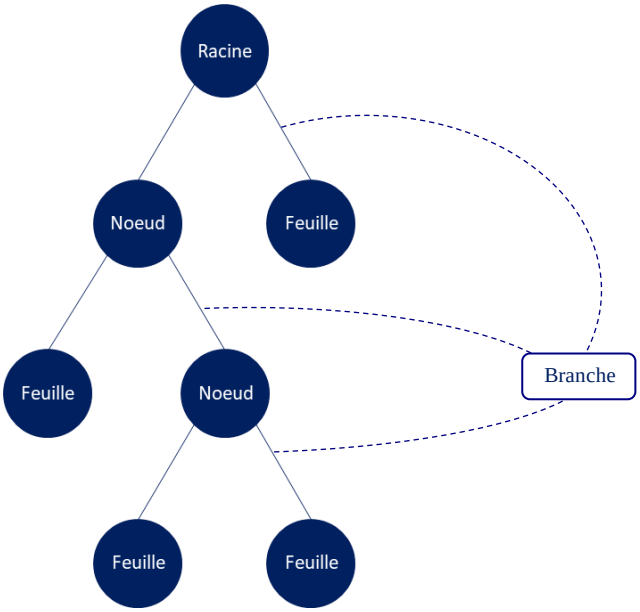


Figure 7 : Structure type d'un arbre CART

Les variables composant le vecteur $\boldsymbol{X}$ sont utilisées pour effectuer une discrimination ou bien une régression des observations sous la forme d'une ossature arborescente. Toutes les observations disponibles sont regroupées à la racine de l'arbre CART. Ensuite, chaque division subdivise chaque nœud de manière à obtenir deux nouveaux nœuds plus harmonieux que le nœud initial au sens d'un critère de division à préciser et qui dépend du type de Y, qualitative ou quantitative.

Une fois que le partitionnement s'interrompt, les nœuds terminaux de l'arbre maximal sont obtenus. Ces nœuds terminaux sont également appelés feuilles. A chacune des feuilles est attribuée :

- Une classe si Y est qualitative
- Une valeur si Y est quantitative

Le principe général de la méthode CART est de partitionner de façon récursive l'espace de manière binaire. Il convient ensuite de déterminer la sous-partition optimale pour la phase de prédiction. En considérant uniquement des variables explicatives quantitatives, la structure de l'arbre se traduit par un partitionnement dyadique de l'espace.

Un exemple d'arbre CART est représenté ci-dessous :

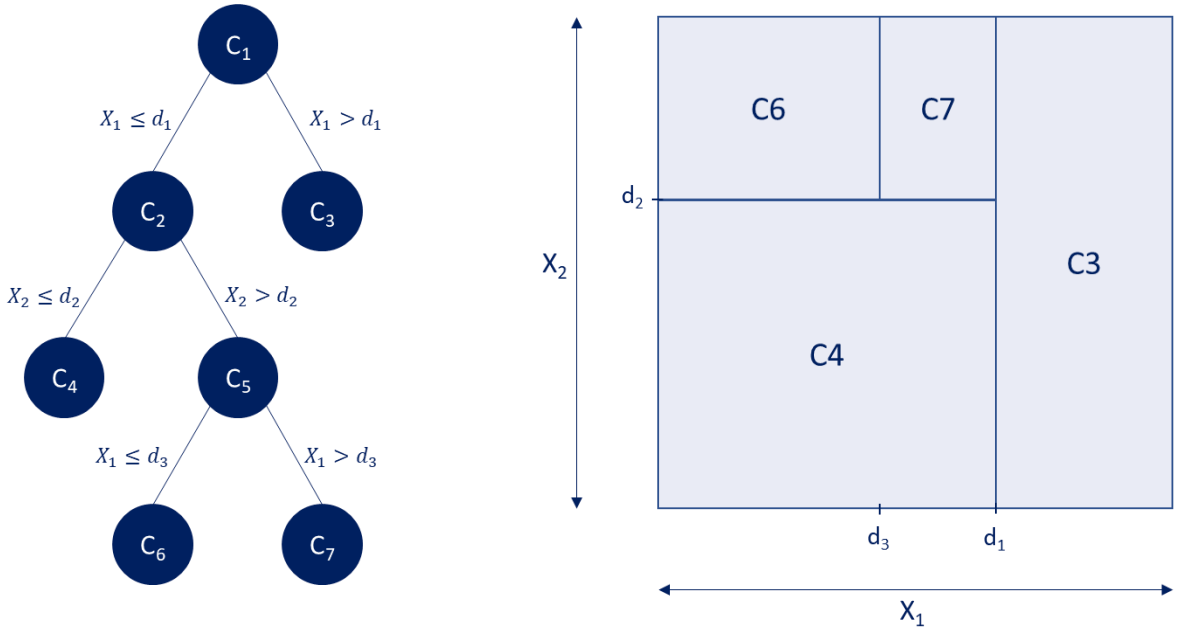


Figure 8 : Exemple d'arbre CART composé de deux variables explicatives quantitatives, X_1 et X_2, avec partitionnement dyadique de l'espace. Chaque nœud initial se sépare en deux branches conduisant à deux nouveaux nœuds par le biais d'une variable X_1ou X_2 et d'une valeur seuil, ici d_1, d_2, d_3 de manière successive. Ainsi à chaque zone C_3, C_4, C_6, C_7 est assimilée une classe ou une valeur de Y.

La construction d'un arbre CART se fait en général en deux étapes. La première est celle qui consiste à élaborer un arbre maximal, aussi appelé arbre saturé, afin de définir une famille de modèles pour laquelle il conviendra de trouver le meilleur. La deuxième étape est l'étape d'élagage. Elle consiste en la construction d'une suite de sous-arbres optimaux dits élagués de l'arbre saturé. La partie suivante portera alors sur l'explication et la présentation de ces deux étapes.

2.2 - La construction de l'arbre saturé

2.2.1 - Le principe

Considérons qu'il est observé p variables qualitatives ou quantitatives explicatives notées x_i et qu'il est à expliquer une variable Y quantitative ou qualitative à m modalités concernant un échantillon de n individus.

Le principe de cette méthode est la détermination d'une suite de nœuds. Pour obtenir un nœud, il est choisi parmi les variables explicatives une variable ainsi qu'une division qui permet une séparation en deux classes. Ainsi, il est appliqué à chaque nœud une dichotomie sur le sous-ensemble de l'échantillon correspondant. Une division est quant à elle déterminée soit par une séparation en deux ensembles des modalités si la variable est qualitative soit par une valeur seuil si la variable sélectionnée est quantitative. Le nœud initial, aussi appelé racine, correspond à l'échantillon initial auquel est appliqué cet algorithme sur chacun des sous-ensembles.

La procédure considérée implique :

- De définir une règle de découpe ou un critère de division qui a pour objectif de sélectionner la « meilleure » division parmi celles admissibles pour toutes les variables.
- De déterminer si un nœud est terminal. Cette mesure s'apparente à une règle d'arrêt.
- D'affecter chaque nœud terminal à une valeur ou à une des classes de la variable Y.

Les trois points ci-dessus seront développés dans la partie suivante.

2.2.2 - La règle de découpe

Une division est considérée admissible si aucun des nœuds fils descendant du nœud père n'est vide. Si la variable explicative x est qualitative nominale le nombre de divisions est de $2(m-1)-1$. Si cette dernière est qualitative ordinale ou quantitative, elle présente $(m-1)$ divisions admissibles.

Un point d'attention particulier est à noter sur le fait que l'algorithme tend à privilégier la sélection de variables explicatives présentant un grand nombre de modalités. L'explication à cette observation réside dans le fait que ces variables explicatives offrent une plus grande souplesse concernant la construction des deux sous-ensembles de l'échantillon.

Il sera donc par la suite conseillé d'utiliser ces variables avec parcimonie puisqu'elles tendent à favoriser un sur-apprentissage. Il est ainsi d'usage de fusionner des modalités afin de réduire leur nombre. La règle de découpe du noeud repose sur la définition d'une fonction de désordre, aussi appelée fonction d'hétérogénéité, qui sera présentée dans la section suivante. L'objectif est de séparer les individus en deux groupes les moins hétérogènes au sens de la variable Y à expliquer. A chaque nœud, l'hétérogénéité d'un nœud se mesure par une fonction positive ayant les caractéristiques suivantes :

- Nulle si, et seulement si, le nœud est homogène, c'est-à-dire que tous les individus prennent la même valeur de Y ou appartiennent à la même modalité.
- Maximale si les valeurs de Y possèdent la même probabilité ou sont très dispersées.

La séparation d'un nœud j crée deux nœuds fils, un nœud gauche et un droit. Dans un souci de simplification, ces deux nœuds sont notés J_g et J_d, cependant afin d'être conforme à la suite de sous-arbres, une renumérotation est nécessaire. L'algorithme retient, parmi les découpes admissibles du nœud

j, celle qui minimise la somme des hétérogénéités des nœuds fils, H_{J_g} et H_{J_d}. Il faut donc qu'à chaque étape j de l'obtention de l'arbre résoudre :

$$\max_{\{divisions\ de\ x_i\ ;i=1,\dots,p\}} H_J - (H_{J_g} + H_{J_g})$$

Avec H ; la fonction d'hétérogénéité.

2.2.3 - Le critère d'arrêt

Le critère d'arrêt est nécessaire afin de stopper la croissance de l'arbre à un nœud donné, devenant donc feuille. Il existe plusieurs cas pour lesquels le développement de l'arbre est stoppé à un certain nœud :

- S'il est homogène
- S'il n'existe plus de partition admissible
- Si le nombre d'observations qu'il comporte est inférieur à une valeur seuil (afin d'éviter un découpage trop fin)

2.2.4 - La règle d'affectation

L'étape d'affectation diffère en fonction de la nature de la variable à expliquer Y.

Lorsque Y est qualitative, chaque feuille est affectée à une classe de l'ensemble des modalités de Y en prenant en compte les critères conditionnels suivants :

- La classe qui est la mieux représentée dans la feuille. Il est ensuite possible de déduire la quantité d'objets mal affectés
- La classe qui est a posteriori la plus probable au sens bayésien si des probabilités sont connues a priori.
- Si des coûts de mauvais classements sont conférés, il est choisi la classe la moins coûteuse.

Lorsque Y est quantitative, chaque feuille correspond à un pavé de l'espace. La valeur affectée à un pavé de l'espace est la moyenne des observations associées à la feuille correspondante.

2.2.5 - Le critère d'homogénéité

Pour le critère d'homogénéité, il convient également de séparer lorsque la variable à expliquer est qualitative ou quantitative.

2.2.5.1 - Le cas de la régression

Dans le cas d'une variable Y quantitative, l'hétérogénéité du nœud J se traduit par la variance :

$$H_J = \frac{1}{\#J}\sum_{i\in J}(y_i - \bar{y_J})^2$$

Avec :

- $\#J$, l'effectif du nœud J.

Pour chaque nœud, il faut rechercher la séparation, à savoir la variable ainsi que la règle de découpe, qui permettra d'obtenir la plus forte baisse d'hétérogénéité pour les deux nœuds fils J_g et J_d. Il convient donc de minimiser l'expression de la variance intra-groupe suivante :

$$\frac{\#J_g}{\#J}\sum_{i\in J_g}(y_i-\bar{y}_{J_g})^2+\frac{\#J_d}{\#J}\sum_{i\in J_d}(y_i-\bar{y}_{J_d})^2$$

2.2.5.2 - Le cas de la classification

Dans le cas d'une variable Y qualitative, où l'ensemble des classes est $\{1,\dots,L\}$, plusieurs fonctions d'hétérogénéité peuvent être définies pour un nœud. La concentration de Gini est la fonction traditionnellement utilisée :

$$H_J=\sum_{c=1}^{L}p_J^c(1-p_J^c)$$

Comme pour le cas de la régression, il est nécessaire de rechercher pour chaque nœud dans toutes les découpes admissibles laquelle maximise la décroissance de l'hétérogénéité.

Ce processus permet d'obtenir un arbre maximal dit « saturé ». Toutefois cet arbre est rarement optimal puisque sujet au sur-apprentissage. Une étape d'élagage est donc nécessaire afin d'améliorer le modèle.

2.3 - L'étape d'élagage

La méthode de construction décrite dans la partie précédente permet d'obtenir un arbre A_{max} à k feuilles qui dans la plupart des cas est trop sophistiqué. Cet arbre A_{max} conduit à un modèle de prédiction très instable puisqu'il est grandement dépendant des échantillons d'observation qui ont permis son estimation. Cela correspond à un problème de surajustement qui est à éviter. Il est préférable de favoriser un modèle plus parcimonieux qui permet d'obtenir des prévisions plus robustes. Pour ce faire, une étape d'élagage de l'arbre maximal A_{max} est effectuée. Il s'agit de trouver parmi tous les sous-arbres un arbre optimal entre l'arbre maximal et l'arbre composé uniquement de la racine qui contient toutes les observations.

Considérer tous les sous-arbres de A_{max} semble peu concevable puisque leurs nombres suivent une croissance exponentielle. Afin d'éviter ce problème, Breiman et al. (1984) ont établi une démarche qui consiste à construire une séquence emboîtée de sous-arbres de l'arbre saturé. Ensuite, parmi cette séquence, l'arbre optimal qui minimise l'erreur de généralisation est choisi.

2.3.1 - La construction de la suite d'arbre

Considérons un arbre A et notons T_A le nombre de feuilles de cet arbre. La valeur de T_A traduit la complexité de l'arbre A.

Il est ainsi possible d'exprimer la qualité d'ajustement de A comme suit :

$$H(A) = \sum_{J=1}^{T_A} H_J$$

Avec :

- H_J ; représentant l'hétérogénéité du nœud terminal J de l'arbre A.

Ainsi, en régression, elle correspond à la variance intra-groupe et en classification la concentration de Gini peut être utilisée.

Pour construire la suite emboîtée de sous-arbres, il est introduit une notion de pénalisation de la complexité de A :

$$C(A) = H(A) + \gamma \times T_A$$

Pour $\gamma = 0$; la relation $C(A) = H(A)$ est retrouvée et ainsi $A_{max} = A_{T_A}$. Dans ce cas, la complexité $C(A)$ est minimisée.

Lorsque $\gamma > 0$, il est repéré une découpe de A_{T_A} pour laquelle la décroissance de l'hétérogénéité H est la plus faible. Cette découpe est ainsi considérée superflue et les deux nœuds terminaux fils sont élagués (regroupés) dans le nœud père qui devient lui-même terminal. Il a été rassemblé deux feuilles en une et ainsi A_{T_A} devient A_{T_A-1}. Ce processus est ainsi itéré et permet de construire la suite emboîtée :

$$A_{max} = A_{T_A} \supset A_{T_A-1} \supset \cdots \supset A_1$$

Avec A_1 ; l'arbre trivial uniquement composé de la racine regroupant l'ensemble des observations

Il convient ensuite de tracer le graphique représentant la décroissance des valeurs de H_J en fonction du nombre croissant de nœuds terminaux dans l'arbre.

2.3.2 - La recherche de l'arbre optimal

Il s'agit ici de rechercher parmi la suite emboîtée de sous-arbres, établie dans la section précédente, celui qui minimalise l'erreur de généralisation.

Si la taille de l'échantillon étudié est assez importante il convient de préalablement extraire un échantillon de validation afin d'estimer l'erreur de généralisation. Dans le cas contraire, une validation croisée est préférée.

2.4 - Les avantages et les inconvénients de l'algorithme CART

L'arbre CART présente de nombreux avantages que ce soit au niveau de l'interprétabilité, de la résistance aux valeurs aberrantes et du temps de calcul. En effet, l'arbre CART permet d'obtenir une grande interprétabilité, contrairement au Random Forest (RF) qui sera présenté dans la section suivante. Dès lors que l'arbre CART est établi, il est à considérer que les variables qui interviennent dans les découpes des nœuds de l'arbre (notamment les nœuds les plus proches de la racine) sont les variables les plus significatives. En effet, plus une division est proche du nœud racine plus la baisse de l'hétérogénéité est importante. En ce sens l'algorithme CART permet d'obtenir une mesure de

l'importance des variables dans le modèle. La résistance naturelle aux valeurs aberrantes s'explique en majeure partie par le fait que la présence d'une donnée aberrante dans l'échantillon d'apprentissage va impacter essentiellement la feuille qui la contient et laisser une trace réduite pour les autres. Enfin l'algorithme CART qui permet de construire l'arbre présente une faible complexité informatique et induit donc un faible temps de calcul.

Il a été vu que la facilité de l'interprétation de l'arbre CART est son atout majeur. Cependant, la contrepartie est que cette méthode fournit des modèles particulièrement instables, sensibles à des variations de l'échantillon. Afin de pallier ce problème, il est introduit dans les sections suivantes plusieurs méthodes ensemblistes construites sur la base d'arbres CART. Ces méthodes peuvent reposer sur un principe d'agrégation (*Random Forest*) ou un principe séquentiel adaptatif (*Gradient Boosting*).

3 - Les méthodes d'agrégations

Dans la section précédente, les arbres CART ont été présentés et il s'est avéré qu'ils pouvaient être instables. Cette caractéristique induit que le modèle estime des prédictions qui varient beaucoup. Cela s'explique par le fait que plus l'arbre est long, plus les estimations données par les feuilles dépendent de l'échantillon de départ et de certains profils en particulier. Dans cette configuration, les estimations seront correctes pour des profils proches de ceux présents dans l'échantillon d'apprentissage mais elles seront majoritairement erronées lorsqu'il s'agira de prédire des observations différentes que celles de l'échantillon d'apprentissage.

La principale problématique des méthodes d'agrégation est de limiter le surajustement aux données de l'échantillon qui sert à construire le modèle en proposant une plus grande stabilité dans les résultats. Il a été pensé pour cela des méthodes dites d'agrégation de modèle. Ces méthodes reposent sur la construction de plusieurs arbres qui sont ensuite agrégés pour obtenir un seul estimateur plus fiable possédant une variance plus faible.

Il sera présenté dans les sections suivantes des méthodes d'agrégation parallèles assemblant des arbres construits de manière indépendante. Il sera ensuite étudié des méthodes d'agrégation adaptatives qui, à la différence de la méthode précédente, construisent des arbres dépendant les uns des autres de manière hiérarchique. Dans les méthodes parallèles, il peut être nommé les forêts aléatoires ainsi que le bagging tandis que les méthodes adaptatives sont regroupées sous le nom de boosting.

L'objectif de cette partie est de présenter les modèles cités précédemment tout en utilisant la méthode d'arbre CART étudiée dans la section précédente.

3.1 - Le bagging

Le *bagging* correspond à un ensemble de méthodes établi en 1996 par Léo Breiman. Le terme *bagging* vient de la contraction des termes *Bootstrap* et *Aggregating*. Cette méthode est, dans un premier temps, présentée en considérant une variable à expliquer quantitative. Dans un second temps, cette méthode sera étendue au cas d'une variable qualitative et donc à une classification.

3.1.1 - Le principe de l'algorithme

Soit (X, Y) un vecteur aléatoire, X appartient à $\mathbb{R}^p$ et Y appartient à $\mathbb{R}$.

Soit, $\mathcal{L}_n = (X_1, Y_1), \dots, (X_n, Y_n)$ un échantillon de taille n indépendant et identiquement distribué et ayant la même loi que (X, Y).

Soit, $\phi(x) = \mathbb{E}(Y|X = x)$, une fonction de régression.

Cette fonction de régression repose sur un classifieur faible qui peut être par exemple l'algorithme d'un arbre CART ou bien l'algorithme de 1 plus proche voisin.

Avec X appartient à $\mathbb{R}^p$, l'erreur quadratique moyenne de l'estimateur $\hat{\phi}$ permet d'obtenir une décomposition biais/variance :

$$\mathbb{E}\left(\left(\hat{\phi}(x) - \phi(x)\right)^2\right) = (\mathbb{E}(\hat{\phi}(x)) - \phi(x))^2 + \mathbb{V}(\hat{\phi}(x))$$

La méthode de *bagging* est une méthode d'agrégation car elle repose sur l'agrégation d'un nombre B d'estimateurs $\hat{\phi}_1, \dots, \hat{\phi}_B$ ayant pour objectif d'obtenir un estimateur agrégé $\hat{\phi} = \frac{1}{B}\sum_{k=1}^{B} \hat{\phi}_k$.

En considérant les estimateurs, $\hat{\phi}_1, \dots, \hat{\phi}_B$, indépendants et identiquement distribués, les relations suivantes sont vérifiées :

$$\mathbb{E}\left(\hat{\phi}(x)\right) = \mathbb{E}\left(\hat{\phi}_1(x)\right)$$

$$\mathbb{V}\left(\hat{\phi}(x)\right) = \frac{1}{B}\mathbb{V}\left(\hat{\phi}_1(x)\right)$$

Ainsi, le biais de l'estimateur agrégé $\hat{\phi}$ est égal à celui des estimateurs $\hat{\phi}_k$. Cependant, la variance de l'estimateur agrégé $\hat{\phi}$ diminue avec le nombre d'estimateurs construits. Ces résultats se vérifient lorsque que les estimateurs $\hat{\phi}_1, \dots, \hat{\phi}_B$ sont indépendants (et de même loi). Néanmoins dans la pratique, il est difficile si ce n'est impossible d'obtenir des estimateurs $\hat{\phi}_k$ indépendants du fait qu'ils résultent tous de l'échantillon $\mathcal{L}_n$. La méthode bagging permet de réduire la dépendance entre les différents estimateurs à agréger en les obtenant via des échantillons *Bootstrap.*

Le graphique représenté ci-dessous représente la structure d'un algorithme de *bagging* :

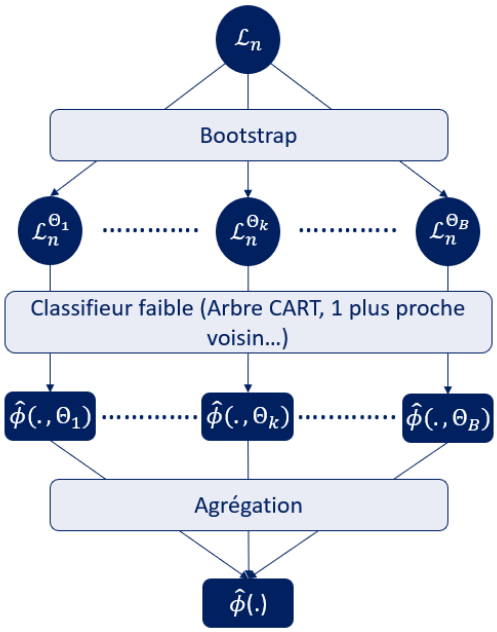


Figure 9 : Structure de l'approche par bagging

Cette structure permet d'exprimer comment l'algorithme est implémenté. Il convient d'abord d'établir B estimateurs $\hat{\phi}(.,\Theta_1),\dots,\hat{\phi}(.,\Theta_B)$ sur des échantillons obtenus par Bootstrap $\mathcal{L}_n^{\Theta_1},\dots,\mathcal{L}_n^{\Theta_p}$ de l'échantillon initial $\mathcal{L}_n$. La notation $\mathcal{L}_n^{\Theta_k}$décrivant l'échantillon Bootstrap de l'étape k permet de prendre en compte la nouvelle source d'aléa introduite.

De la même manière, l'estimateur agrégé final s'écrit :

$$\hat{\phi}^B(\boldsymbol{X}) = \frac{1}{B}\sum_{k=1}^{B}\hat{\phi}(\boldsymbol{X},\Theta_k)$$

Avec :

- $\hat{\phi}(\boldsymbol{X},\Theta_k)$; l'estimateur construit à l'étape k.
- $\hat{\phi}^B(\boldsymbol{X})$; l'estimateur agrégé final pour B classifieurs faibles.

Les échantillons *bootstrap* sont réalisés de telle sorte qu'ils sont indépendants deux à deux. Ces échantillons permettent ainsi d'obtenir des estimateurs moins dépendants que s'ils étaient issus du même échantillon d'apprentissage.

Ces résultats sont valables dans un contexte de régression. Dans le cas d'une variable qualitative, l'agrégation de l'estimateur baggé diffère. En effet, en classification, l'agrégation se traduit par un vote à la majorité des prédicteurs.

3.1.2 - La dualité biais/variance

Dans cette section, la comparaison des estimateurs que l'on agrège à celui agrégé est présentée. Les égalités suivantes sont vérifiées :

- $\hat{\phi}^B(\boldsymbol{X}) = \frac{1}{B}\sum_{k=1}^{B}\hat{\phi}(\boldsymbol{X},\Theta_k)$; l'estimateur agrégé final pour B classifieurs faibles.
- $\hat{\phi}(\boldsymbol{X}) = \lim_{B\to+\infty}\hat{\phi}^B(\boldsymbol{X})$; l'estimateur agrégé final pour un nombre infini de classifieurs faibles.
- $\sigma^2(\boldsymbol{X}) = \mathbb{V}(\hat{\phi}(\boldsymbol{X},\Theta_k))$; la variance des estimateurs qui sont agrégés.
- $\rho(\boldsymbol{X}) = cor\left(\hat{\phi}(\boldsymbol{X},\Theta_k),\hat{\phi}(\boldsymbol{X},\Theta_{k+1})\right)$; le coefficient de corrélation entre deux estimateurs qui sont agrégés.

La variance ainsi que la corrélation sont établies en fonction des lois de Θ et de $\mathcal{L}_n$. En supposant que, $\forall k \in [1\!:\!B]$, $\hat{\phi}(\boldsymbol{X},\Theta_k)$ sont identiquement distribués, le biais des estimateurs à agréger est égal au biais de l'estimateur agrégé. Ainsi, l'agrégation n'influe pas sur le biais.

Concernant la variance, l'égalité suivante est vérifiée :

$$\mathbb{V}\left(\hat{\phi}^B(\boldsymbol{X})\right) = \rho(\boldsymbol{X})\sigma^2(\boldsymbol{X}) + \frac{1-\rho(\boldsymbol{X})}{B}\sigma^2(\boldsymbol{X})$$

En faisant tendre le nombre d'échantillons *bootstrap* B vers l'infini, l'égalité suivante est obtenue :

$$\mathbb{V}\left(\hat{\phi}(\boldsymbol{X})\right) = \rho(\boldsymbol{X})\sigma^2(\boldsymbol{X})$$

La formule ci-dessus induit que pour un nombre d'échantillons *bootstrap* B suffisamment grand, l'estimateur agrégé possède une variance inférieure à celles des estimateurs qui sont agrégés si le

coefficient de corrélation $\rho(\boldsymbol{X})$ est inférieur à 1. Une solution intéressante pour obtenir un bon prédicteur serait de choisir des estimateurs possédant un biais réduit. Néanmoins cette solution ne parait pas pertinente puisque cela reviendrait à choisir des estimateurs avec une forte variance. Or, comme cette équation le démontre, un estimateur baggé avec une forte variance serait obtenu car la variance ne serait pas assez réduite par le coefficient de corrélation $\rho(\boldsymbol{X}) < 1$.

Il en découle que le gain de la procédure d'agrégation réside dans le coefficient de corrélation $\rho(\boldsymbol{X})$. En effet, plus les estimateurs à agréger seront décorrélés plus la réduction de la variance sera grande. L'utilisation de la méthode *bootstrap* dans la construction des échantillons corrobore cette idée.

Un point d'attention existe concernant les estimateurs à agréger. Ces derniers nécessitent d'être modifiés du fait des différents échantillons *bootstrap*. En effet, si les prédicteurs sont robustes à des modifications de l'échantillon d'apprentissage, la décorrélation des estimateurs reste réduite et donc l'approche par *bagging* ne peut apporter d'amélioration.

Les arbres CART étant des prédicteurs très instables, la méthode par *bagging* se révèle être efficace pour ce cas précis. Afin d'augmenter la décorrélation des différents arbres CART dans la méthode de *bagging*, et donc la performance de l'algorithme, la forêt aléatoire ajoute une étape appelée randomisation. Cette étape permet à l'algorithme de gagner en performance, elle sera développée plus en détail dans la partie suivante.

3.2 - La forêt aléatoire

La forêt aléatoire, ou, sous sa traduction anglaise plus communément utilisée, le *random forest*, consiste de manière générique en la construction d'une multitude d'arbres de décision. Le résultat est ensuite obtenu en effectuant l'agrégation des estimateurs correspondant à chacun des arbres.

3.2.1 - Le principe de l'algorithme

Le graphique ci-dessous représente la structure d'une forêt aléatoire triviale :

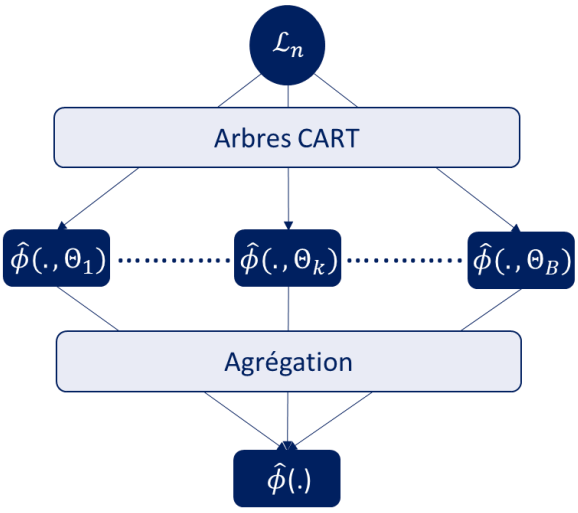


Figure 10 : Structure d'une forêt aléatoire triviale

Une forêt aléatoire se définit donc simplement comme une agrégation d'arbres de décision qui dépendent de certaines variables aléatoires. En reprenant ce qui a été fait dans la section précédente, le bagging permet d'obtenir une forêt aléatoire dans le cas où le classifieur faible choisi est l'algorithme de l'arbre CART.

Il existe donc plusieurs structures différentes de forêts aléatoires. Cependant, une structure de forêt aléatoire s'est fait une place particulière par rapport aux autres du fait de sa qualité d'apprentissage et de prédiction. Cette structure est le *random forest RI*. Étant donné son utilisation répandue, l'appellation de forêt aléatoire est souvent utilisée pour uniquement désigner la structure spécifique des *random forests RI*. C'est ce choix qui a été retenu dans les parties suivantes.

L'algorithme de *random forest RI* relève du même principe que la méthode du bagging. En effet, un *bootstrap* sur l'échantillon d'apprentissage est également effectué de sorte à obtenir de nouvelles observations. Comme c'est le cas pour toutes les forêts aléatoires, le classifieur faible retenu est l'algorithme de l'arbre CART.

La subtilité de cette approche se situe dans la construction de l'arbre maximal. A chaque nœud de l'arbre, au lieu de considérer toutes les variables explicatives disponibles, m variables sont choisies aléatoirement. Ensuite, la meilleure division est sélectionnée parmi ces m variables uniquement. Ce tirage est effectué avant chaque division des nœuds des arbres.

L'objectif de ce procédé est de faire décroître la corrélation entre les différents arbres construits et ainsi d'obtenir une variance finale réduite. En effet, l'introduction de cet aléa dans la construction des estimateurs proposés par Breiman réduit leur corrélation. De plus, cette méthode prenant en compte moins de variables explicatives à chaque division de l'arbre permet de réduire le temps de calcul par rapport au bagging.

Le graphique ci-dessous représente la structure d'une forêt aléatoire avec randomisation de m variables explicatives à chaque nœud par les p variables disponibles :

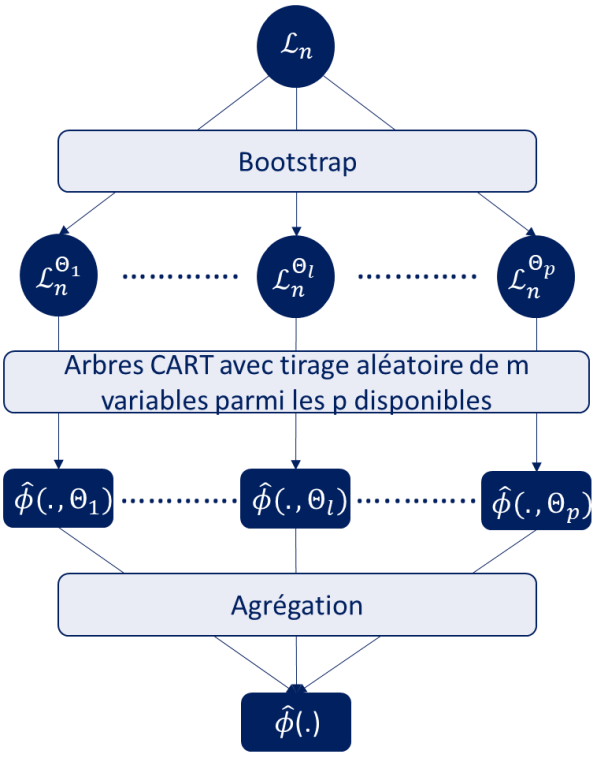


Figure 11 : Structure d'une forêt aléatoire avec tirage aléatoire de m *inputs*

La dualité biais/variance évoquée pour la méthode du bagging se retrouve dans cette structure de Forêt aléatoire. En effet, le choix du nombre de variables m à considérer à chaque division est au cœur de cet enjeu.

Plus m est choisi petit, moins les variables choisies pour effectuer les divisions à chaque nœud se révèlent pertinentes. En prenant par exemple le cas extrême où une seule variable est sélectionnée à chaque nœud pour établir la division (m = 1), cette situation revient à choisir des axes de découpes des arbres de manière aléatoire. Il est également à noter que plus m est petit plus les arbres sont décorrélés les uns des autres plus la variance de l'estimateur baggé diminue. L'effet recherché sur la variance est de cette manière atteint. Cependant le choix d'un m trop petit dégrade la qualité d'ajustement des prédicteurs aux données de l'échantillon d'apprentissage Bootstrap. Cela a pour impact d'augmenter le biais de chaque prédicteur et par conséquent le biais du prédicteur agrégé. Les évolutions inverses sont observées sur le biais et la variance lorsque le nombre m choisi augmente.

Les paramètres des arbres de décision exercent une influence sur le choix du nombre de variables m à considérer. En effet, à titre d'exemple, la règle d'arrêt qui correspond au nombre d'observations dans les feuilles impacte le biais et la variance des arbres. Plus le nombre d'observations dans les feuilles est grand plus le biais est grand, moins la variance est élevée et donc moins l'agrégation est performante. Ainsi le nombre maximal d'observations dans les feuilles est choisi petit pour maximiser la variance dans un contexte d'agrégation.

Considérons p, le nombre de variables explicatives, la fonction *randomForest* de R choisit par défaut :

- Dans le cas d'une régression ; $m = \frac{p}{3}$
- Dans le cas d'une classification ; $m \approx \sqrt{p}$

Il est à souligner que l'instabilité de l'algorithme CART, qui est son plus gros défaut dans le cadre d'un arbre de décision seul, est contre intuitivement l'élément pour lequel l'agrégation permet une amélioration des performances.

L'interprétabilité d'une forêt aléatoire n'est pas directe comme c'est le cas pour un arbre CART. Des indicateurs d'analyses ont donc été développés afin de pouvoir interpréter le modèle construit. Ils sont explicités dans la partie suivante.

3.2.2 - La définition des indicateurs d'analyse

Il existe deux indicateurs qui permettent de donner une analyse pertinente d'une forêt aléatoire :

- L'erreur *Out of Bag*
- L'importance des variables

3.2.2.1 - L'erreur *Out Of Bag*

L'erreur *Out of Bag* (OOB) est une méthode qui construit un prédicteur des erreurs. Cette procédure peut être utilisée pour toutes les méthodes de *Bagging*. L'estimateur est différent en fonction de la nature de la variable à expliquer Y :

- Dans le cas d'une classification, l'estimateur s'écrit $\mathbb{P}(\widehat{m}(\boldsymbol{X}) \neq Y)$
- Dans le cas d'une régression, l'estimateur s'écrit $\mathbb{E}[(\widehat{m}(\boldsymbol{X}) - Y)^2]$

L'approche *Out Of Bag* possède cet avantage qu'elle ne demande pas une découpe de l'échantillon. Elle s'appuie sur le fait que les arbres sont construits sur des échantillons baggés. Ils n'utilisent donc pas toute l'information de l'échantillon d'apprentissage.

Soit (Y_i, X_i), une observation de $\mathcal{L}_n$. Il convient de définir un ensemble $\mathcal{E}_B$ qui contient uniquement les arbres de la forêt n'ayant pas l'individu (Y_i, X_i) dans leur échantillon Bootstrap.

Dans un contexte de régression, les arbres appartenant à l'ensemble $\mathcal{E}_B$ sont agrégés de la manière suivante :

$$\hat{Y}_i = \frac{1}{\#\mathcal{E}_B} \sum_{k \in \mathcal{E}_B} \phi(X_i, \Theta_k)$$

Dans la situation d'une classification, la prédiction est obtenue sur un vote à la majorité en considérant uniquement les arbres appartenant à l'ensemble $\mathcal{E}_B$.

Finalement, l'erreur *Out of Bag* notée E_{OOB} s'écrit :

- Si Y est qualitative, $E_{OOB} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$
- Si Y est quantitative, $E_{OOB} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\hat{Y}_i \neq Y_i}$

Le second indicateur d'analyse présenté dans la partie suivante est l'importance des variables.

3.2.2.2 - L'importance des variables

Contrairement à l'algorithme CART où il est possible de connaître l'importance des variables dans le modèle à la simple lecture de l'arbre, la forêt aléatoire ne permet pas une telle interprétation. Une définition d'un critère permettant d'expliciter l'importance de chaque variable dans la construction du modèle est donc nécessaire.

Il convient pour cela d'introduire l'échantillon OOB assimilé au k^{ème} arbre du randomForest qui sera noté OOB_k. Cet ensemble est composé des individus qui ne sont pas dans l'échantillon Bootstrap du k^{ème} arbre de la forêt.

L'erreur E_{OOB_k} mesurée sur l'échantillon k s'écrit comme suit :

$$E_{OOB_k} = \frac{1}{\#OOB_k} \sum_{i \in OOB_k} (\phi(X_i, \Theta_k) - Y_i)^2$$

Il est également nécessaire de définir l'échantillon OOB_k^j qui correspond à l'ensemble OOB_k avec une variable j dont les valeurs ont été modifiées de manière aléatoire. De la même façon $E_{OOB_k^j}$ désigne l'erreur de prévision de l'arbre $\phi(., \Theta_k)$ sur l'échantillon OOB_k^j :

$$E_{OOB_k^j} = \frac{1}{\#OOB_k} \sum_{i \in OOB_k^j} (\phi(X_i^j, \Theta_k) - Y_i)^2$$

Avec X_i^j ; les données perturbées de l'échantillon OOB_k^j.

L'idée est d'utiliser l'instabilité de l'arbre CART pour mesurer l'importance d'une variable. En effet, si une variable j est importante dans la construction de l'arbre alors la perturbation de cette dernière va augmenter de manière considérable l'erreur OOB. Il convient alors d'effectuer la différence entre les erreurs $E_{OOB_k^j}$ et E_{OOB_k} pour obtenir l'importance de la variable j dans l'arbre k. Enfin, afin d'obtenir l'importance de la variable j dans la forêt, la moyenne de toutes ces différences est calculée de la manière suivante :

$$Importance(X_j) = \frac{1}{B}\sum_{k=1}^{B}(E_{OOB_k^j} - E_{OOB_k})$$

De cette manière, l'importance des variables permet d'interpréter l'influence de chaque variable dans le modèle construit.

La forêt aléatoire, une méthode ensembliste d'apprentissage supervisé, vient d'être présentée. Elle repose sur l'agrégation de classifieurs faibles que sont les arbres CART. Dans la partie suivante, une autre méthode ensembliste qui repose non pas sur l'agrégation mais sur une approche séquentielle adaptative sera explicitée.

4 - Le *gradient boosting*

La méthode de boosting est une autre méthode ensembliste d'apprentissage supervisé. Cette technique ne repose pas sur la réduction de la variance des classifieurs faibles comme dans la méthode de bagging sur lesquels reposent les forêts aléatoires. Une différence majeure entre ces deux techniques est également le fait que concernant les méthodes de bagging, les différents classifieurs faibles sont générés de manière indépendante et leur poids est égal d'un classifieur à l'autre. A l'inverse, la méthode du boosting est un processus séquentiel dans lequel chaque modèle est issu d'un précédent modèle avec pour objectif d'améliorer les performances de l'ensemble des modèles précédents. Le boosting utilise des moyennes pondérées des différentes fonctions de perte. Les deux méthodes ont ceci en commun qu'elles génèrent une grande quantité de modèles sur l'échantillon d'apprentissage. Ces différents modèles sont ensuite utilisés pour améliorer la précision du modèle final en les combinant les uns aux autres.

4.1 - L'approche séquentielle de l'algorithme

De manière graphique, les méthodes de gradient boosting peuvent être représentées de la manière suivante :

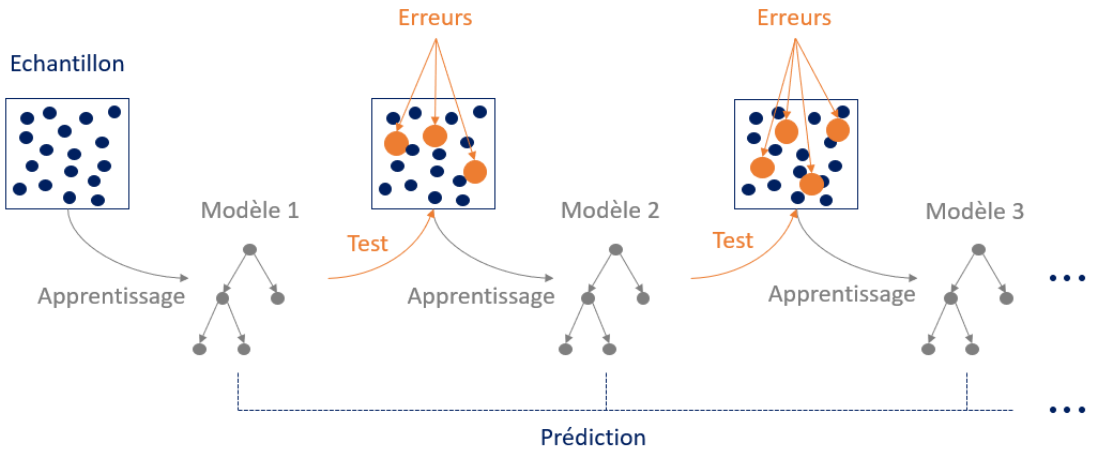


Figure 12 : Représentation de l'approche séquentielle du gradient boosting

En théorie, il est possible de choisir n'importe quel classifieur faible dans la méthode du gradient boosting. En pratique, il s'avère que le classifieur faible classiquement utilisé est l'arbre de régression. Il a également été fait le choix de l'arbre CART comme classifieur faible par la suite. De manière générale, les arbres utilisés dans la méthode du *gradient boosting* ne comptent qu'un à six nœuds. L'intérêt d'utiliser des classifieurs faibles à la place de classifieurs robustes est multiple :

- La rapidité : La construction des classifieurs faibles ne requiert pas un important temps de calcul.
- L'amélioration de la précision : Les classifieurs faibles permettent à l'algorithme GBM d'apprendre relativement lentement. Ceci permet de faire uniquement des ajustements mineurs pour les zones pour lesquelles l'algorithme ne performe pas bien.
- Limitation du surapprentissage : Ce point est lié au point précédent sur le fait que les petites incrémentations d'amélioration pour chaque modèle sur l'ensemble permettent d'arrêter la phase d'apprentissage dès lors que le surapprentissage est détecté. Le surapprentissage est repéré de la même façon que pour la méthode *random forest*, c'est-à-dire par validation croisée.

Dans l'algorithme de *gradient boosting*, les arbres CART sont construits de manière séquentielle. Chaque arbre utilise les informations des arbres construits précédemment. L'algorithme pour les arbres de régression boostés peut être généralisé par les étapes suivantes :

1. Construire un arbre de décision sur l'échantillon initial : $F_1(x) = \hat{y}$
2. Construire l'arbre de décision suivant en fonction des résidus du précédent : $h_1(x) = y - F_1(x)$
3. Ajouter le nouvel arbre à l'algorithme :$F_2(x) = F_1(x) + h_1(x)$
4. Construire l'arbre de décision suivant en fonction des résidus de F_2 : $h_2(x) = y - F_2(x)$
5. Ajouter le nouvel arbre à l'algorithme :$F_3(x) = F_2(x) + h_2(x)$
6. Répéter cette méthodologie tant que la validation croisée n'indique pas d'arrêter.

L'algorithme basique pour les arbres de régression boosté peut donc être généralisé de la manière suivante :

$$f(x) = \sum_{k=1}^{K} f^k(x)$$

De cette façon le modèle final est simplement un modèle additif par étape de k arbres de régression.

Les algorithmes tels que les arbres CART ainsi que les forêts aléatoires reposent sur la réduction de la fonction de perte *Mean Squared Error*. Cependant il peut être utilisé une autre fonction de perte comme la fonction de perte *Mean Average Error*. Le nom *gradient boosting* vient du fait que cette procédure peut être généralisée pour des fonctions de perte différentes de *Mean Squared Error*. Il a été tout de même fait le choix d'utiliser la fonction de perte *Mean Squared Error* par la suite.

4.2 - La descente du gradient

La méthode du *gradient boosting* s'inscrit dans la catégorie des algorithmes de descente de gradient. La descente de gradient est un algorithme générique qui a pour caractéristique de pouvoir trouver des solutions optimales pour une grande sélection de problèmes. L'idée fondamentale de la descente de gradient est de tordre les paramètres de manière itérative afin de minimiser une fonction de perte. Ainsi, la descente du gradient mesure la fonction de perte locale du gradient pour un ensemble donné de paramètres puis d'itération en itération se dirige vers la direction du gradient descendant. Le minimum est atteint lorsque le gradient atteint zéro.

Le graphique ci-dessous représente le phénomène de descente du gradient :

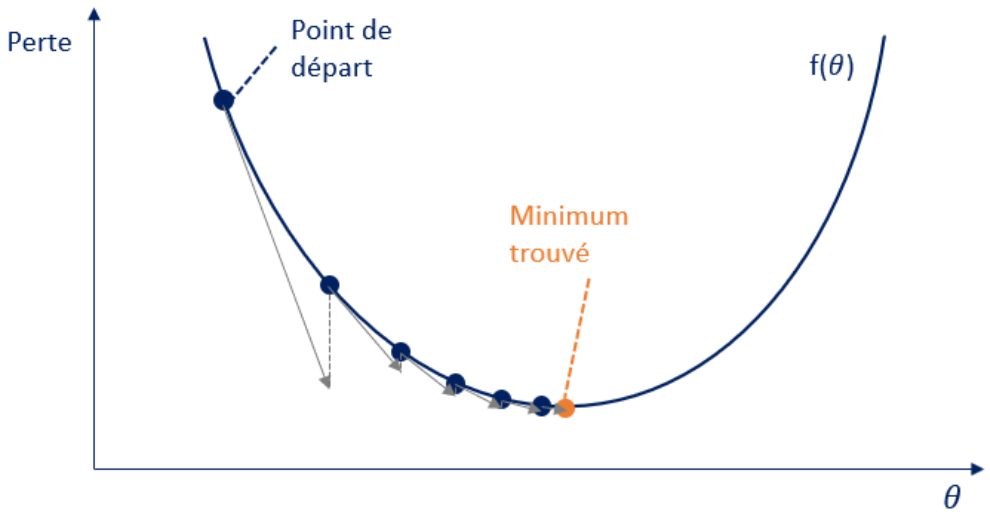


Figure 13 : Processus de la descente du gradient

Les flèches représentent les vecteurs de gradient dont la direction, le sens et la valeur indiquent la position du prochain point qui produit la plus grande décroissance sur la courbe de la fonction de perte. La méthode de descente du gradient peut être appliquée à n'importe quelle fonction de perte différentiable. Un paramètre important dans le processus de descente du gradient est le poids des corrections apportées à chaque étape. Ce paramètre est appelé le taux d'apprentissage ou en anglais *learning rate*. La calibration de ce taux est fondamentale. S'il est choisi trop grand le minimum de la fonction de perte ne pourrait être dépassé sans le repérer et même pire pourrait obtenir un résultat moins

satisfaisant que le point de départ. A l'inverse, s'il est trop petit, le temps d'exécution et le nombre d'itérations pourraient être trop conséquents pour trouver le minimum.

Les graphiques ci-dessous représentent ces deux situations :

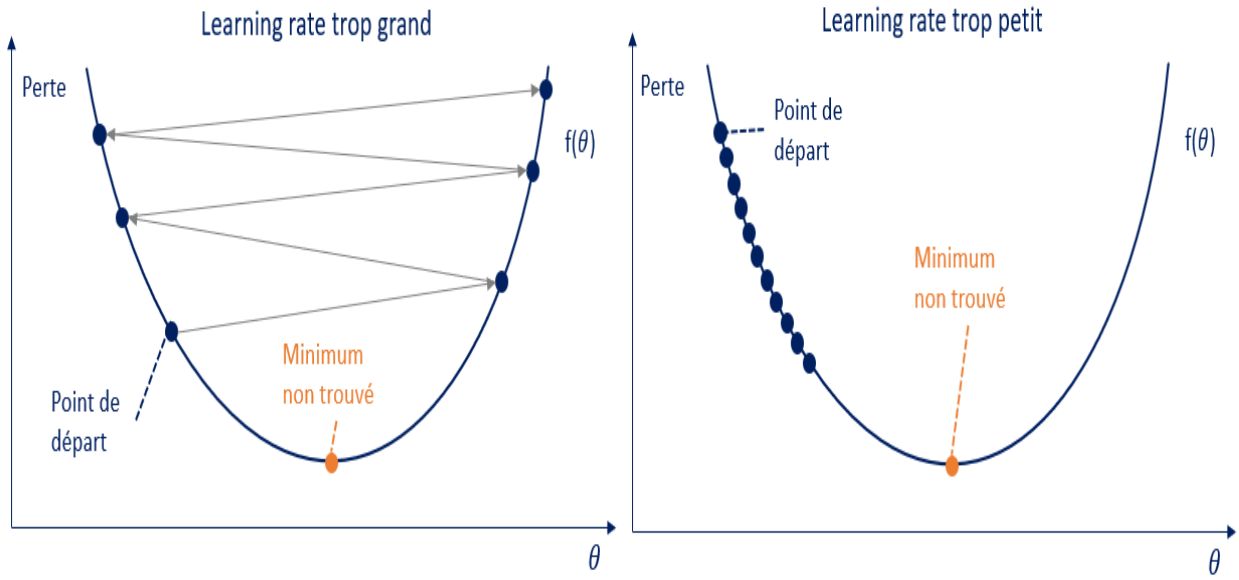


Figure 14 : Influence du réglage du *learning rate*

4.3 - Le *stochastic gradient boosting*

Il peut arriver que le minimum global soit difficile à atteindre par la méthode du gradient boosting classique du fait des irrégularités de la fonction de perte. Le *stochastic gradient boosting* a été développé afin de surmonter cette difficulté. Cet algorithme plus élaboré consiste à n'utiliser qu'un sous-échantillon de la base d'apprentissage pour construire l'arbre CART suivant. Cela permet à l'algorithme de gagner en temps d'exécution mais la nature stochastique de l'échantillonnage aléatoire ajoute de l'aléa dans la descente du gradient de la fonction de perte. Cette méthode ne permet pas forcément de trouver le minimal global mais permet de passer d'un minimum local à l'autre afin de se rapprocher de ce minimum global. La fonction utilisée pour faire appel à cet algorithme est la fonction *gbm* sous R.

Il convient d'ajouter quelques remarques générales sur les caractéristiques de cet algorithme. Le gradient boosting s'avère être en général l'algorithme qui permet d'apporter les meilleures prédictions en comparaison aux modèles précédemment introduits. Il s'adapte particulièrement bien aux données qu'elles soient numériques ou catégorielles. Cet algorithme présente également une robustesse aux valeurs manquantes ainsi aucune imputation n'est requise. Il s'agit enfin d'un algorithme qui présente l'avantage d'être très flexible. Il peut considérer différentes fonctions de perte et prévoit de nombreuses options sur le réglage des hyperparamètres. La flexibilité possède également un coût en ce sens que lors de la phase de réglage des hyperparamètres, une grande et longue recherche par quadrillage est requise. La problématique de temps de calcul existe également au vu du nombre d'arbres que nécessite l'algorithme. La méthode du *gradient boosting* s'avère moins interprétable que le GLM ou que l'arbre CART même si l'importance des variables ainsi que les graphiques de dépendance partielle permettent d'obtenir une vision du modèle. Pour conclure, le gradient boosting est particulièrement adapté pour réduire le biais des modèles possédant une forte variance.

Les différents algorithmes qui vont ensuite être implémentés puis comparés au regard des certaines métriques définies dans la partie suivante ont ainsi été présentés. Il convient d'abord de compléter la

présentation de ces modèles par des éléments complémentaires et nécessaires avant le début de l'étude technique.

IV - Quelques notions complémentaires

1 - La sélection de modèles pour GLM

Un estimateur permettant de sélectionner un modèle parmi plusieurs modèles candidats est l'AIC (*Akaike Information Criterion*). Cet estimateur permet d'étudier l'ajustement et ainsi renseigne la qualité relative d'un modèle par rapport à chacun des autres modèles étudiés. L'AIC repose sur la théorie de l'information en ce sens que ce critère estime l'information perdue pour chaque modèle considéré. Le critère AIC doit être minimisé afin que le modèle perde le moins d'informations et renseigne ainsi une meilleure qualité.

L'AIC d'un modèle est défini par l'expression suivante :

$$AIC = 2k - 2\ln(L)$$

Avec :

- k, le nombre de paramètres du modèle
- L, le maximum de la fonction de vraisemblance du modèle

Le meilleur modèle parmi un ensemble de modèle au sens de l'AIC est celui qui minimise sa valeur. L'expression est composée de deux termes. Le terme $-2\ln(L)$ récompense par la fonction de vraisemblance la qualité d'ajustement tandis que le terme $2k$ pénalise l'ajout de nouveaux paramètres dans le modèle. Cette pénalisation permet de limiter le surajustement car l'ajout d'un nouveau paramètre se traduit généralement par une meilleure qualité d'ajustement. L'AIC représente ainsi un compromis entre le biais qui diminue avec le nombre de paramètres et la parcimonie qui est la volonté d'établir un modèle avec le moins de paramètres possible.

Dans le cas d'un échantillon de petite taille, l'AIC a tendance à sélectionner des modèles sur-ajustés. Pour pallier ce problème, il a été considéré un nouveau critère lorsque la base de données est relativement petite. Il s'agit de l'AICc et son expression est :

$$AICc = AIC + \frac{2k(k+1)}{N-k-1}$$

Avec :

- k, le nombre de paramètres du modèle
- N, le nombre d'individus dans l'échantillon

L'AICc correspond finalement à l'AIC avec un terme supplémentaire de pénalité. Il est à noter que lorsque n est grand, l'AICc est équivalent à l'AIC.

L'AIC et l'AICc n'apportent qu'une information de qualité d'ajustement relative aux autres modèles mais n'apportent pas d'information sur la qualité absolue. Il convient donc d'effectuer une étude des résidus pour obtenir cette information.

2 - Le sous et sur-apprentissage

Lors de la modélisation, deux types d'effets non désirés sont à vérifier. Il s'agit du sur-apprentissage ou bien du sous-apprentissage.

Le premier apparaît lorsque l'algorithme correspond de manière trop précise à l'échantillon d'apprentissage. C'est généralement le cas dans des modèle complexes lorsque la modélisation tend à apprendre une trop grande quantité d'information par rapport à la quantité de données disponibles. Le sur-apprentissage est trompeur puisque les critères d'ajustement aux données d'apprentissage paraissent satisfaisants. Cependant, cela entraîne une mauvaise qualité prédictive du fait de la dépendance aux données d'entraînement et aux petites fluctuations aléatoires non représentative du risque sous-jacent.

A l'inverse, le sous-apprentissage apparaît lorsque le modèle construit ne reflète pas assez la complexité des données. Le sous-apprentissage est également problématique puisqu'il biaise le résultat du fait d'un modèle trop simple.

Cette problématique peut être représentée graphiquement de la manière suivante :

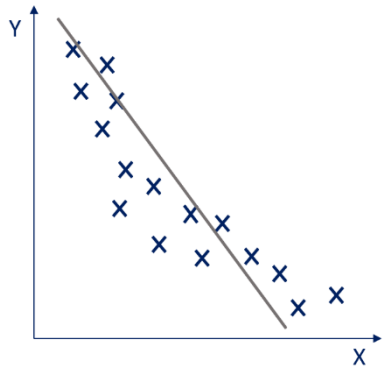


Figure 15 : Sous-apprentissage

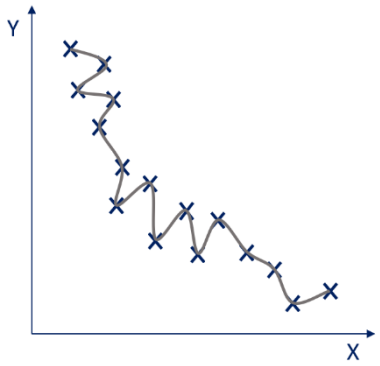


Figure 16 : Sur-apprentissage

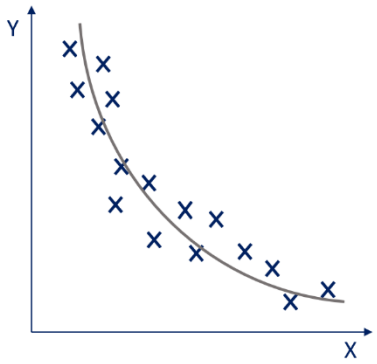


Figure 17 : Bon compromis

Il s'agit finalement de trouver un bon compromis entre ces deux situations. Cette situation est représentée dans le dernier graphique.

Finalement, le risque d'un modèle peut être décomposé comme la somme de deux termes :

- Le biais du modèle.
- La variance du modèle.

Ces deux types d'erreur varient en sens contraire avec la complexité du modèle. Il s'agit ainsi de minimiser la somme de ces deux types d'erreur pour trouver un bon compromis. D'une manière générale, les modélisations parcimonieuses sont à privilégier.

Le graphique ci-dessous permet de représenter le lien entre la dualité biais/variance et le sous et sur-apprentissage :

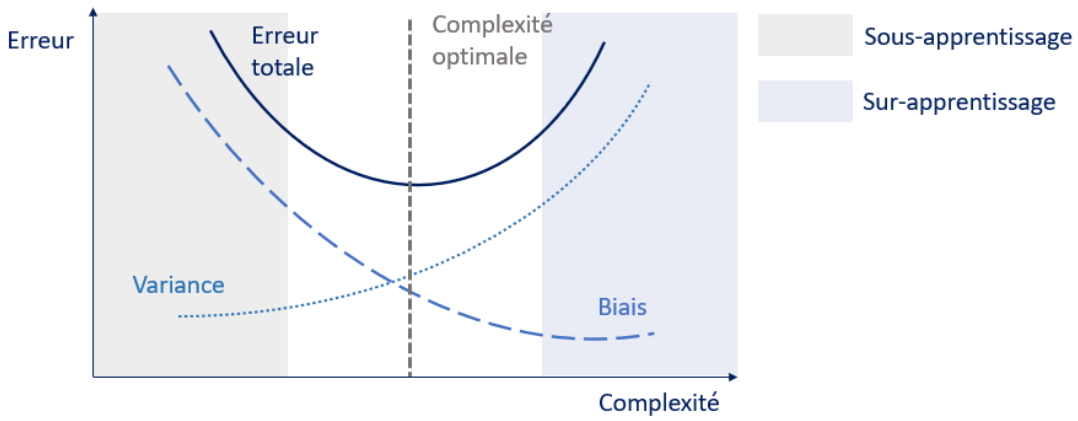


Figure 18 : Représentation de l'erreur totale en fonction de la complexité du modèle

Différentes techniques ont été développées pour éviter le sous et sur-apprentissage. Dans le cadre de la régression, il existe les régressions pénalisées ou bien la diminution du nombre de variables explicatives dans le modèle. En apprentissage, il s'agit de sélectionner le jeu de paramètres optimal à l'aide de la méthode de validation croisée présentée dans la partie suivante.

3 - La validation croisée

La validation croisée est notamment utilisée dans le cadre des techniques d'apprentissage supervisées. La technique de validation croisée a pour but de sélectionner un modèle optimal parmi différents modèles obtenus lors de la phase d'apprentissage. Cette méthode permet ainsi de détecter le jeu de paramètres le plus adapté aux données.

La validation croisée également nommée *k-folds cross-validation* consiste à partitionner en k sous échantillons distincts de la base d'apprentissage. $k-1$ sous échantillons sont utiles à l'entraînement et un sous échantillon est conservé pour évaluer le modèle.

Au vu de la taille de l'échantillon d'apprentissage, il sera considéré pour chacun des algorithmes une validation croisée de 5-folds. Ce processus peut être représenté de la manière suivante :

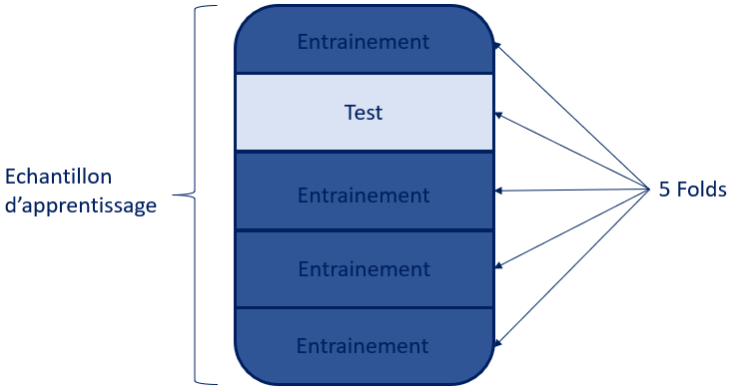


Figure 19 : Découpage de l'échantillon d'apprentissage lors de la validation croisée

Il s'agit ensuite de répéter ce processus 5 fois en modifiant à chaque répétition le *fold* test considéré.

Après avoir réalisé les 5-tests croisés, l'erreur finale de validation croisée est obtenue en effectuant la moyenne des erreurs sur chaque sous-échantillon. L'erreur finale est couramment appelée erreur de généralisation.

De cette manière, un test de validation croisée est effectué pour chaque jeu de paramètres considéré. Il est ensuite sélectionné le jeu de paramètres pour lequel l'erreur de généralisation obtenue est la plus faible. Ainsi, le modèle sera construit sur la base d'apprentissage totale en utilisant le jeu de paramètres optimal au sens de la validation croisée. La validation croisée s'avère alors nécessaire pour réduire la variance et améliorer la précision.

4 - Les hyperparamètres

En apprentissage supervisé, il convient de sélectionner un hyperparamètre, qui à la différence d'un paramètre, est externe au processus d'apprentissage mais en définit les propriétés. La détermination des

hyperparamètres intervient donc avant la phase d'apprentissage. Cette phase de réglage des hyperparamètres, aussi appelée *hyperparameter tuning* en anglais, est alors une étape influente en apprentissage supervisé.

Une méthode de recherche par quadrillage, appelée *Grid Search*, est traditionnellement utilisée. Il s'agit de déterminer une liste de possibilités pour chaque hyper-paramètre. Le score obtenu pour chaque combinaison d'hyper-paramètres est ensuite calculé à l'aide d'une métrique d'erreur qui sera introduite dans la partie suivante. Il est enfin sélectionné le vecteur d'hyperparamètres optimal par rapport à cette métrique d'erreur. Cette technique s'avère performante mais possède un inconvénient, à savoir son temps d'exécution car tester chaque combinaison possible d'hyper-paramètres est chronophage. Il convient alors de conserver une quantité limitée d'hyper-paramètres, de tester un nombre réduit de valeurs possibles pour chaque hyperparamètre et d'opter pour une approche par dichotomie afin de réduire ce temps d'exécution.

Le nombre d'hyperparamètres possible pour chaque algorithme est important. Cependant, comme suggéré précédemment, il convient d'en conserver un nombre réduit dans un souci de temps d'exécution. Pour chaque algorithme, les hyper-paramètres les plus influents ont été conservés. Ces derniers sont résumés dans le tableau ci-dessous :

Algorithme	Hyper-paramètres retenus
Arbre CART	max_depth
Forêt aléatoire	ntree mtry
Gradient boosting	n.trees interaction.depth shrinkage

Tableau 4 : Hyperparamètres considérés pour chaque algorithme

Concernant l'algorithme de forêt aléatoire, la phase de réglage des hyperparamètres consiste à déterminer le nombre d'arbres que compose la forêt, ainsi que le nombre de variables à utiliser à chaque division de nœud. Ces deux hyperparamètres seront respectivement appelés *ntrees* et *mtry*.

Concernant l'algorithme de gradient boosting, les hyperparamètres retenus sont au nombre de trois :

- *interaction.depth* : Il s'agit de la profondeur maximale autorisée pour chaque arbre.
- *n.trees* : Cet hyperparamètre correspond au nombre d'arbres considérés par l'algorithme.
- *shrinkage* : Le shrinkage est compris entre 0 et 1. Il correspond à la proportion de correction qui va être apportée à la prédiction en considérant un arbre supplémentaire. Une explication plus détaillée sera réalisée dans la partie de mise en application.

Cette étape d'étude et d'optimisation des hyperparamètres ajoutée au processus du retraitement des données de la base initiale ainsi qu'à celui des statistiques descriptives restent relativement simples à comprendre mais représentent le travail le plus fastidieux et chronophage de l'étude. En effet, cette phase demande le déploiement de nombreuses recherches et vérifications qui sont nécessaires et fondamentales afin d'obtenir de bons modèles de prédictions pour chaque algorithme implémenté.

5 - Les métriques d'erreur

Les métriques d'erreur ont pour objectif de déterminer l'efficacité d'un modèle en comparant les valeurs prédites aux valeurs réellement observées sur l'échantillon test. Il s'agit alors de définir des critères permettant de noter la qualité de prévision.

5.1 - Le critère RMSE

Ce critère est le critère traditionnellement utilisé dans le cadre d'une régression. Il s'agit de la racine carrée de l'erreur moyenne quadratique, d'où son acronyme RMSE (*Root Mean Squared Error*). L'erreur est définie comme l'écart entre la prévision du modèle effectuée sur l'échantillon test et les valeurs réellement observées sur ce même échantillon. Il s'agit ici d'obtenir le modèle minimisant le critère RMSE. Cette métrique s'écrit alors de la façon suivante :

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

Avec :

- y_i, la valeur observée pour l'individu i
- $\hat{y}_i$, la valeur prédite pour l'individu i
- N, la taille de l'échantillon

Le critère RMSE présente l'avantage d'être facilement interprétable et les valeurs obtenues sont de la même unité que celle de la variable à expliquer (par application de la racine carrée). Cette métrique pénalise de manière plus forte les grandes erreurs de prédiction par rapport aux petites, contrairement au critère MAE qui sera présenté dans la partie qui suit.

5.2 - L'Erreur moyenne absolue

La deuxième métrique considérée est la MAE (*Mean Absolute Error*). Tout comme le critère RMSE, la MAE permet de mesurer l'écart entre les prévisions et les observations sur l'échantillon test. L'expression de la MAE est la suivante :

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

Avec :

- y_i, la valeur observée pour l'individu i
- $\hat{y}_i$, la valeur prédite pour l'individu i
- N, la taille de l'échantillon

De la même manière que pour la métrique RMSE, la MAE permet une interprétation relativement facile et les valeurs obtenues sont de la même unité que celle de la variable à expliquer. Cette métrique confère le même poids aux grands et aux petits écarts de prédiction. La métrique MAE est toujours inférieure ou égale au critère RMSE.

5.3 - La moyenne obtenue sur les 10% des plus grands écarts de prédiction

La troisième métrique d'erreur est un critère d'évaluation qui se concentre uniquement sur le quantile 10% des plus grands écarts entre les valeurs observées et les valeurs prédites.

Afin de calculer cette métrique, il faut isoler le quantile 10% des prédictions qui présentent les plus grands écarts. Pour ce faire, le vecteur suivant est introduit :

$$w = (w_1, \dots, w_N)$$

Avec :

- $w_i = |y_i - \hat{y}_i|$
- y_i, la valeur observée pour l'individu i
- $\hat{y}_i$, la valeur prédite pour l'individu i
- N, la taille de l'échantillon

Il convient ensuite d'ordonner les composantes w_i en ordre décroissant dans un nouveau vecteur z :

$$z = (z_1, \dots, z_i, \dots, z_N) = (w_{(N)}, \dots, w_{(N-i+1)}, \dots, w_{(1)})$$

Avec :

- $w_{(i)}$, le i^{ème} plus petit écart de prédiction
- $z_i = w_{(N-i+1)}$

De cette manière, la moyenne du quantile 10% des plus grands écarts de prédiction est calculée :

$$Erreur_{10\%} = \frac{1}{K}\sum_{i=1}^{K} z_i$$

Avec :

- $K = \left\lfloor \frac{N}{10} \right\rfloor$, la taille de l'échantillon du quantile 10%.

Cette méthode s'avère finalement correspondre à la méthode MAE appliquée au quantile 10% des « pires » prédictions. Cette métrique présente l'avantage d'obtenir une certaine vision des plus grands écarts de prédictions et vient en ce sens compléter les deux premières métriques utilisées.

6 - Le modèle individuel et le modèle collectif

Le coût global des sinistres des assurés pour une année de survenance peut s'exprimer de deux manières :

- La somme, sur le nombre d'assurés, du coût total engendré par chaque assuré.

- La somme, sur le nombre de sinistres, des coûts de chaque sinistre.

La première approche est appelée modèle individuel tandis que la seconde correspond au modèle collectif.

6.1 - Le modèle individuel

Soient :

- $Y_{i,j}$; la variable aléatoire qui décrit le montant du $j^{ème}$ sinistre de l'assuré i.
- X_i ; la variable aléatoire représentant le coût annuel de l'assuré i.
- N_i ; le nombre de sinistres pour une année de l'assuré i.
- X ; le montant total des sinistres de tous les assurés du portefeuille sur une année, appelé charge globale.
- n ; le nombre d'assurés dans le portefeuille.

Le coût des arrêts de travail de l'assuré i sur une année s'écrit :

$$X_i = \sum_{j=1}^{N_i} Y_{i,j}$$

Le coût des sinistres de tous les assurés sur cette même année s'exprime :

$$X^{ind} = \sum_{i=1}^{n} \sum_{j=1}^{N_i} Y_{i,j} = \sum_{i=1}^{n} X_i$$

6.2 - Le modèle collectif

Il convient d'introduire la variable aléatoire N du nombre annuel de sinistres. Cette variable permet d'écrire :

$$X^{coll} = \sum_{i=1}^{N} X_i$$

Dans l'hypothèse où les X_i sont indépendantes, identiquement distribuées et indépendantes de N, les deux premiers moments du coût global annuel des sinistres des assurés s'écrivent de la manière suivante :

$$\mathbb{E}\left(X^{coll}\right) = \mathbb{E}(N)\mathbb{E}(X_1)$$

Et

$$\mathbb{V}\left(X^{coll}\right) = \mathbb{E}(N)\mathbb{V}(X_1) + \mathbb{V}(N)\mathbb{E}(X_1)^2$$

Cette décomposition de l'espérance et de la variance de la charge globale X^{coll} justifie l'introduction de deux modèles :

- Un modèle de fréquence
- Un modèle de coût moyen

Les éléments nécessaires à la modélisation des différents algorithmes ont été introduits. Dans la partie suivante, les méthodes expliquées précédemment vont être appliquées avec le jeu de données étudié dans ce mémoire.

V - La construction du modèle de fréquence des arrêts de travail

1 - La préparation de la base est essentielle

Avant d'appliquer les différents algorithmes, il est nécessaire de procéder à une phase de retraitement des données afin d'obtenir une base de bonne qualité. Cette étape constitue un point primordial pour obtenir un modèle performant. Peu importe la complexité de l'algorithme utilisé, si les données ne sont pas de bonne qualité alors les résultats finaux ne pourront pas être satisfaisants. En effet, ce sont les algorithmes qui s'adaptent aux données et non l'inverse.

1.1 - La qualité et le retraitement de la base de données

Une base de données est jugée de bonne qualité lorsque ces 3 critères peuvent être vérifiés :

- Pertinence
- Exhaustivité
- Exactitude

La base de données étudiée ne soulève pas de problème sur les critères de pertinence ou d'exactitude des données. Néanmoins, quelques problématiques concernant l'exhaustivité des données ont dû être résolues dans la phase de préparation des données. Il convient donc d'expliciter les points pour lesquels les données ont présenté des difficultés particulières ou ont nécessité des retraitements.

Comme indiqué dans la partie concernant les statistiques descriptives, le portefeuille est composé de deux produits distincts de type Madelin afin de gagner en quantité de données. La contrepartie de cette augmentation du nombre de ligne dans le portefeuille est qu'il est de cette manière introduit un biais. Il sera explicité par la suite la méthode qui a permis de limiter ce biais.

Comme expliqué dans une partie précédente, les TNS ont des profils très différents présentant notamment de fortes disparités au niveau des revenus et des secteurs d'activités. Ces différentes caractéristiques ont présenté un enjeu dans la modélisation expliqué par la suite.

Tout d'abord, il est important de noter que la base de données ne comprend que des TNS classiques et ne présente donc pas de micro-entrepreneur. Ensuite, la base de données contient les informations sur les grandes catégories professionnelles mais la granularité n'est pas assez fine pour pouvoir construire un modèle de prédiction précis. Ainsi, pour obtenir une meilleure précision et donc une plus petite granularité, il a été nécessaire de retraiter la variable des libellés des métiers. Cette variable est intéressante car elle présente la plus faible granularité sur la catégorie socio-professionnelle et donne également l'information sur le secteur d'activité du TNS. Ensuite, s'est posé la question du regroupement des attributs de la variable des libellés profession dans différents sous-groupes de risques homogènes. Une variable nommée *codeCSP* a donc été créée, elle regroupe à la fois le secteur d'activité et la catégorie socio-professionnelle.

La segmentation qui a donc été retenue permet un compromis entre le statut, le régime obligatoire ainsi que la catégorie professionnelle du TNS. Cette nouvelle segmentation est présentée ci-dessous :

	Code CSP	Catégorie professionnelle	Régime obligatoire	Libellé regoupement métier
Gérants majoritaires Artisans / commerçants	GM_1	Gérant Majoritaire	CNAV	Métier de l'hôtellerie et restauration
	GM_2	Gérant Majoritaire	CNAV	Métier de bouche
	GM_3	Gérant Majoritaire	CNAV	Métier du commerce
	GM_4	Gérant Majoritaire	CNAV	Métier de l'artisanat
	GM_5	Gérant Majoritaire	CNAV	Métier de la beauté et bien être
	GM_6	Gérant Majoritaire	CNAV	Métier du transport et de l'automobile
	GM_7	Gérant Majoritaire	CNAV	Métier du bâtiment/Gros œuvre
	GM_8	Gérant Majoritaire	MSA	Métier de la culture, élevage, pêche et travaux forestiers

	Code CSP	Catégorie professionnelle	Régime obligatoire	Libellé regoupement métiers
Entrepreneurs Individiuels Artisans / commerçants	EI_1	Entrepreneur Individuel	CNAV	Métier de l'hôtellerie et restauration
	EI_2	Entrepreneur Individuel	CNAV	Métier de bouche
	EI_3	Entrepreneur Individuel	CNAV	Métier du commerce
	EI_4	Entrepreneur Individuel	CNAV	Métier de l'artisanat
	EI_5	Entrepreneur Individuel	CNAV	Métier de la beauté et bien être
	EI_6	Entrepreneur Individuel	CNAV	Métier du transport et de l'automobile
	EI_7	Entrepreneur Individuel	CNAV	Métier du bâtiment/Gros œuvre
	EI_8	Entrepreneur Individuel	MSA	Métier de la culture, élevage, pêche et travaux forestiers

	Code CSP	Catégorie professionnelle	Régime obligatoire	Libellé regoupement métier
Professions libérales réglementées	PL_R_1	PL réglementées	CARMF	Médecin
	PL_R_2	PL réglementées	CARCDSF1	Dentiste et Stomatologue
	PL_R_3	PL réglementées	CARCDSF2	Sage Femme
	PL_R_4	PL réglementées	CARPV	Vétérinaires
	PL_R_5	PL réglementées	CAVP	Pharmacien
	PL_R_6	PL réglementées	CAVAMAC	Agents généraux d'assurances
	PL_R_7	PL réglementées	CAVEC	Expert comptable
	PL_R_8	PL réglementées	CAVOM	Officiers publiques et ministériels
	PL_R_9	PL réglementées	CNBF&LPA	Avocats
	PL_R_10	PL réglementées	CRN	Notaires
	PL_R_11	PL réglementées	Maison des artistes&AGESSA	uniquement les architectes, infographistes, graphistes et photographes

	Code CSP	Catégorie professionnelle	Régime obligatoire	Libellé regoupement métiers
Professions libérales réglementées paramédicales	PL_RP_1	PL réglementées paramédicales	CARPIMKO	Paramédicaux rattachés à la CARPIMKO
	PL_RP_2	PL réglementées paramédicales	CARPIMKO	Infirmiers

	Code CSP	Catégorie professionnelle	Régime obligatoire	Libellé regoupement métiers
Professions libérales NON réglementées	PL_NR_1	PL non réglementées	CIPAV ou CNAV	Métier de l'assurance, de l'immobilier et du patrimoine, commercial
	PL_NR_2	PL non réglementées	CIPAV ou CNAV	Métier paramédical rattaché à la CIPAV ou CNAV
	PL_NR_3	PL non réglementées	CIPAV ou CNAV	Autres Professions libérales 1
	PL_NR_4	PL non réglementées	CIPAV ou CNAV	Autres Professions libérales 2

Tableau 5 : Tableau récapitulatif de la segmentation de la nouvelle variable codeCSP

Les différents libellés métiers ont été regroupés dans l'une des 33 modalités de la variable *codeCSP*. Il est intéressant de noter que la distinction entre les gérants majoritaires et les entrepreneurs individuels a été conservé afin de prendre en compte cette distinction dans la modélisation. Cette distinction permet ainsi de réduire en partie le biais intégré par le regroupement des deux produits de prévoyance en une seule base.

La variable *codeCSP* comprend ainsi cinq grandes catégories professionnelles :

- Les gérants majoritaires : 8 modalités
- Les entrepreneurs individuels : 8 modalités
- Les professions libérales réglementées : 11 modalités
- Les professions libérales réglementées paramédicales : 2 modalités
- Les professions libérales non réglementées : 4 modalités

Cette variable permet ainsi de prendre en compte de manière précise les différentes hétérogénéités existantes au sein des travailleurs non salariés avec dans chaque modalité un nombre suffisant d'individus réduisant ainsi le biais dû au manque d'exposition.

1.2 - L'obtention de la base des sinistres attritionnels

Après avoir retraité les données, il est nécessaire de créer une base des sinistres attritionnels afin de les séparer des sinistres extrêmes. Les sinistres extrêmes sont à écarter lors de la modélisation car ils peuvent conduire à représenter une part trop importante dans la charge de sinistralité sur une catégorie spécifique. Le quantile 1% des lignes sinistrées a donc été écarté et sera réintégré a posteriori de la modélisation sur l'ensemble de la population.

Une faible proportion des lignes sinistrées est observée. Cet ordre de grandeur semble néanmoins cohérent avec les deux éléments suivants :

- Le risque arrêt de travail est un risque restant relativement peu fréquent.
- La fréquence des sinistres des contrats possédant une grande franchise est sous-estimée du fait de la troncature à gauche. Les sinistres possédant des durées d'arrêt inférieures à la franchise ne sont pas observés puisqu'il n'existe pas d'intérêt à les déclarer pour l'adhérent.

1.3 - Le découpage de la base des sinistres attritionnels

La base des sinistres attritionnels est séparée en deux échantillons, à savoir un échantillon d'apprentissage et un échantillon test :

- La base d'apprentissage correspond à l'échantillon qui sert à construire et ajuster le modèle. Dans le cas du GLM, cet échantillon permet d'estimer les coefficients du modèle. Pour les méthodes d'apprentissage supervisées, il s'agit de la base sur laquelle les algorithmes ajustent leur modèle.
- La base test est, quant à elle, utilisée afin d'évaluer la qualité de prédiction du modèle. Cette base n'ayant pas été utilisée pour l'apprentissage du modèle, est indépendante du modèle construit. Cela permet de simuler l'obtention de nouvelles données, avec pour objectif, la prédiction de la variable à expliquer à partir du modèle ajusté sur la base d'apprentissage. Ces prédictions sont ensuite comparées aux valeurs réellement observées de la variable à expliquer sur la base test.

L'échantillon d'apprentissage représente ici 70% de la base initiale et l'échantillon test représente donc 30% de la base de données. Une attention particulière a été portée sur le fait d'effectuer ces affections sur l'une ou l'autre base de manière aléatoire afin de n'introduire aucun biais.

Les fréquences résultantes dans chacun des deux échantillons sont proches. Le biais induit par cette séparation de la base initiale dans la construction du modèle reste donc limité. Cette répartition est alors satisfaisante.

1.4 - La procédure de modélisation

Après avoir réalisé le découpage de la base de données, trois étapes décrites ci-dessous sont nécessaires pour la modélisation :

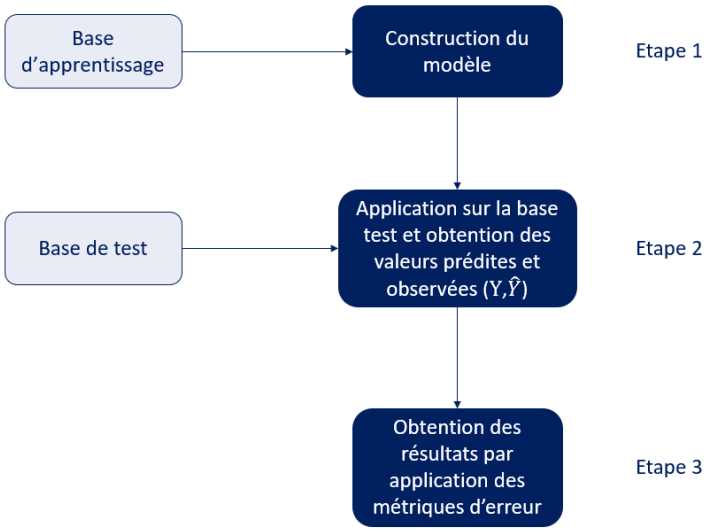


Figure 20 : Procédure de modélisation appliquée

Dans le cadre de l'apprentissage supervisé, un échantillon supplémentaire doit être considéré, à savoir l'échantillon de validation. Cet échantillon permet d'optimiser les paramètres de la méthode considérée. En effet, lors de la mise en place des méthodes d'apprentissage supervisées, il est nécessaire de tester différents jeux de paramètres afin de choisir celui qui est optimal. Cet échantillon permet ainsi de comparer les modèles issus des différents jeux de paramètres. Dans cette étude, la base de données n'est pas de taille suffisante pour extraire un troisième échantillon de la base initiale. Il convient alors de considérer l'échantillon de validation comme un sous échantillon de la base d'apprentissage. Il s'agit de la méthode de validation croisée.

Avec la création de la variable explicative $codeCSP$, les variables considérées dans les différents algorithmes sont les suivantes :

- $classeAge$
- $codeCSP$
- $équipementSanté$
- $Franchise$
- $Genre$
- $zoneGéographique$

Dans chacun des algorithmes, la variable à expliquer peut être la variable $nombreSinistres$ si l'algorithme présente la possibilité de mettre en offset la variable $duréeExposition$. Dans le cas contraire, la variable à expliquer considérée est la variable $Fréquence = \frac{nombreSinistres}{duréeExposition}$. Il s'agit désormais de mettre en place chacun de ces algorithmes.

2 - Première méthode : le GLM

Dans un premier temps, la méthode traditionnelle est appliquée, il s'agit du modèle linéaire généralisé. La première étape est celle qui permet d'évaluer la loi de distribution des sinistres. Traditionnellement, pour représenter la fréquence, il est considéré la loi de Poisson ou la loi binomiale négative en cas de surdispersion. La fonction lien ln est adaptée pour ces deux lois. Ainsi, l'expression du modèle est la suivante :

$$\ln(nombresSinistres) = \beta_0 + \beta_1 classeAge + \beta_2 codeCSP + \beta_3 équipementSanté + \beta_4 Franchise + \beta_5 Genre + \beta_6 zoneGéographique + offset(\ln(duréeExposition)) + \varepsilon$$

La prédiction moyenne est ensuite obtenue en appliquant la fonction inverse du logarithme népérien c'est-à-dire la fonction exponentielle.

Le choix de la loi de distribution est obtenu selon l'étude de la surdispersion ou sousdispersion. En effet, la loi de Poisson suppose que la variance est égale à l'espérance. Ainsi, si les données sont surdispersées alors la loi binomiale négative est préférée puisqu'elle autorise l'utilisation d'un paramètre de dispersion différent de 1. Le ratio entre la déviance résiduelle du modèle et son degré de liberté permet de calculer la dispersion. La loi de Poisson présente un ratio égal à 0,25. Il n'apparaît donc pas de surdispersion cependant le ratio s'avère éloigné de 1. Afin d'étudier quelle loi de distribution permet d'obtenir un meilleur modèle, les critères AIC des deux modèles obtenus sont comparés en considérant les mêmes variables explicatives.

Après avoir construit les modèles, les AIC sont obtenus avec la commande *summary()* :

Loi de distribution	AIC
Poisson	35011
Binomiale négative	34812

Tableau 6 : Comparaison des AIC

La loi binomiale négative est donc mieux adaptée aux données que la loi de Poisson au sens du critère d'AIC. Pour conforter cette approche, un test graphique de conformité est réalisé afin de vérifier l'adéquation du modèle aux données des deux lois de distribution.

Cette comparaison est effectuée avec la commande *goodfit* disponible sur R et est représentée ci-dessous :

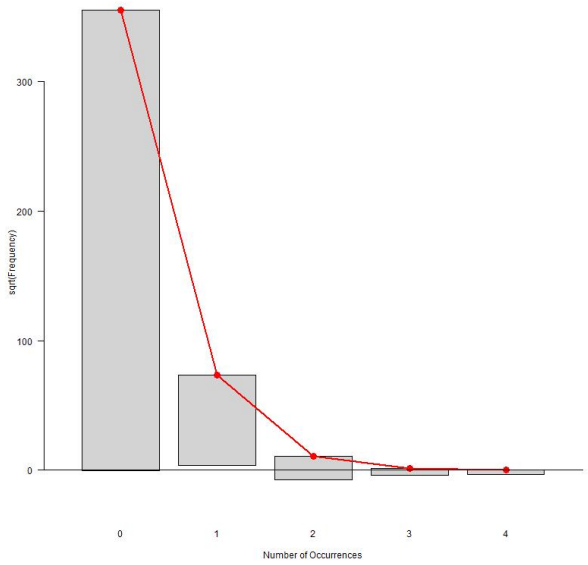


Figure 21 : Adéquation du nombre de sinistres observés et théoriques d'une distribution de loi de Poisson

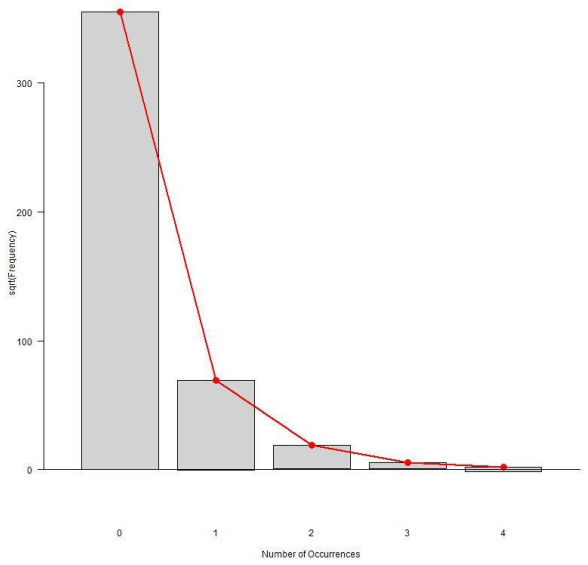


Figure 22 : Adéquation du nombre de sinistres observés et théoriques d'une distribution de loi binomiale négative

Ce graphique permet de conforter visuellement la meilleure adéquation des données à la loi binomiale négative en comparaison à la loi de Poisson. Il est ainsi retenu la loi binomiale négative comme loi de distribution par la suite.

2.1 - L'adéquation du modèle aux données

Après avoir réalisé le GLM, la bonne qualité d'ajustement du modèle doit être vérifiée.

Concernant celle-ci, les méthodes permettant de mesurer l'adéquation du modèle aux données sont explicitées en annexe, à savoir :

- La méthode de la déviance
- Le test de Pearson

Les p-value obtenues pour ces deux tests sont proches de 1 et donc supérieures à 5 %. Ainsi, l'hypothèse nulle ne peut pas être rejetée au risque de 5%.

2.2 - Les tests de significativité

Un premier test de significativité est le test de significativité globale du modèle. Il est effectué via le test du rapport de vraisemblance. L'hypothèse nulle *H0* est :

$$\beta_1 = \beta_2 = \cdots = \beta_p = 0$$

La p-value obtenue est proche de 0. Le modèle est donc globalement significatif.

Le second test est celui de la significativité individuelle des coefficients. Il s'agit d'effectuer le test de Wald sur chacun des coefficients. La commande *summary()* prenant en argument le GLM permet

d’observer les différentes p-value de ce test. Les probabilités critiques étant pour chaque coefficient inférieures à 5%, il en est déduit que chaque coefficient est significativement différent de zéro au risque 5%.

Les tests de significativité ainsi obtenus permettent de valider la cohérence du modèle ainsi que le regroupement effectué des modalités de chaque variable.

2.3 - L’analyse des résidus

Après avoir effectué les tests d’adéquation et de significativité, l’étape qui suit est celle de l’analyse des résidus. Il est représenté ci-dessous le diagramme Quantile-Quantile des résidus de déviance :

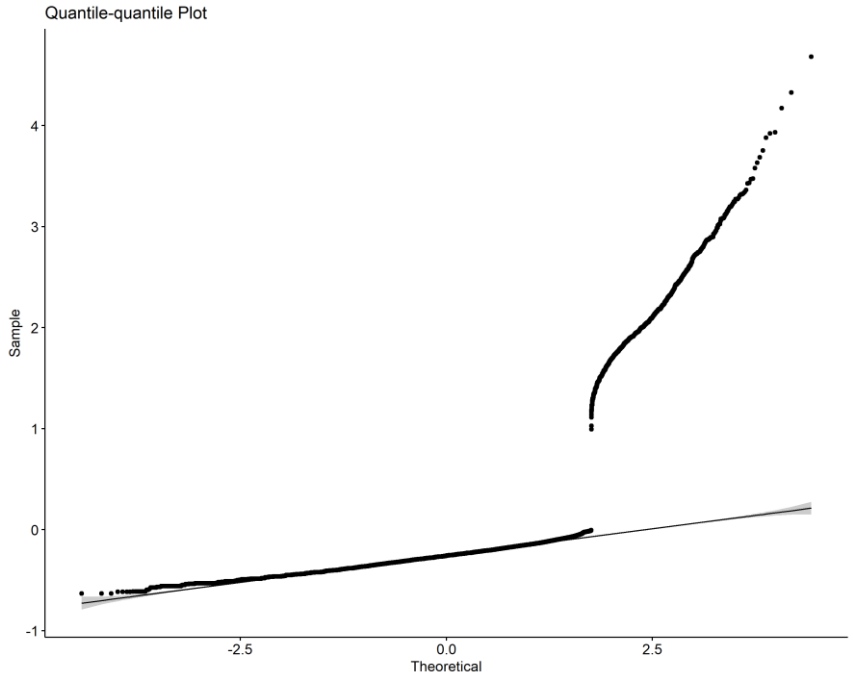


Figure 23 : Q-Q plot des résidus de déviance

Le diagramme Quantile-Quantile, aussi appelé Q-Q plot, permet d’évaluer graphiquement la pertinence de l’ajustement de la distribution des résidus de déviance à une loi normale. Le Q-Q Plot présente un mauvais alignement des points principalement sur les données sinistrées. Cela s’explique principalement par le fait que les données sinistrées sont relativement peu fréquentes dans le portefeuille. Elles en deviennent donc pour le GLM plus difficiles à prédire. Cette étude des résidus ne permet donc pas de valider la distribution normale des résidus de déviance.

2.4 - La performance du GLM

En appliquant le modèle construit sur l’échantillon d’apprentissage à l’échantillon test, il est obtenu les valeurs prédites. Ainsi, les différentes métriques d’erreur du GLM sont les suivantes :

Métrique d'erreur	RMSE	MAE	Erreur 10%
GLM	0,224	0,078	0,525

Tableau 7 : Résultats des métriques d'erreur du modèle obtenu par GLM

Ces métriques d'erreur sont à comparer avec celles obtenues avec les algorithmes d'apprentissage supervisé.

La fréquence prédite avec la méthode GLM est inférieure de 3,06% à la fréquence réelle de l'échantillon test. Cette observation peut être expliquée par l'étude des résidus qui montre comment les prédictions des données sinistrées sont sous-estimées. Il semble donc que le GLM sous évalue la fréquence, ce qui s'avère problématique dans le cadre d'une tarification. Il s'agit ainsi d'effectuer une approche par apprentissage supervisé afin de tenter d'obtenir une fréquence prédite plus proche de la fréquence réelle et de meilleures métriques d'erreur.

3 - Deuxième méthode : L'arbre CART

La première méthode d'apprentissage qui est modélisée est l'arbre CART. Dans le modèle GLM, il est directement modélisé la fréquence avec le nombre d'arrêts de travail dans l'année considérée en prenant en compte la variable d'exposition au risque en offset. L'algorithme CART n'offre pas la possibilité de considérer la variable *duréeExposition* en offset néanmoins il est nécessaire de prendre en compte cette variable. Il convient donc d'intégrer la durée d'exposition dans la définition de la fréquence de la manière suivante :

$$Fréquence = \frac{Nombre\ d'arrêts}{Durée\ d'exposition}$$

Par souci de comparaison, il est utilisé les mêmes échantillons d'apprentissage et de test que ceux ayant servi pour le GLM. L'arbre CART est construit avec la commande *rpart* sur le logiciel de programmation R. Dans un premier temps, il est nécessaire de construire l'arbre saturé. Il ne peut être représenté du fait de sa trop grande dimension. L'arbre CART optimal doit ensuite être construit. Pour cela, il convient de sélectionner la complexité optimale qui minimise l'erreur de validation croisée. Le graphique ci-dessous représente l'erreur de validation croisée en fonction de la complexité :

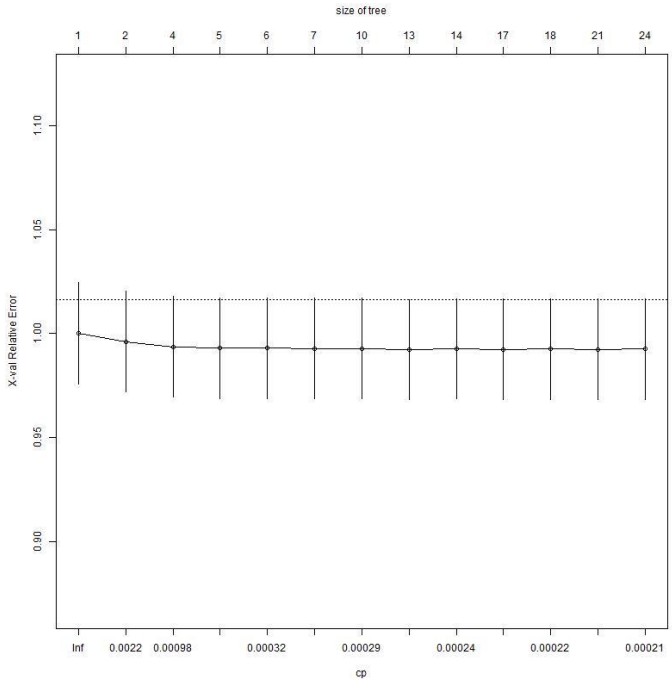


Figure 24 : Représentation de l'erreur de validation croisée en fonction de la complexité

L'erreur de validation croisée ne varie et ne diminue que très peu en fonction de la complexité en restant proche de 1. Cela traduit une relative mauvaise performance de cet algorithme. Il est retenu le paramètre cp qui minimise cette erreur de validation croisée, c'est-à-dire, $cp = 0{,}00025$. L'arbre CART élagué est obtenu en construisant l'arbre CART avec ce paramètre cp optimal.

Après avoir obtenu l'arbre optimal, les prédictions sur l'échantillon test puis l'évaluation des métriques d'erreur sont effectuées. De cette manière il est calculé :

Métrique d'erreur	RMSE	MAE	Erreur 10%
GLM	**0,224**	**0,078**	**0,525**
Arbre CART	0,280	0,098	0,569

Tableau 8 : Résultats des métriques d'erreur du modèle obtenu par arbre CART

Au regard de chacun des métriques d'erreur, l'arbre CART est un modèle moins performant que le GLM.

La fréquence obtenue par prédiction sur l'échantillon test est légèrement supérieure de 2,95% à celle réellement observée. La méthode CART surévalue donc la fréquence mais présente un plus faible biais que l'algorithme GLM. Néanmoins l'arbre CART présente une variance élevée, cet effet n'est pas recherché et doit au contraire être diminué. L'arbre CART est traditionnellement utilisé dans les méthodes ensemblistes afin de construire des classifieurs robustes et stables.

Deux méthodes d'ensembles permettent de réduire la variance du modèle, à savoir les méthodes d'agrégation et de boosting.

4 - Troisième méthode : La forêt aléatoire

La forêt aléatoire est une méthode d'agrégation de modèle qui s'appuie sur la variance inhérente à l'algorithme CART afin de construire un modèle plus robuste. Tout d'abord, il est nécessaire d'effectuer un réglage des hyperparamètres. Les hyperparamètres retenus dans le cadre de l'algorithme *random forest* sont les suivants :

- *ntree*, il s'agit du nombre d'arbres que composent la forêt.
- *mtry*, il s'agit du nombre de variables m testées à chaque nœud parmi les 6 variables explicatives. Plus *mtry* est choisi petit, plus le choix des variables s'approche du hasard, ce qui se traduit par une faible variance mais un fort biais. L'effet inverse est observé lorsque *mtry* devient grand.

Cette étape est effectuée par une recherche par quadrillage dont la méthode est nommée *grid search*. Les valeurs suivantes ont été testées :

- *ntree* : de 100 à 500 avec un pas de 100
- *mtry* : de 1 à 4

Le calcul de la validation croisée a été effectué sur un 5-*folds*. La métrique d'erreur utilisée pour cette étape est le RMSE. Il est ci-dessous exprimé les résultats de l'étape de *tuning* effectuée à l'aide des fonctions *train()* et *expand.grid()* de R.

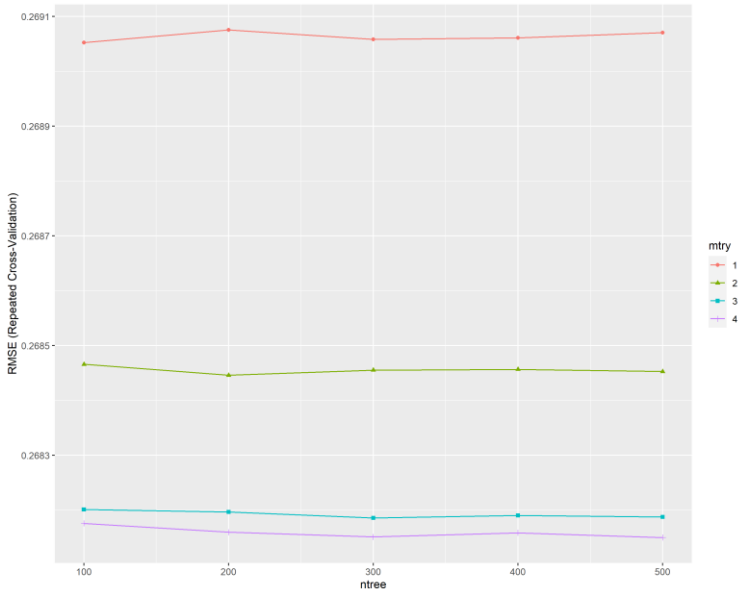


Figure 25 : Résultats de la recherche par quadrillage des hyperparamètres de la forêt aléatoire

Pour une même valeur de mtry, le nombre d'arbres n'impacte que très peu les RMSE de validation croisée. En revanche, la variable mtry possède un fort impact sur le résultat. Finalement le meilleur modèle au regard du critère RMSE est obtenu pour une forêt composée de 500 arbres et effectuant une randomisation de 4 variables à chaque nœud. Le choix de mtry relativement grand semble signifier ici qu'un modèle possédant un plus faible biais et une plus forte variance permettra de meilleures prédictions.

Après avoir estimé les hyperparamètres optimaux, il convient ensuite de calculer cette forêt à l'aide de la fonction *randomForest* implémentée sur R. L'interprétation des résultats n'est pas immédiate comme c'est le cas pour l'algorithme CART. En effet, pour l'algorithme des forêts aléatoires la randomisation des arbres empêche cette interprétation. La notion d'importance des variables a été introduite afin de pallier cette perte d'information. Le graphique ci-dessous représente quelle variable possède le plus d'influence dans la construction du modèle :

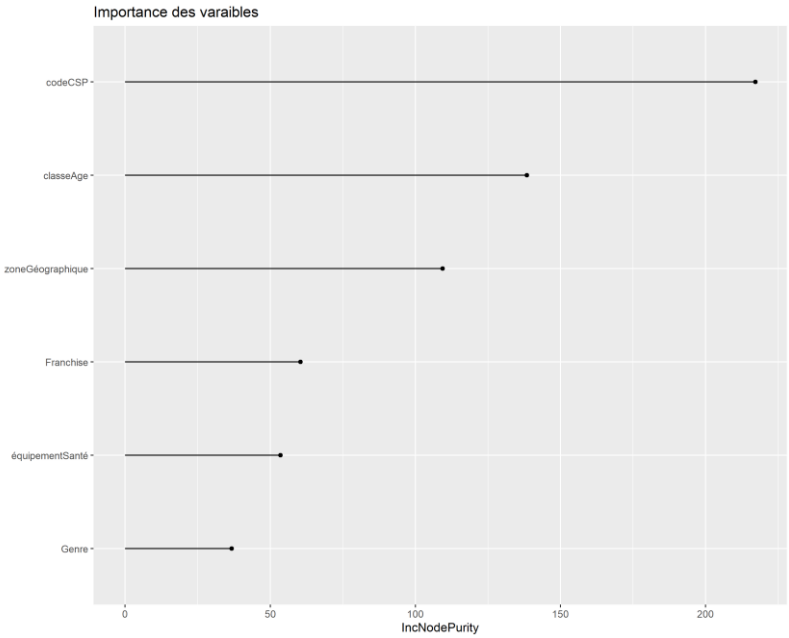


Figure 26 : Importance des variables de la forêt aléatoire

La mesure utilisée pour évaluer l'importance des variables est notée *IncNodePurity*. Il s'agit de l'augmentation de la pureté dans les nœuds. Cette mesure est analogue à la décroissance de l'impureté de Gini dans le cadre de la classification et s'appuie sur la réduction de la somme des erreurs au carré lorsque la variable est sélectionnée. Il est à noter que la variable la plus importante au regard de l'*IncNodePurity* est la variable *codeCSP*. Cela suggère que la catégorie professionnelle ainsi que le statut du TNS possèdent une forte influence dans la sinistralité. A l'inverse, la variable la moins importante est la variable *Genre*. Le graphique ci-dessous présente l'impact du nombre d'arbres sur la stabilisation de la MSE :

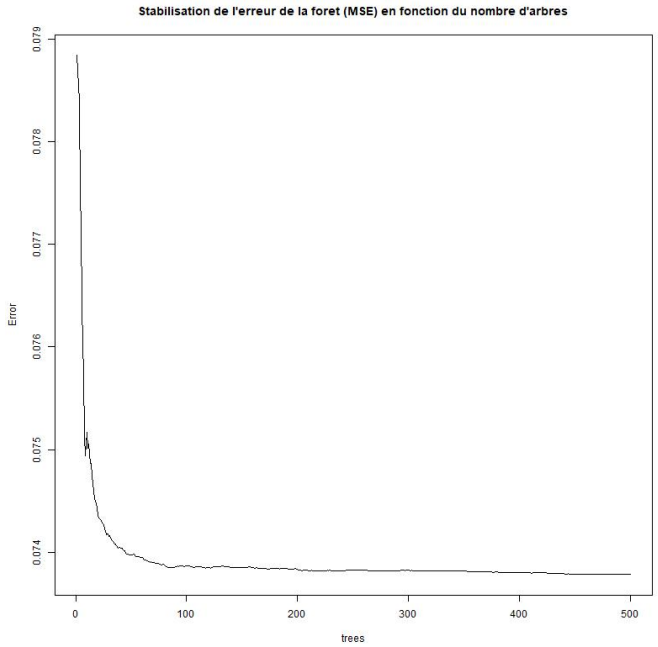


Figure 27 : Stabilisation de la MSE en fonction de ntree

Le graphique concorde avec le résultat de la phase de réglage des hyperparamètres qui suggère qu'entre 100 et 500 arbres la réduction de la MSE (et donc de la RMSE) est relativement faible.

Les prédictions sont ensuite effectués sur l'échantillon test. La répartition des fréquences prédites est représentée ci-dessous :

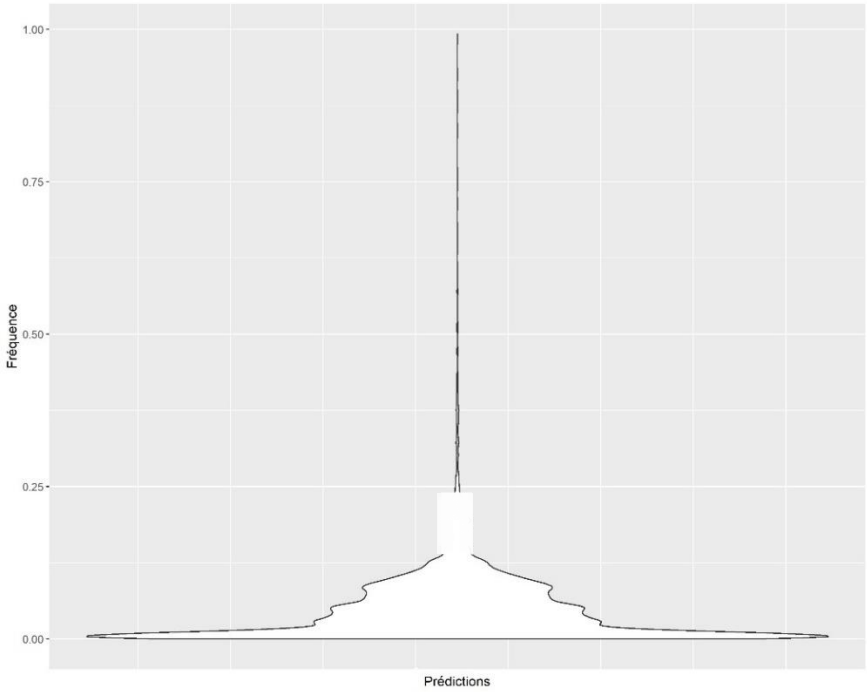


Figure 28 : Répartition des prédictions de la fréquence

Ce graphique montre une très forte concentration des petites fréquences prédites et une grande étendue des fréquences élevées avec un maximum à 0,987. La moyenne résultante des prédictions est proche de celle réellement observée sur l'échantillon test. La fréquence moyenne prédite est 1,8% plus grande que celle de l'échantillon test. Le biais du modèle par forêt aléatoire paraît ainsi réduit.

Les différentes métriques d'erreur obtenues sont résumées dans le tableau suivant :

Métrique d'erreur	RMSE	MAE	Erreur 10%
GLM	**0,224**	**0,078**	**0,525**
Arbre CART	0,280	0,098	0,569
Forêt aléatoire	0,268	0,092	0,557

Figure 29 : Résultats des métriques d'erreur du modèle obtenu par random forest

Le critère RMSE est légèrement amélioré par rapport à celui de l'arbre CART mais il reste plus élevé que celui du GLM.

5 - Quatrième méthode : Le *stochastic gradient boosting*

Comparons cette méthode à une autre méthode d'ensemble, le gradient boosting. Cette méthode est utilisée pour la modélisation de la fonction *gbm* de R. Cette fonction présente l'avantage de pouvoir intégrer une variable en offset. Ainsi, comme pour le GLM, la variable à expliquer considérée est le *nombreSinistres* et il est intégré la variable *duréeExposition* en offset.

Tout d'abord, il est nécessaire d'effectuer un réglage des hyperparamètres. Les hyperparamètres retenus pour l'algorithme *stochastic gradient boosting* sont :

- *n.trees*, il s'agit du nombre d'arbres qui va être calculé.
- *interaction.depth*, il s'agit du nombre maximal de nœuds par arbre.
- *shrinkage*, il s'agit du taux d'apprentissage.

De la même manière que pour la forêt aléatoire, cette étape est effectuée par une recherche par quadrillage, *grid search*. Après avoir effectué une première approche par dichotomie afin de réduire les possibilités, les valeurs suivantes ont été testées :

- *n.trees* : 250 ; 500 ; 1000
- *interaction.depth* : 3 ; 4
- *shrinkage* : 0,01 ; 0,1

Il va ainsi être testé les 12 combinaisons suivantes :

Hyperparamètres	n.trees	interaction.depth	shrinkage
Combinaison 1	250	3	0.01
Combinaison 2	250	3	0.1
Combinaison 3	250	4	0.01
Combinaison 4	250	4	0.1
Combinaison 5	500	3	0.01
Combinaison 6	500	3	0.1
Combinaison 7	500	4	0.01
Combinaison 8	500	4	0.1
Combinaison 9	1000	3	0.01
Combinaison 10	1000	3	0.1
Combinaison 11	1000	4	0.01
Combinaison 12	1000	4	0.1

Figure 30 : Représentation des différentes combinaisons d'hyperparamètres testés

Le calcul de la validation croisée a été effectué sur un 5-*folds*. La métrique d'erreur utilisée pour cette étape est le RMSE. Les résultats de l'étape de *tuning* effectuée à l'aide des fonctions *train()* et *expand.grid()* de R sont exprimés dans les graphiques ci-dessous :

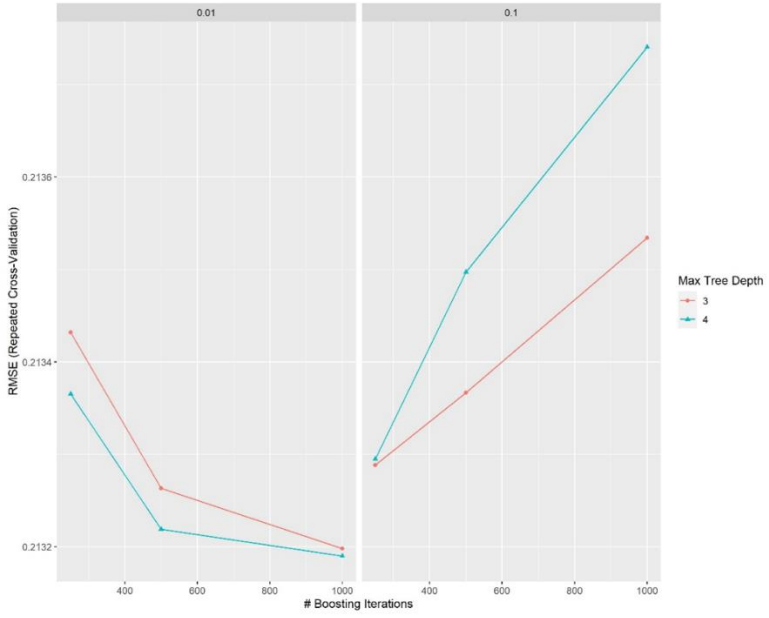


Figure 31 : Résultat de la recherche par quadrillage des hyperparamètres retenus

Le graphique de gauche représente les courbes d'évolution de la RMSE en fonction du nombre d'arbres pour un *learning rate* de 0,01. Celui de droite représente les mêmes informations pour un *learning rate* de 0,1. Le graphique du taux d'apprentissage de 0,1 montre ici que le RMSE augmente avec le nombre d'itérations. Cela est expliqué par le fait que le learning rate considéré est trop grand. Le minimum de la fonction de perte n'est donc pas trouvé.

La meilleure combinaison d'hyperparamètres au sens de la validation croisée du critère RMSE retenue est la suivante :

- *n.trees* = 1000
- *interaction.depth* = 4
- *shrinkage* = 0,01

Le critère RMSE ainsi obtenu sur l'échantillon d'apprentissage par validation croisée est de 0,2132.

Il est ensuite calculé l'algorithme du *stochastic gradient boosting*, avec la combinaison des hyperparamètres optimaux. De la même manière que pour l'algorithme de forêt aléatoire il est possible de visualiser l'importance des variables dans le modèle.

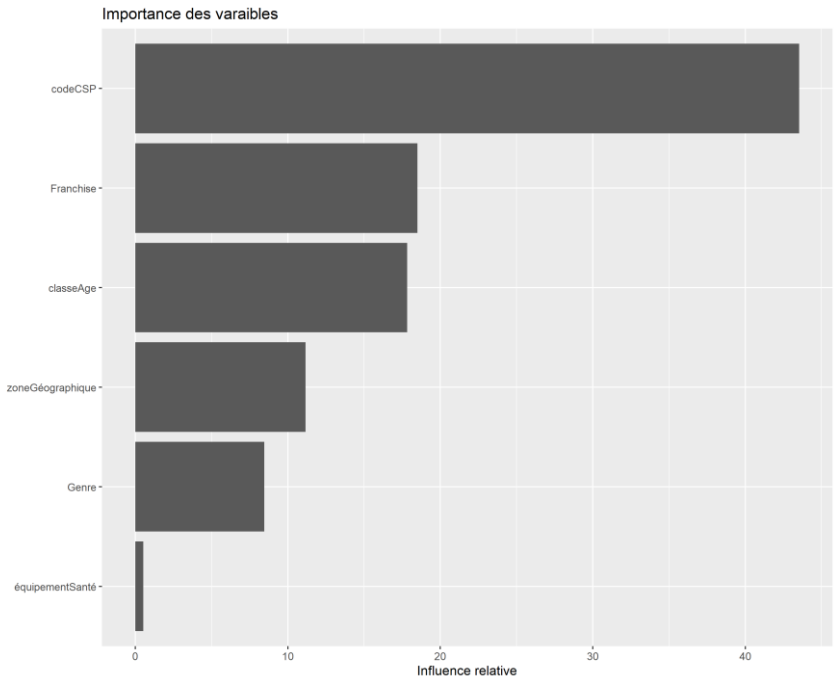


Figure 32 : Importance des variables explicatives pour l'algorithme GBM

A chaque division de chaque arbre, l'algorithme calcule l'amélioration en MSE et fait ensuite la moyenne sur chaque variable de l'amélioration faite sur tous les arbres construits lors de la descente du gradient. Les 6 variables explicatives du modèle exercent une influence plus ou moins forte. Il ressort ici que la variable *codeCSP* est à nouveau la variable la plus importante. La variable *équipementSanté,* quant à elle, semble exercer une influence limitée dans ce modèle. L'ordre ainsi que la valeur de l'importance des différentes variables varient significativement entre celles observées pour la *forêt aléatoire* et celle du *gradient boosting*.

Ensuite, le modèle construit sur la base d'apprentissage est appliqué sur la base test.

Afin de visualiser la descente du gradient il a été représenté l'évolution du critère RMSE en fonction du nombre d'arbres considérés.

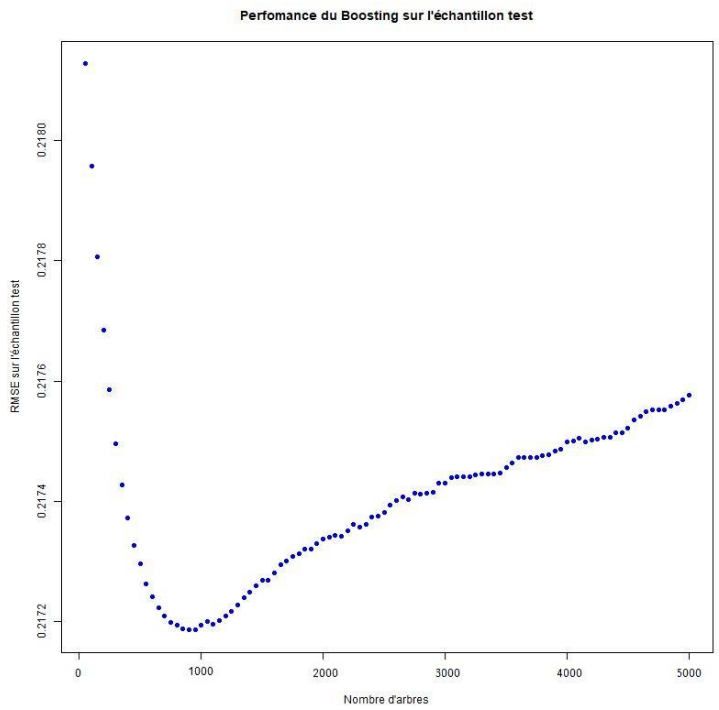


Figure 33 : Représentation de la descente du gradient sur l'échantillon test

Le minimum de la métrique d'erreur RMSE sur l'échantillon test est trouvé pour un algorithme comprenant 900 arbres. Les différentes métriques d'erreur sont exprimées dans le tableau suivant :

Métrique d'erreur	RMSE	MAE	Erreur 10%
Stochastic gradient boosting	0,217	0,069	0,516

Tableau 9 : Résultats des métriques d'erreur du modèle obtenu par gradient boosting

Concernant la fréquence moyenne prédite, le modèle construit à partir de l'algorithme du *gradient boosting* présente une prédiction moyenne proche de celle réellement observée. La fréquence prédite est 0,8% plus grande que la fréquence réelle. Le biais de prédiction du modèle par *gradient boosting* est donc négligeable.

Les différents algorithmes considérés ont été appliqués. Il convient désormais de comparer les différents modèles construits entre eux dans la partie suivante.

6 - Les résultats et comparaisons

Il a ainsi été appliqué quatre algorithmes :

- Le GLM, méthode de référence en assurance.
- L'arbre CART, classifieur fiable utilisé dans les méthodes ensemblistes.
- La forêt aléatoire, méthode ensembliste d'agrégation de classifieurs faibles.
- Le gradient boosting, méthode ensembliste adaptative de descente de gradient.

6.1 - Le récapitulatif des métriques d'erreur

Dans un premier temps, les différentes métriques d'erreur obtenues après application de chacun de ces algorithmes sont récapitulées dans le tableau suivant :

Métrique d'erreur	RMSE	MAE	Erreur 10%
GLM	0,224	0,078	0,525
Arbre CART	0,280	0,098	0,569
Forêt aléatoire	0,268	0,092	0,557
Stochastic gradient boosting	**0,217**	**0,069**	**0,516**

Tableau 10 : Comparaison des métriques d'erreur selon les algorithmes utilisés

Ce tableau indique que l'algorithme le plus performant sur chacune des métriques d'erreur est l'algorithme du *stochastic gradient boosting*. Les modèles obtenus par les algorithmes CART et de forêt aléatoire ne dépassent pas la performance du modèle obtenu par GLM.

6.2 - La fréquence moyenne prédite

Il est également intéressant de calculer la fréquence moyenne prédite par chaque modèle et de comparer cette fréquence avec la fréquence moyenne réelle de l'échantillon test. Les écarts relatifs par rapport à la fréquence réelle sont résumés dans le tableau suivant :

	GLM	Arbre CART	Forêt aléatoire	Stochastic Gradient boosting
Ecart relatif fréquence	-3,06%	2,95%	1,77%	**0,79%**

Tableau 11 : Ecarts relatifs des fréquences moyennes prédites par rapport à la fréquence moyenne réelle observée sur l'échantillon test

Cette comparaison permet de compléter l'analyse effectuée à l'aide des métriques d'erreur. En effet, les métriques d'erreur permettent d'estimer la performance d'un modèle en comparant individuellement pour chaque observation l'écart entre la valeur prédite et la valeur réelle. La comparaison de la fréquence moyenne de l'échantillon test permet d'estimer si, au global, le modèle présente une fréquence moyenne proche de celle de l'échantillon test.

Malgré ses relatives bonnes métriques d'erreur, le modèle obtenu par GLM est celui qui s'éloigne le plus de la fréquence moyenne réellement observée. Le modèle GLM a ainsi tendance à prédire une fréquence moyenne 3,06% inférieure à celle réellement observée. Cela est problématique car dans le cadre d'un tarificateur, ce biais résulte à une sous-tarification du risque et ainsi à une potentielle perte pour l'organisme d'assurance. Sur ce point les algorithmes CART et de forêt aléatoire sont plus performantes que le GLM. Ces algorithmes prédisent une fréquence moyenne respectivement 2,95% et 1,77% supérieures à la fréquence moyenne réelle. L'algorithme *stochastic gradient boosting* est, tout comme pour les métriques d'erreur, le plus performant au regard de la fréquence moyenne prédite avec une prédiction qui est 0,79% supérieure à celle observée sur l'échantillon test. Le modèle obtenu à partir de l'algorithme de *stochastic gradient boosting* est ainsi le meilleur modèle au sens des métriques d'erreur ainsi que de la fréquence moyenne prédite.

Le modèle de fréquence est ainsi finalisé. Il s'agit désormais d'établir le modèle de coût moyen. Le coût moyen correspond ici à la durée moyenne en arrêt des sinistres sur la base des observations sinistrées. L'objectif est d'allier les prédictions des deux modèles afin d'estimer la prime pure résultante.

VI - La construction du modèle de coût moyen des arrêts de travail

1 - La préparation des données et le retraitement des variables

1.1 - L'adaptation du modèle de coût moyen au risque arrêt de travail

La prime pure dans un modèle de coût moyen x fréquence est définie de la manière suivante :

$$Prime\ pure = Fréquence\ du\ sinsitre * Montant\ du\ sinistre$$

Dans le cadre de l'arrêt de travail, le montant du sinistre se décompose comme suit :

$$Montant\ du\ sinistre = Durée\ de\ l'indemnisation * Montant\ de\ la\ prestation\ journalière$$

Pour rappel, le montant de la prestation journalière peut être de nature :

- Forfaitaire
- Indemnitaire

Dans le cas d'une prestation journalière forfaitaire le montant de la prestation journalière est constant dans le temps et donc indépendant de la durée d'indemnisation. Dans ce cas, la durée de l'indemnisation est directement proportionnelle au montant du sinistre.

Dans le cas où la prestation journalière est de nature indemnitaire, le montant de la prestation servie est dépendant du remboursement du régime obligatoire qui est lui-même dépendant de la durée d'indemnisation. Dans ce cas, le montant de la prestation journalière n'est plus constant mais devient variable et dépendant de la durée d'indemnisation. Il en résulte que le lien entre la durée d'indemnisation et le montant du sinistre n'est plus directement proportionnel. En pratique, les différents régimes obligatoires des TNS présentent des durées de couverture ayant des valeurs seuils similaires. L'intégralité des différentes caisses proposent des indemnités journalières constantes sur les périodes suivantes :

- De 1 à 90 jours
- De 91 jours à 365 jours
- De 366 jours à 1095 jours

De cette manière, pour chacune de ces périodes, le montant de l'indemnité journalière est constant. Une solution afin de prendre en compte le caractère indemnitaire de la prestation est de tordre les primes pures sur chacun de ces paliers.

Finalement, l'enjeu du modèle de coût moyen est donc celui de modéliser la durée moyenne de l'indemnisation d'un sinistre pour des situations où il existe une relation de proportionnalité entre ces deux variables.

Il sera, dans ce mémoire, uniquement présenté le modèle de coût moyen dans le cas où la prestation journalière versée est de nature forfaitaire. Dans cette situation, la relation de proportionnalité entre la durée d'indemnisation et le coût moyen est toujours vérifiée.

La durée d'un arrêt de travail n'est pas directement proportionnelle au coût moyen car ce dernier dépend également de la franchise du contrat. La durée d'indemnisation, quant à elle, correspond de manière directement proportionnelle au coût que l'arrêt représente pour l'organisme d'assurance.

La durée d'indemnisation pour un arrêt est calculée comme suit :

$$Durée_{Indemnisation} = Durée_{Arrêt} - Franchise$$

Cette durée d'indemnisation correspond à la durée de l'arrêt de travail à laquelle est retranchée la durée de la franchise.

1.2 - La problématique de la censure à droite

Dans le cadre de l'arrêt de travail, il existe une problématique concernant les sinistres en cours lors de l'extraction des données. La date de fin d'observation simule alors une fin d'arrêt de travail pour les sinistres en cours. Il s'agit d'une censure à droite car la durée d'arrêt pour certaines données sont majorées à la date d'extraction des données. Or, l'adhérent peut en réalité rester en arrêt pour une durée non connue supérieure à celle à la date de fin d'observation.

La base de données étudiée possède cette particularité qu'elle intègre les provisions mathématiques dans la durée d'indemnisation. La problématique de la censure à droite est ainsi résolue puisque la durée retenue comprend la durée probable future des sinistres en cours. Cette information supplémentaire dans l'expression de la durée permet ainsi de contourner cette problématique et de considérer la situation comme un modèle de coût avec des durées ne représentant que des charges ultimes. Par cette méthode, la durée totale indemnisée est la durée effectivement indemnisée jusqu'à la date de fin d'observation additionnée à la durée de la provision mathématique constituée. La contrepartie de cette méthode est qu'elle introduit un biais dans la modélisation en considérant la provision dans la variable à expliquer qui est par nature estimée. Il est donc effectué une modélisation de la durée estimée du coût et non une modélisation de la durée réelle. Ce biais est cependant à mettre en perspective car il ne concerne que 4,16% des sinistres du portefeuille, son impact reste donc relativement limité. Il a ainsi été considéré que dans ces conditions le biais intégré par la durée d'indemnisation comprenant les provisions mathématiques est acceptable et donc que sa modélisation reste pertinente.

1.3 - La définition de la variable à expliquer

La variable à expliquer correspond à la durée moyenne indemnisée par sinistre pour un adhérent donné et une année de survenance donnée. Cette variable s'écrit :

$$Durée_{Moyenne} = \frac{Durée_{indemnisation_{totale}}}{nombre\ d'arrêts}$$

Avec :

- $Durée_{indemnisation_{totale}}$, correspond à la somme des durées d'indemnisation des arrêts de travail d'un adhérent sur une année de survenance.
- $nombre\ d'arrêts$, correspond aux nombres d'arrêts de travail d'un adhérent sur une année de survenance.

La répartition de la durée moyenne des sinistres peut être représentée par une boîte à moustaches qui renseigne les déciles, les quantiles ainsi que la médiane :

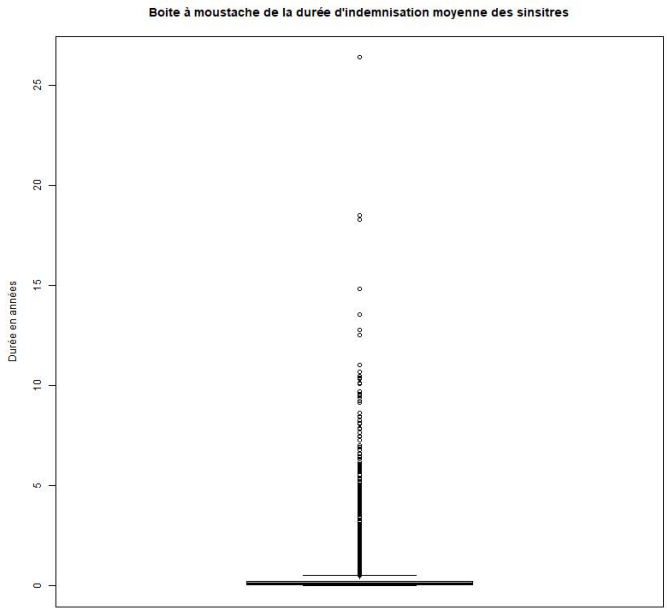


Figure 34 : Boite à moustache de la variable duréeMoyenne

Ce graphique démontre une très forte concentration des durées moyennes inférieures à 1 an. De sinistres extrêmes sont également présentés, ils correspondent principalement à des invalidités. Le maximum s'élève à une durée d'indemnisation de 26,4 ans. Ces sinistres extrêmes créent un écart important entre la moyenne et la médiane de la durée moyenne d'indemnisation d'un sinistre.

1.4 - Le retraitement et la description des variables

La partie concernant la préparation des données a été effectuée lors de la construction du modèle de fréquence. Les données ne nécessitent donc pas de retraitement spécifique. Cependant, la taille de l'échantillon utilisé dans le modèle de coût moyen est considérablement plus faible puisqu'il est retenu uniquement les données sinistrées. Certaines modalités nécessitent d'être regroupées afin de réduire le nombre de modalités dans chaque variable et ainsi limiter le biais lié à la faible exposition.

1.4.1 - La franchise

A titre d'exemple, les modalités *60/60* et *90/90* de la variable *Franchise* présentent chacune un nombre trop faible de sinistres pour pouvoir être intégrées à la modélisation. Deux facteurs peuvent expliquer ce nombre réduit de sinistres. D'une part les modalités *60/60* et *90/90* font partie des modalités possédant le moins d'effectifs dans le portefeuille. D'autre part la variable *Franchise* induit une troncature à gauche car les sinistres de durée inférieure à la franchise ne sont pas observés. Il est ainsi uniquement observé les sinistres de longue durée pour les franchises *60/60* et *90/90* ce qui explique en partie leur faible représentation dans la base des sinistres. Les modalités *60/60* et *90/90* sont donc exclues de la variable *Franchise* pour le modèle de coût moyen. Il s'agira de trouver une méthode alternative de modélisation pour ces deux modalités.

De cette manière la variable *Franchise* est composée des 4 modalités suivantes :

- *7/3*
- *15/3*
- *30/3*
- *30/30*

De fortes disparités de représentation sont observées. Les modalités *7/3* et *30/30* sont sous-représentées dans la base des sinistres. Une possibilité de biais de faible exposition demeure donc sur ces deux modalités. Elles sont néanmoins conservées car les statistiques descriptives ainsi que les résultats obtenus dans les différents modèles restent cohérents.

1.4.2 - La catégorie professionnelle

Au vu de la taille de la base des sinistres, la variable *codeCSP* créée pour le modèle de fréquence, n'est plus adaptée car elle possède une trop forte segmentation. Si cette segmentation était conservée en l'état, certaines modalités comporteraient trop peu d'années d'observation. Un choix consistant à conserver le niveau inférieur de détail, à savoir la variable *catégorieProfessionnelle*, a donc été effectué.

Elle est composée des 5 modalités suivantes :

- *GM*, les gérants majoritaires
- *EI*, les entrepreneurs individuels
- *PL_R*, les professions libérales réglementées
- *PL_RP*, les professions libérales réglementées paramédicales
- *PL_NR*, les professions libérales non réglementées

La modalité la moins représentée correspond aux professions libérales réglementées. Cette catégorie professionnelle comporte néanmoins un nombre conséquent d'observations. Le biais de faible exposition se retrouve ainsi fortement réduit.

1.4.3 - La classe d'âge

Le même raisonnement est appliqué sur les modalités de la variable *classeAge*. Les regroupements de modalités suivants sont effectués :

- *Moins de 29 ans*
- *30-39 ans*
- *40-49 ans*
- *50-59 ans*
- *60 ans et plus*

Les classes d'âge extrêmes sont logiquement les moins représentés dans la base des données sinistrées du fait de leur sous-représentation dans le portefeuille étudié. La modalité la moins représentée est celle des plus de 60 ans. Cette modalité n'est pas regroupée avec une autre classe d'âge car elle présente des résultats pertinents dans les différentes modélisations.

1.4.4 - La zone géographique

La variable *zoneGéographique* créée pour le modèle de fréquence est à adapter pour le modèle de coût moyen. Il est observé que les modalités *Région_1* et *Région_2* ne présente pas de différence notable sur le coût moyen. La même observation est remarquée sur les modalités *Région_3* et *Région_4.* Il a donc été construit des modalités regroupant les anciennes modalités, à savoir, *Région 1&2* et *Région 3&4*. La modalité *Région 5* n'est pas modifiée par rapport au modèle de fréquence.

Le portefeuille étudié représente une base de données conséquente de prévoyance sur le marché des TNS. Il subsiste néanmoins une problématique de quantité de données sur certaines modalités du fait de la nature du risque et de la taille de la population des TNS en France. Il a donc été nécessaire d'adapter la construction du modèle de coût moyen aux données disponibles. Cela se traduit notamment par une plus faible segmentation des variables explicatives afin de limiter le risque de faible exposition. Cette obligation de regroupement des modalités est au détriment de la précision finale du modèle de coût moyen et des estimateurs.

De la même manière que pour le modèle de fréquence, avant de commencer la modélisation, une étape de détermination de la base des sinistres attritionnels est nécessaire. Ensuite il sera également nécessaire de séparer cette base en deux afin d'obtenir un échantillon d'apprentissage et un échantillon test.

1.5 - L'obtention de la base des sinistres attritionnels

Les sinistres extrêmes sont à exclure de la modélisation pour éviter de faire peser le coût sur une partie spécifique de la population. Tout comme pour le modèle de fréquence, le coût de ces sinistres est réintroduit *a posteriori* de la modélisation. Il a été fait le choix de conserver dans la base des sinistres attritionnels le quantile 95% afin de conserver une quantité de données pertinente tout en réduisant l'impact des sinistres extrêmes sur la modélisation.

Finalement, la modélisation du coût moyen consiste ici dans un premier temps à modéliser les sinistres d'incapacité puis dans un second temps d'intégrer le coût des sinistres d'invalidité.

1.6 - Le découpage de la base des sinistres attritionnels

Le découpage de la base des sinistres attritionnels est effectué en prenant une proportion 70/30. Par souci de cohérence, ce ratio est identique à celui choisi pour le modèle de fréquence. Ainsi, l'échantillon d'apprentissage est constitué de 70% des données et l'échantillon test est donc composé de 30% des données. La sélection de ces données est aléatoire. Il est a posteriori vérifié que les durées moyennes sont similaires dans les deux échantillons afin de s'assurer qu'aucun biais n'est introduit à cette étape.

Le découpage considéré afin d'obtenir les deux échantillons d'apprentissage et de test, présente un écart entre les durées moyennes inférieur à 1%. Ce découpage est considéré satisfaisant et est donc celui retenu pour la modélisation de la durée moyenne des sinistres.

La procédure utilisée pour le modèle de fréquence est reconduite pour le modèle de coût moyen. Maintenant que les échantillons d'apprentissage et de test ont été créés, il convient d'établir les modèles à partir de chacun des algorithmes. Il sera calculé les métriques d'erreur pour chacun de ces modèles afin d'opter pour le modèle considéré optimal.

2 - Le modèle GLM

2.1 - Le choix de la loi de distribution des durées moyennes des sinistres

Dans un premier temps, il est nécessaire de sélectionner une loi de distribution de la durée des sinistres. La loi de distribution classiquement utilisée pour un modèle de coût est la loi Gamma. Elle repose sur 2 paramètres $\alpha > 0$ et $\beta > 0$. Sa densité s'écrit pour $x \geq 0$:

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

L'espérance et la variance s'exprime :

$$\mathbb{E}(Y) = \frac{\alpha}{\beta}$$

$$\mathbb{V}(Y) = \frac{\alpha}{\beta^2}$$

Le coefficient de variation est défini comme étant le rapport entre l'écart-type et la moyenne :

$$Coeff_{var} = \frac{\sqrt{\mathbb{V}(Y)}}{\mathbb{E}(Y)} = \frac{1}{\sqrt{\alpha}}$$

Il peut être considéré comme un coefficient de dispersion en posant $\phi = \frac{1}{\alpha}$, il est obtenu :

$$\mathbb{V}(Y) = \phi \mathbb{E}(Y)^2$$

La variance de la loi Gamma possède une forme quadratique. Dans le cas où $\alpha = 1$, la loi Gamma correspond à la loi exponentielle.

2.2 - L'adéquation du modèle aux données

Une fois la loi de distribution choisie, il est nécessaire de vérifier la bonne qualité d'ajustement du modèle.

Concernant celle-ci, les méthodes permettant de mesurer l'adéquation du modèle aux données sont explicitées en annexe, à savoir :

- La méthode de la déviance
- Le test de Pearson

Les p-value obtenues pour ces deux tests sont proches de 1 et donc supérieures à 5 %. Ainsi, il ne peut être rejeté l'hypothèse nulle au risque de 5%. L'adéquation du modèle aux données est donc validée avec une certitude de 95%.

2.3 - Les tests de significativité

Un premier test de significativité est le test de significativité globale du modèle. La p-value obtenue est inférieur au seuil des 5%, le modèle est donc globalement significatif.

Les tests de significativité de Wald permettent de remarquer que les variables *Genre* et *équipementSanté* ne sont pas significatives avec une confiance de 95%. Il est également comparé les critères AIC des différents modèles obtenus avec différentes combinaisons de variables. Il s'avère que le critère AIC est minimal en considérant les variables ci-dessous :

- *Franchise*
- *catégorieProfessionnelle*
- *classeAge*
- *zoneGéographique*

Au vu de ces éléments, les variables *Genre* et *équipementSanté* ne sont pas retenues pour le modèle de coût moyen. Les tests de significativité ainsi obtenus permettent de valider la cohérence du modèle ainsi que le regroupement effectué des modalités pour chaque variable.

2.4 - L'analyse des résidus

Il est tracé ci-dessous le diagramme quantile-quantile des résidus de déviance du modèle obtenu pour la loi de distribution Gamma :

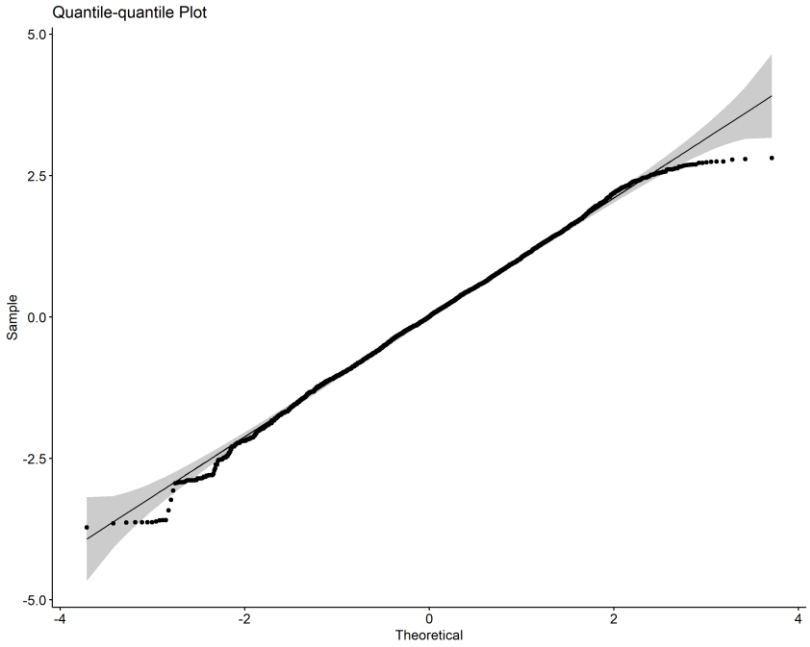


Figure 35 : Diagramme quantile-quantile des résidus de déviance du modèle

La p-value du test de Shapiro-Wilk est inférieure à 5%. Ainsi l'hypothèse de la distribution normale des résidus est rejetée avec un niveau de confiance de 95% pour ce modèle. Cependant, la distribution des résidus graphiquement représentés sur le Q-Q plot reste relativement pertinente.

2.5 - La performance du GLM

Les métriques d'erreur obtenues sont les suivantes :

Métrique d'erreur	RMSE	MAE	Erreur 10%
GLM	0,212	0,135	0,523

Tableau 12 : Résultats des métriques d'erreur du modèle obtenu par GLM

La durée moyenne prédite sur l'échantillon test est seulement 0,2% supérieure à la durée moyenne réellement observée. Il n'existe donc pas de biais entre la durée moyenne prédite et la moyenne observée sur l'échantillon test.

Le rejet de la normalité des résidus du modèle par le test de Shapiro-Wilk n'est pas satisfaisant pour pouvoir valider ce modèle. Cet algorithme est à comparer aux méthodes d'apprentissage supervisé avec pour objectif d'obtenir de meilleures métriques d'erreur.

3 - L'Arbre CART

La première méthode d'apprentissage supervisé qui est modélisée est l'algorithme CART. Afin de ne pas introduire de biais, il est utilisé les mêmes échantillons d'apprentissage et de test que ceux ayant servi pour le GLM. Les variables explicatives considérées pour l'ensemble des méthodes d'apprentissage supervisé sont celles sélectionnées par le GLM.

Il est appliqué la même procédure que celle suivie pour le modèle de la fréquence des sinistres. Dans un premier temps, il faut donc construire l'arbre saturé. Il convient ensuite de sélectionner le paramètre de complexité optimal au sens de l'erreur de validation croisée afin d'établir l'arbre élagué.

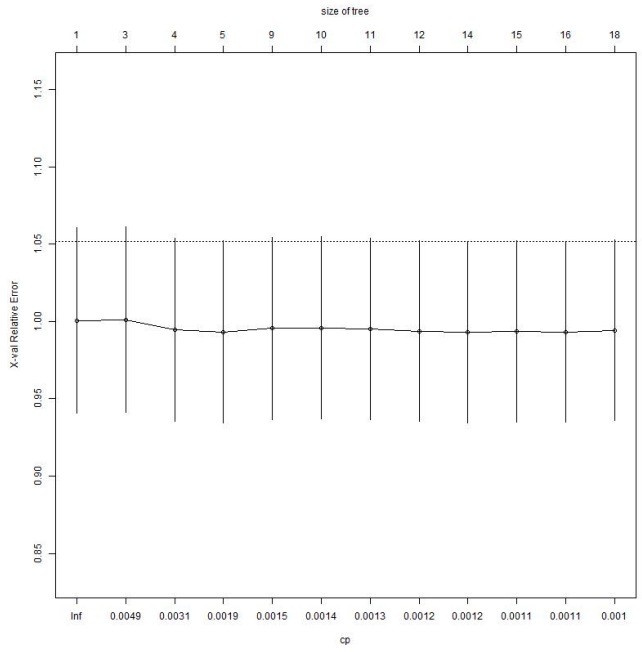


Figure 36 : Représentation de l'erreur de validation croisée en fonction de la complexité

Ce graphique montre comment l'erreur de validation croisée ne varie et ne diminue que très peu en fonction du paramètre de complexité. Elle reste en effet proche de 1. Le paramètre de complexité *cp*

retenu est celui qui minimise l'erreur de validation croisée, c'est-à-dire, $cp = 0{,}0019$. L'arbre élagué est donc obtenu en choisissant ce *cp* optimal.

Après avoir obtenu l'arbre élagué, les prédictions sur l'échantillon test peuvent être estimées et les métriques d'erreur évaluées. De cette manière, les métriques d'erreur sont indiquées dans le tableau ci-dessous :

Métrique d'erreur	RMSE	MAE	Erreur 10%
GLM	**0,212**	**0,135**	**0,523**
Arbre CART	0,216	**0,135**	0,531

Tableau 13 : Résultats des métriques d'erreur du modèle obtenu par arbre CART

Au regard des trois métriques d'erreur, le modèle construit par l'algorithme CART est moins performant que le GLM. Tout comme pour la méthode GLM, la durée moyenne prédite d'un sinistre par le modèle d'arbre CART ne présente pas d'écart significatif avec celle observée sur l'échantillon test.

Utilisé de cette manière, le modèle par algorithme CART n'est donc pas meilleur que celui par GLM. De la même façon que pour le modèle de fréquence, le classifieur faible qu'est l'algorithme CART peut cependant être utilisé pour les méthodes ensemblistes. En effet, les méthodes d'agrégation et adaptatives peuvent permettre d'obtenir des modèles plus stables et robustes. La partie suivante établit le modèle de la méthode d'agrégation par forêt aléatoire.

4 - La forêt aléatoire

Dans un premier temps, une recherche par quadrillage des hyperparamètres est effectuée. es valeurs suivantes ont été testées :

- *ntree* : de 100 à 1 000 avec un pas de 100
- *mtry* : de 1 à 4

Les résultats de cette recherche par quadrillage sont résumés dans le graphique suivant :

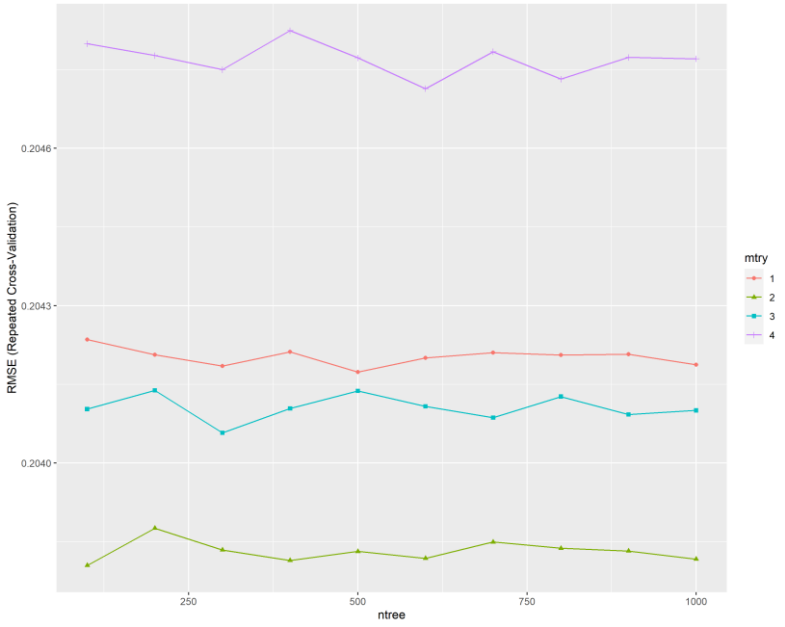


Figure 37 : Recherche par quadrillage des hyperparamètres de la forêt aléatoire

Pour une même valeur de mtry, le nombre d'arbres n'impacte que très peu les RMSE de validation croisée. En revanche, la variable mtry possède une forte influence sur le résultat. Le meilleur modèle au regard du critère RMSE est alors obtenu pour une forêt composée de 100 arbres et effectuant une randomisation de 2 variables à chaque nœud. Après avoir obtenu les hyperparamètres optimaux, il convient ensuite de calculer la forêt aléatoire qui en découle. Le niveau d'influence de chaque variable dans la construction du modèle est représenté ci-dessous :

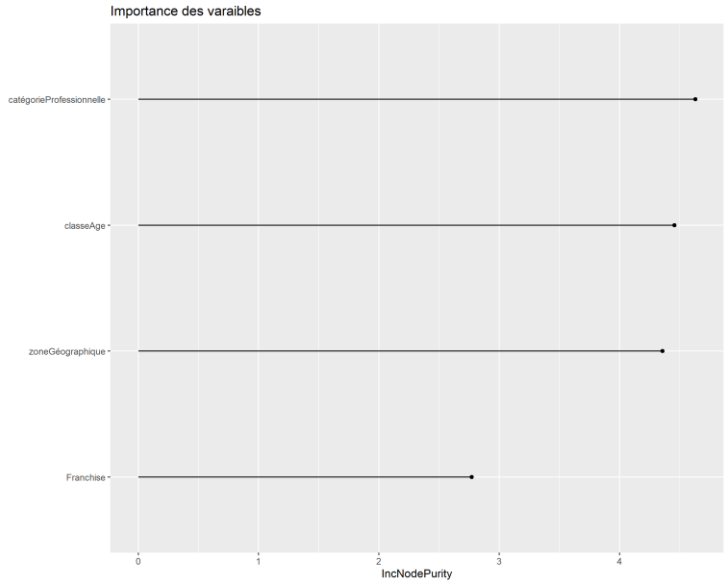


Figure 38 : Importance des variables pour l'algorithme de forêt aléatoire

La variable relative à la catégorie professionnelle du TNS est la variable qui possède le plus d'importance dans la construction du modèle. Une vérification a été menée afin de vérifier que les variables *Genre* et *équipementSanté* ne sont pas explicatives dans le modèle de durée moyenne. L'importance des deux variables s'avérait être effectivement proche de zéro. Cela conforte le fait de ne pas avoir conservé ces variables dans les différents algorithmes.

Le graphique ci-dessous représente l'impact du nombre d'arbres sur la stabilisation de la MSE.

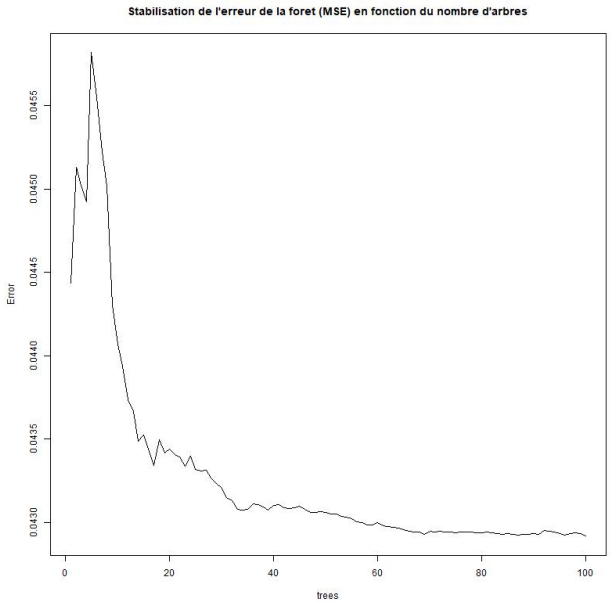


Figure 39 : Stabilisation de la MSE en fonction du nombre d'arbres

Ce graphique conforte l'idée selon laquelle une forêt constituée de 100 arbres semble être optimale au regard du critère MSE et donc également pour la métrique RMSE. Il convient de mesurer la performance du modèle à partir des différentes métriques d'erreur :

Métrique d'erreur	RMSE	MAE	Erreur 10%
GLM	**0,212**	**0,135**	**0,523**
Arbre CART	0,216	**0,135**	0,531
Forêt aléatoire	0,214	0,136	0,529

Tableau 14 : Résultats des métriques d'erreur du modèle obtenu par forêt aléatoire

Les métriques sont meilleures que celles obtenues sur le modèle de l'algorithme CART excepté pour la MAE qui est légèrement supérieure dans le modèle obtenu par forêt aléatoire en comparaison avec le GLM. Les métriques d'erreur restent en dessous de celles du modèle GLM. Il est enfin développé un modèle de durée moyenne construit sur l'algorithme de *gradient boosting*.

5 - Le *stochastic gradient boosting*

Dans un premier temps, une sélection des meilleurs hyperparamètres au sens de la validation croisée est nécessaire. Une recherche par quadrillage est à nouveau utilisée en considérant différentes combinaisons d'hyperparamètres. Après avoir effectué une première approche par dichotomie afin de réduire les possibilités, il a été testé les hyperparamètres suivantes :

- *n.trees* : 100 ; 200 ; 300 ; 400 ; 500
- *interaction.depth* : 2 ; 4 ; 6
- *shrinkage* : 0,01; 0,1

Il va ainsi être testé les 30 combinaisons suivantes :

Hyperparamètres	n.trees	interaction.depth	shrinkage
Combinaison 1	100	2	0.01
Combinaison 2	100	2	0.1
Combinaison 3	100	4	0.01
Combinaison 4	100	4	0.1
Combinaison 5	100	6	0.01
Combinaison 6	100	6	0.1
Combinaison 7	200	2	0.01
Combinaison 8	200	2	0.1
Combinaison 9	200	4	0.01
Combinaison 10	200	4	0.1
Combinaison 11	200	6	0.01
Combinaison 12	200	6	0.1
…	…	…	…
Combinaison 25	500	2	0.01
Combinaison 26	500	2	0.1
Combinaison 27	500	4	0.01
Combinaison 28	500	4	0.1
Combinaison 29	500	6	0.01
Combinaison 30	500	6	0.1

Tableau 15 : Représentation des différentes combinaisons d'hyperparamètres testés

Le calcul de la validation croisée a été effectué sur un 5-*folds*. La métrique d'erreur utilisée pour cette étape reste le RMSE.

Il est ci-dessous exprimé les résultats de l'étape de *tuning* effectué à l'aide des fonctions *train()* et *expand.grid()* de R.

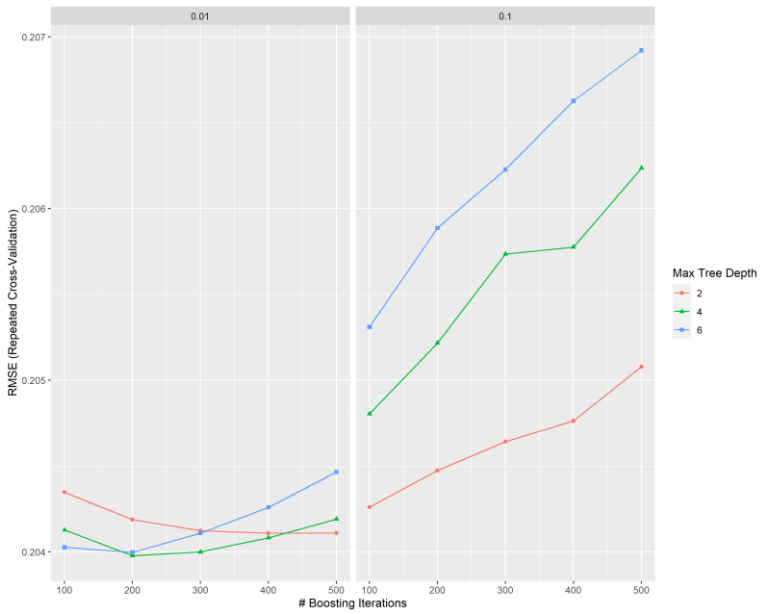


Figure 40 : Recherche par quadrillage des hyperparamètres du modèle par gradient boosting

Le graphique de gauche représente les courbes d'évolution de la RMSE en fonction du nombre d'arbres pour un *learning rate* de 0,01. Celui de droite représente les mêmes informations pour un *learning rate* de 0,1. Le graphique du taux d'apprentissage de 0,1 montre ici que le RMSE augmente avec le nombre d'itérations. Cela est expliqué par le fait que le learning rate considéré est trop grand. Le minimum de la fonction de perte n'est donc pas trouvé. Il s'agit du même cas que celui observé pour le modèle de fréquence.

La meilleure combinaison d'hyperparamètres au sens de la validation croisée du critère RMSE est retenue :

- *n.trees* = 200
- *interaction.depth* = 4
- *shrinkage* = 0,01

Le critère RMSE ainsi obtenu sur l'échantillon d'apprentissage par validation croisée est de 0,204.

L'algorithme du *stochastic gradient boosting*, avec la combinaison des hyperparamètres optimaux, est ensuite calculé.

De la même manière que pour l'algorithme de forêt aléatoire, le graphique ci-dessous permet de visualiser l'importance des variables dans le modèle.

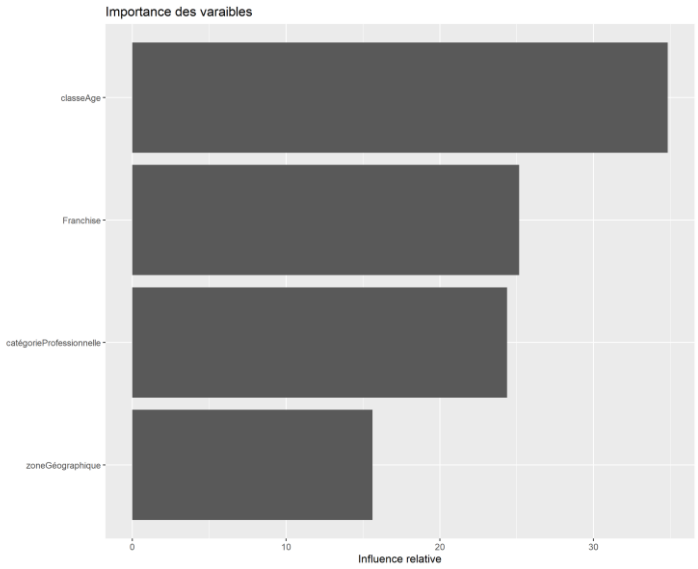


Figure 41 : Importance des variables dans le modèle par gradient boosting

Les 4 variables explicatives du modèle exercent effectivement une influence plus ou moins importante. La variable *classeAge* est la variable la plus importante. La variable *zoneGéographique* quant à elle exerce une influence plus limitée dans ce modèle. Il est intéressant d'observer les différences notables sur le classement de l'importance des variables entre le modèle obtenu par forêt aléatoire et le modèle obtenu par gradient boosting.

Ensuite, le modèle construit est appliqué sur la base test. Afin de visualiser la descente du gradient, le graphique ci-dessous représente l'évolution du critère RMSE en fonction du nombre d'arbres retenu :

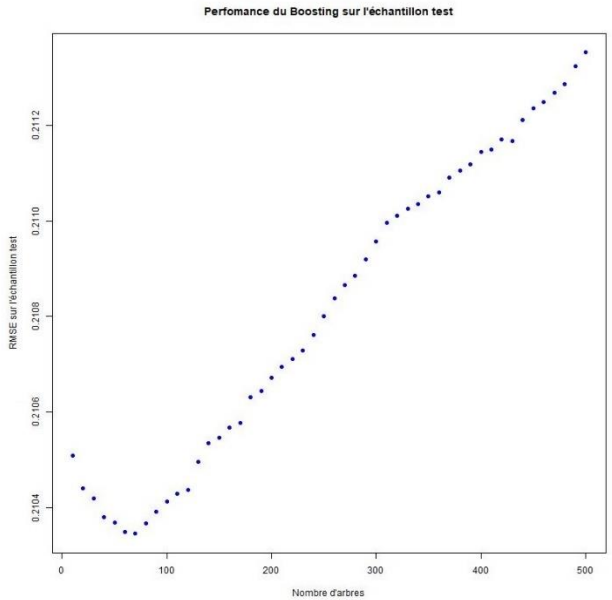


Figure 42 : Représentation de la descente du gradient boosting sur l'échantillon test

Le minimum de la métrique d'erreur RMSE sur l'échantillon test est trouvé pour un algorithme comprenant 70 arbres. Ce nombre d'arbres est relativement faible pour une méthode comme le gradient

boosting mais il semble pertinent au regard du nombre de données que comporte la base de durée moyenne des sinistres.

Les différentes métriques d'erreur sont résumées dans le tableau suivant :

Métrique d'erreur	RMSE	MAE	Erreur 10%
Stochastic gradient boosting	0,210	0,134	0,522

Tableau 16 : Résultats des métriques d'erreur du modèle obtenu par gradient boosting

Le modèle obtenu par gradient boosting ne présente, tout comme les autres algorithmes, un biais de prédiction nul par rapport à la durée moyenne observée sur l'échantillon test.

L'ultime partie consiste à comparer les différents modèles obtenus entre eux afin de déterminer lequel est le plus adapté aux données du portefeuille.

6 - La comparaison des modèles

Cette section a pour but de comparer les différents algorithmes entre eux. Cette comparaison va permettre de sélectionner un modèle optimal de durée moyenne en arrêt de travail.

6.1 - Le récapitulatif des métriques d'erreur

Dans un premier temps, il est récapitulé les différentes métriques d'erreur obtenues après application de chacun de ces algorithmes :

Métrique d'erreur	RMSE	MAE	Erreur 10%
GLM	0,212	0,135	0,523
Arbre CART	0,216	0,135	0,531
Forêt aléatoire	0,214	0,136	0,529
Stochastic gradient boosting	**0,210**	**0,134**	**0,522**

Tableau 17 : Comparaison des métriques d'erreur selon les algorithmes utilisés

L'algorithme le plus performant sur chacune des métriques d'erreur est l'algorithme du *stochastic gradient boosting*. Il est néanmoins observé que l'écart entre les différentes métriques d'erreur s'avère réduit. L'intérêt de la comparaison des différents modèles entre eux pour le modèle de coût moyen est donc limité. Cette observation est principalement expliquée par le fait que la segmentation peu fine ne permet pas un niveau de granularité suffisant pour obtenir une modélisation précise du coût moyen.

6.2 - La durée moyenne prédite

Il est ici pertinent de calculer la durée d'arrêt moyenne prédite par chaque modèle et de la comparer avec la durée d'arrêt moyenne réelle de l'échantillon test. Les écarts relatifs par rapport à la durée moyenne d'un sinistre observée sur l'échantillon test sont exprimés dans le tableau suivant :

	GLM	Arbre CART	Forêt aléatoire	Stochastic Gradient boosting
Ecart relatif durée moyenne	0,16%	-1,13%	-0,16%	-0,16%

Tableau 18 : Ecarts relatifs des durées moyennes prédites par rapport à la durée moyenne réelle observée sur l'échantillon test

Tout comme pour les métriques d'erreur, il est remarqué que la durée moyenne prédite ne varie que très peu entre les différents modèles. La valeur maximale prédite et la valeur minimale prédite correspondent respectivement au modèle GLM et au modèle utilisant l'arbre CART. Néanmoins, chacun des modèles reste proche de la valeur moyenne réelle de l'échantillon test.

En combinant l'analyse effectuée sur les métriques d'erreur à celle sur la durée moyenne prédite, le modèle qui est le plus performant est à nouveau celui construit avec l'algorithme de *stochastic gradient boosting.* Cependant, le gain en performance par rapport au GLM est plus limité que sur le modèle de durée moyenne.

VII - Les estimations des primes pures et comparaisons

Les modèles de fréquence et de coût moyen étant construits, il est, dans cette section, comparé les primes pures résultantes. Il est dans un premier temps calculé la prime pure qui va être désignée prime pure hors IJ (Indemnité Journalière). Il s'agit d'une prime pure avant prise en compte du montant des indemnités journalières. Cette prime pure est calculée de la manière suivante :

$$Prime\ pure\ hors\ IJ = Fréquence\ du\ sinistre * Durée\ moyenne\ du\ sinistre$$

Il sera d'abord présenté la prime pure sans, puis avec la prise en compte des sinistres extrêmes. Les caractéristiques des différents algorithmes seront ensuite synthétisés.

1 - La prime pure hors IJ des sinistres attritionnels

Les écarts des primes pures obtenues pour chaque modèle par rapport à la prime pure du portefeuille sont résumés dans le tableau suivant :

	GLM	Arbre CART	Forêt aléatoire	Stochastic gradient boosting
Ecart relatif prime pure	-2,91%	1,68%	1,55%	0,70%

Figure 43 : Ecart relatif des primes pures des sinistres attritionnels par rapport à la prime pure moyenne observée sur l'échantillon test

Ces primes pures sont celles obtenues en ne considérant que les sinistres attritionnels que ce soit en fréquence ou en coût moyen. Les écarts sur la prime pure obtenue pour chaque modèle sont essentiellement expliqués par le modèle de fréquence car le modèle de coût moyen ne varie que très peu en fonction de l'algorithme utilisé.

Cette comparaison permet plusieurs observations. Le modèle GLM présente la prime pure la plus faible. Le GLM sous-tarifie ainsi le risque des sinistres attritionnels de 2,91%. Les modèles obtenus par arbre de décision et par forêt aléatoire présentent des primes pures similaires. Elles surévaluent la prime pure respectivement de 1,68% et 1,55%. Le modèle le plus proche de la prime pure observée est celui du *gradient boosting* avec un écart limité de 0,70%.

A cela s'ajoute l'analyse préalable sur les métriques d'erreur. Cette analyse a déterminé que le *stochastic gradient boosting* est l'algorithme qui présente les métriques d'erreur les plus performantes sur les modèles de fréquence et de coût moyen. Au regard de ces éléments, le *stochastic gradient boosting* s'impose comme l'algorithme qui produit le meilleur modèle.

2 - L'intégration des sinistres extrêmes dans la prime pure hors IJ

Les sinistres extrêmes restent à être introduits dans la prime pure. Le choix qui a été fait est celui de répercuter le coût des sinistres extrêmes sur l'ensemble de la population sans distinction particulière. Il aurait également été possible d'impacter les sinistres extrêmes de manière différente sur certains groupes de risques homogènes spécifiques. Les sinistres extrêmes comprennent ceux liés à la fréquence et ceux liés à la durée moyenne.

D’une part, les fréquences extrêmes retirées *a priori* de la modélisation de la fréquence sont introduites. Dans l’hypothèse d’une répartition des fréquences extrêmes sur l’ensemble du portefeuille, l’impact sur la fréquence correspond à une augmentation de 4,75% par rapport à la fréquence des sinistres attritionnels. L’impact sur les fréquences reste limité car il s’agit uniquement du quantile 1% des sinistres extrêmes et ces sinistres extrêmes restent relativement proches des sinistres attritionnels.

D’autre part, les durées moyennes extrêmes retirées *a priori* de la modélisation du coût moyen sont intégrées. Le quantile 5% des sinistres extrêmes représente une durée moyenne totale importante. S’il est décidé de répartir ces sinistres extrêmes sur l’ensemble de la base sinistrée alors l’impact sur le coût moyen est une augmentation de la durée moyenne de 91,2%. L’impact de la prise en compte des sinistres extrêmes est plus important sur la durée moyenne que sur la fréquence. La première explication à cette observation est qu’à la différence du modèle de fréquence, il a été choisi le quantile 5% et non le quantile 1%. La seconde explication provient du fait que les sinistres extrêmes sont considérablement plus éloignés des sinistres attritionnels dans le modèle de durée moyenne. La conséquence est que l’impact de l’ajout des sinistres extrêmes vient plus que doubler la durée moyenne d’indemnisation.

Dans l’hypothèse où ces sinistres sont répartis sur l’ensemble du portefeuille, les nouveaux écarts par rapport à la prime pure du portefeuille sont résumés dans le tableau ci-dessous :

	GLM	Arbre CART	Forêt aléatoire	Stochastic gradient boosting
Ecart relatif prime pure	-2,84%	2,12%	1,54%	0,73%

Figure 44 : Ecart relatif des primes pures après intégration des sinistres extrêmes par rapport à la prime pure moyenne observée sur l’échantillon test

3 - L’adaptation des primes pures hors IJ à la politique de souscription

Les primes pures ainsi obtenues ne prennent en compte uniquement que des considérations techniques. En pratique, une phase d’adaptation des primes pures à la politique de souscription est nécessaire.

A titre d’exemple, le fait de tarifer par classe d’âge en considérant des intervalles allant de 5 à 10 ans peut s’avérer peu adapté dans le cadre d’une tarification. Il semble préférable d’appliquer un lissage des primes pures afin d’obtenir des primes pures qui évoluent d’une année à l’autre. Ce lissage a pour objectif de limiter l’effet de seuil que les classes d’âge peuvent créer sur les primes pures. Ainsi, cette phase de lissage permet de trouver un bon compromis entre la fidélité des tarifs aux données et la régularité recherchée dans les tarifs appliqués par la politique de souscription. Ce lissage peut être réalisé avec des méthodes telles que celle de Whittaker-Henderson.

A cela s’ajoute le retraitement spécifique lié à la variable *Genre*. Comme évoqué lors des statistiques descriptives, la variable *Genre* ne peut être considérée comme discriminante dans la tarification. Il convient toutefois de vérifier que la répartition Homme/Femme n’évolue pas de manière significative dans le temps par rapport à celle du portefeuille ayant servi à la construction du modèle. Dans le cas où la répartition Homme/Femme évoluerait de manière significative dans le temps, il conviendrait d’impacter les primes pures en conséquence. Une méthode pourrait être celle d’utiliser les coefficients de la variable *Genre* obtenus dans les modèles et de les appliquer à l’échelle du portefeuille en fonction de la répartition *Homme/Femme*.

D’autres ajustements peuvent également être appliqués en fonction des résultats obtenus si ces derniers s’avèrent contraire à la politique de souscription.

Conclusion

Plusieurs problématiques actuarielles ont été soulevées et abordées dans la conception de ce tarificateur en prévoyance destiné au marché des travailleurs non salariés.

Le premier enjeu est l'étude des spécificités de la population des travailleurs non salariés en France. Cette étude a révélé une grande diversité propre au profil du TNS. Ces différents éléments d'hétérogénéité intrinsèques à la population des TNS se révèlent être un enjeu majeur dans la tarification. Les variables explicatives sélectionnées dans les modèles nécessitent de refléter ces caractéristiques spécifiques afin d'obtenir une meilleure performance dans la modélisation. Il a pour cela été construit la variable *codeCSP*. Son rôle est de capter les informations concernant la catégorie professionnelle ainsi que le secteur d'activité du TNS avec une segmentation relativement fine tout en conservant une exposition suffisante dans chacune des modalités. Cette variable s'est révélée être très significative dans le GLM et être la variable la plus influente pour les méthodes d'apprentissage supervisé. Ces éléments ont ainsi confirmé la pertinence de cette variable et l'importance des informations qui la composent.

La deuxième problématique est l'adaptation de la tarification de type fréquence et coût moyen au risque arrêt de travail. Le modèle de fréquence n'a pas posé de difficulté particulière sur ce point. En effet, pour une année de survenance, la fréquence est simplement identifiée comme le nombre de sinistres sur l'année pondéré par la durée d'exposition. L'enjeu réside cependant dans la détermination du modèle de coût moyen. Dans un premier temps, la définition de la décomposition du coût moyen comme la durée moyenne d'un sinistre à laquelle est multipliée le montant de la prestation journalière a été introduite. La distinction du cas où la prestation journalière servie est forfaitaire ou indemnitaire a été abordée. La prestation forfaitaire s'est révélée être le type de prestation le plus adapté au modèle de coût moyen. La prestation indemnitaire présente une plus grande complexité par sa dépendance au remboursement du régime obligatoire du TNS. Il a été présenté des solutions afin de pouvoir adapter la méthodologie utilisée pour une prestation forfaitaire dans le cas indemnitaire.

Le troisième enjeu de la tarification du risque arrêt de travail sur la population des TNS réside dans l'adaptation à la quantité de données disponibles pour construire un tarificateur fiable et robuste. Cette base de données représente l'une des plus grandes bases de données sur la prévoyance des TNS. Cependant, la nature du risque arrêt de travail combinée à la taille de la population des TNS en France a montré quelques limites dans la modélisation. Le modèle de fréquence n'a pas présenté de difficulté particulière sur ce point puisqu'il est construit sur l'ensemble du portefeuille composé de plus de 100 000 observations. A l'inverse, le modèle de durée moyenne d'un arrêt est construit sur les années d'observations sinistrées. Or, l'arrêt de travail étant un risque relativement rare, la taille de la base de données sinistrées se réduit drastiquement. A cela, vient s'ajouter l'influence de la variable *Franchise* sur le nombre de sinistres observés. La franchise implique une troncature à gauche qui réduit le nombre de sinistres observés et donc la taille de la base sur laquelle peut être construit le modèle de coût moyen. La solution apportée a été de réduire le nombre de variables explicatives ainsi que le nombre de modalités des variables par regroupement des anciennes modalités. De cette manière, en contrepartie d'une segmentation moins fine du modèle de coût moyen par rapport au modèle de fréquence, ces deux choix ont permis d'obtenir une certaine fiabilité dans les résultats. Toutefois, cette méthode est limitée car il n'a parfois pas été possible de regrouper des modalités entre elles pour obtenir une exposition satisfaisante. A titre d'exemple, pour cette raison, les franchises *60/60* et *90/90* qui n'ont pas pu être intégrées au modèle de coût moyen.

La dernière problématique évoquée est la comparaison des modèles obtenus par les différents algorithmes. Le premier algorithme qui a été appliqué est l'algorithme classique en assurance qui est le modèle linéaire généralisé. Cet algorithme présente l'avantage d'être dans une certaine mesure simple à implémenter ainsi qu'à interpréter. Cependant le GLM est une méthode paramétrique et exige donc de

déterminer une loi de distribution des données du portefeuille. Cette loi de distribution peut ne pas présenter une bonne adéquation du modèle aux données. Les hypothèses sur lesquelles reposent l'algorithme sont fortes et peuvent dans certains cas être difficilement vérifiées. A titre d'exemple, les résidus dans les modèles de fréquence et de durée moyenne ne suivent pas une loi normale avec un niveau de confiance de 95%. Cela vient remettre en cause la pertinence du modèle GLM. De plus,le modèle de fréquence GLM présente une fréquence moyenne prédite inférieure de 3% à la fréquence réellement observée. Ces différents éléments ont nourri l'intérêt de développer des méthodes alternatives de tarification à savoir des méthodes d'apprentissage supervisé. Ces méthodes présentent l'avantage d'être non paramétriques. La première méthode considérée est celle de l'arbre CART. Elle s'avère rapide à mettre en place et à interpréter. Elle présente également l'avantage d'être une méthode non paramétrique. L'algorithme CART implique des modèles ayant un faible biais mais une forte variance. Des solutions ensemblistes reposant sur la combinaison de plusieurs arbres CART permettent de réduire la variance. La méthode de la forêt aléatoire repose sur l'agrégation de ces classifieurs faibles tandis que le gradient boosting utilise une approche séquentielle. Ces deux algorithmes sont chronophages lors de la phase d'entraînement et lors du réglage des hyperparamètres. Ils perdent également en interprétabilité des résultats par rapport à l'algorithme CART. Néanmoins, ces deux algorithmes permettent de gagner en stabilité ainsi qu'en précision. Le gain en performance est essentiellement observé sur le modèle de fréquence. Les différents algorithmes présentent des modèles de durée moyenne ayant des résultats relativement proches.

Finalement, le meilleur modèle au sens des métriques d'erreur ainsi que de l'écart par rapport à la prime pure observée est le *stochastic gradient boosting*. Cet algorithme est alors la méthode retenue pour la conception du tarificateur des arrêts de travail du marché des TNS. Cette méthode présente l'inconvénient d'être chronophage que ce soit dans la phase de réglage des hyperparamètres ou bien lors de la phase d'apprentissage. Toutefois, dans le cadre de la construction des bases techniques, cette opération n'est à réaliser qu'une à deux fois par an. Le temps d'exécution de cet algorithme n'est donc pas un frein à son utilisation dans le cadre du tarificateur.

Le modèle ainsi construit présente toutefois certaines limites. En effet, une première limite apparaît sur les franchises longues, notamment concernant les modalités *60/60* et *90/90*, qui demandent d'être calculées en dehors du cadre de ce portefeuille par manque d'effectifs. Une table de maintien d'expérience pourrait, par exemple, permettre cette tarification. Une seconde limite existe sur les coefficients du zonier qui sont calculés sur une échelle régionale et sont donc restreints en termes de précision. Un affinage du tarif est alors nécessaire afin de mieux correspondre au marché. Une solution pourrait être d'ajouter des zones de bonus ou de malus plus fines en fonction des informations du marché. L'idée général de ces retraitements serait alors de tordre le modèle mathématique pure pour l'adapter au marché du risque arrêt de travail des TNS.

Bibliographie

- Articles et ouvrages :

1. Breiman L., Friedman J., Olshen R., Stone C. (1984) *CART: Classification and Regression Trees*, Wadsworth International, New York.
2. McCullagh P., Nelder John A. (1989) *Generalized Linear Models, second edition.* Chapman and Hall/CRC, New York.
3. Friedman J., (1999) *Stochastic gradient boosting*. Stanford University, Standford.
4. Breiman L., (2001) *Random Forests*, Machine Learning, 45, 5-32.
5. Salembier L, Théron G, (2020) *Panorama de l'emploi et des revenus des non-salariés*, INSEE références, INSEE.

- Sites Internet :

6. CNAVPL : https://www.cnavpl.fr/statistiques/ (Statistiques sur les professions libérales, description des sections de la CNAVPL), site consulté le 15 juillet 2021.
7. http://https//www.secu-independants.fr/cpsti/documentation/lessentiel-en-chiffres/ (l'essentiel des chiffres sur les indépendants, catégories professionnels et secteurs d'activité), site consulté le 18 octobre 2021.
8. https://www.metlife.fr/blog/etude-csa-prevoyance-tns-2021/ (étude 2021 CSA/Metlife, prévoyance des TNS), site consulté le 8 janvier 2022.

- Cours :

9. Milhaud X., (2020-2021) *Pratiques avancées de tarification et de provisionnement*, ISFA, Lyon.

Annexe

1 - Complément partie théorique GLM

1.1 - Principe d'estimation des paramètres du GLM

Dans cette partie, le principe d'estimation des paramètres β_j d'un modèle linéaire généralisé est explicité.

Il convient de rappeler que le vecteur à expliquer $Y = (Y_1, \dots, Y_n)$ possède la structure d'une famille exponentielle suivante :

$$f_Y(y) = exp\left(\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)}\omega_i + \sum_{i=1}^{n} c_i(y_i, \phi) \right)$$

Par ailleurs, le modèle s'écrit pour $i \in [\![1; n]\!]$:

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

Et il a été prouvé la relation suivante :

$$\mu = b'(\theta)$$

Le vecteur des paramètres β du modèle linéaire généralisé est alors compris dans θ.

Les paramètres β_j sont estimés en utilisant la méthode du maximum de vraisemblance sur le modèle linéaire généralisé. Après avoir obtenu les estimations du vecteur β, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$, il est défini la valeur ajustée par le modèle :

$$\hat{Y} = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p)$$

Il convient d'expliciter plus en détail la méthode du maximum de vraisemblance. Pour cela, la notion de log-vraisemblance est introduite de la manière suivante :

$$\ell(\theta(\beta), y, \phi) = \ln f_Y(y) = \sum_{i=1}^{n} \ln\ f(y_i, \phi, \theta_i)$$

L'objectif est d'estimer β par maximum de vraisemblance, il convient donc de chercher les $\beta_0, \beta_1, \dots, \beta_p$ qui vérifient les équations suivantes pour $j \in [\![0\ ; p]\!]$:

$$\frac{\partial \ell(\theta(\beta), y, \phi)}{\partial \beta_j} = 0$$

En développant le calcul, il est obtenu :

$$\frac{\partial \ell(\theta(\beta), y, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\frac{\partial \mu_i}{\partial \eta_i} x_{ji}}{\frac{a(\phi)}{\omega_i} b''(\theta_i)}$$

Or $g(\mu_i) = \eta_i$ d'où $\frac{\partial \mu_i}{\partial \eta_i} = g'(\mu_i)$,

Ainsi, les équations de vraisemblance deviennent pour $j \in [\![0\,;p]\!]$:

$$\sum_{i=1}^{n} \omega_i (y_i - \mu_i) \frac{x_{ji}}{b''(\theta_i) g'(\mu_i)} = 0$$

Le paramètre ϕ n'apparaît plus dans les équations de vraisemblance. Il est donc possible d'estimer β sans se soucier de ϕ.

A la différence du modèle gaussien, il n'existe, en général, pas de solution explicite pour β. Il existe cependant une solution numérique en utilisant des méthodes itératives telles que :

- La méthode de Newton-Raphson qui est fondée sur le Hessien $[H]_{jk} = \frac{\partial^2 \mathcal{L}}{\partial \beta_j \partial \beta_k}$
- La méthode du score de Fisher qui est fondée sur la matrice d'information $I = X'WX$ dont le terme général est $[I]_{jk} = -\mathbb{E}\left(\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_j \partial \beta_k}\right) = -\sum_{i=1}^{n} \frac{x_{ji} x_{ki}}{\mathbb{V}(Y_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2$ et où W est la matrice diagonale de pondération $[W]_{ii} = \frac{1}{\mathbb{V}(Y_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2$

Si la fonction lien canonique est choisie alors les équations de vraisemblance peuvent se simplifier. En effet, la fonction lien canonique est telle que $g(\mu_i) = \eta_i = \theta_i = x_i{}'\beta$. Ainsi :

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i)$$

Les équations de vraisemblance du choix de la fonction lien canonique s'écrivent pour $j \in [\![0\,;p]\!]$:

$$\sum_{i=1}^{n} \omega_i (y_i - \hat{\mu}_i) x_{ji} = 0$$

Avec $\varepsilon_i = y_i - \hat{\mu}_i$; représentant les résidus empiriques

1.2 - Construction d'un modèle linéaire généralisé

Dans cette partie, la construction pratique d'un modèle linéaire généralisé est explicité. La construction peut se répartir en cinq étapes :

- Les statistiques qui permettent d'évaluer l'adéquation du modèle aux données
- Le test des hypothèses sur les coefficients du modèle
- Les intervalles de confiance relatifs aux coefficients du modèle
- L'intervalle de confiance de la moyenne
- L'analyse des différents résidus

1.2.1 - Le choix et l'adéquation du modèle

Le choix de la loi de probabilité de la fonction de réponse découle le plus souvent de la nature du problème étudié.

Il est nécessaire d'introduire les deux statistiques qui permettent d'apprécier l'adéquation du modèle aux données :

- La déviance réduite (aussi nommée Scaled Deviance)
- La statistique du khi-deux de Pearson

1.2.1.1 - La méthode de la déviance

L'objectif est de construire un indicateur qui permet d'évaluer la qualité d'ajustement du modèle estimé aux données observées. Pour ce faire, il convient de comparer le modèle estimé au modèle avec le meilleur ajustement possible, c'est-à-dire le modèle qui contient autant d'observations que de paramètres. Ce modèle est appelé le modèle saturé. Pour comparer les deux modèles, les vraisemblances sont utilisées.

La déviance réduite est définie comme étant :

$$D^* = 2\ln\lambda = 2\ln\left(\frac{\mathcal{L}(b_{max};y)}{\mathcal{L}(b;y)}\right) = 2[\ell(b_{max};y) - \ell(b;y)]$$

avec :

- $\mathcal{L}$; correspond à la vraisemblance.
- $\ell = \ln(\mathcal{L})$; correspond à la log-vraisemblance.
- $D = \phi D^*$; correspond à la déviance non réduite avec ϕ le paramètre de dispersion de la famille exponentielle.

Si le modèle choisi est correct alors la déviance réduite (aussi appelée déviance normalisée) D* suit de manière approximative une loi du khi-deux à n-k degrés de liberté.

Le tableau ci-dessous présente les déviances de quelques lois appartenant à la famille exponentielle :

Loi de probabilité	Déviance D*
Normale	$\sum_{i=1}^{n} \omega_i (y_i - \mu_i)^2$
Poisson	$2\sum_{i=1}^{n} \omega_i \left(y_i \ln\frac{y_i}{\mu_i} - (y_i - \mu_i)\right)$
Gamma	$2\sum_{i=1}^{n} \omega_i \left(-\ln\frac{y_i}{\mu_i} + \frac{y_i - \mu_i}{\mu_i}\right)$
Inverse gaussienne	$\sum_{i=1}^{n} \omega_i \frac{(y_i - \mu_i)^2}{y_i \mu_i^2}$
Binomiale	$2\sum_{i=1}^{n} \omega_i m_i \left(y_i \ln\frac{y_i}{\mu_i} + (1 - y_i)\ln\frac{y_i - \mu_i}{\mu_i}\right)$

Figure 45 : Tableau explicitant la déviance de quelques lois appartenant à la famille exponentielle

1.2.1.2 - La méthode du khi-deux de Pearson

L'idée est de comparer les valeurs ajustées par le modèle aux valeurs observées.

La statistique du khi-deux de Pearson peut s'écrire sous la forme :

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

Il existe également la statistique du khi-deux de Pearson normalisé qui s'écrit :

$$\frac{X^2}{\varphi}$$

De la même façon que pour la déviance réduite, lorsque le modèle est correct cette statistique est également distribuée approximativement selon une loi du khi-deux à n-k degrés de liberté.

1.2.2 - Les tests d'hypothèse sur les coefficients du modèle

Deux tests permettent de vérifier les hypothèses sur les coefficients du modèle, savoir le test de Wald et le test LRT.

1.2.2.1 - Le test de Wald

Sous l'hypothèse $H_0 : L'\beta = 0$ la statistique :

$$S = (L'b)'(L'J^{-1}L)^{-1}(Lb)$$

Suit une loi du khi-deux à k degrés de liberté, où k est le rang de L.

En notant z = $\frac{\hat{\beta}}{ASE}$

Où :

ASE est l'erreur standard asymptotique de $\hat{\beta}$, il en résulte que le test d'un coefficient du modèle se base sur la statistique z^2 distribuée sous l'hypothèse de nullité du coefficient selon une loi de khi-deux à un degré de liberté.

1.2.2.2 - Le test LRT (*Likelihood Ratio Test*)

Notons b* l'estimation du maximum de vraisemblance de β sous l'hypothèse $H_0 : L'\beta = 0$:

$$l(b*, y) = \max_{H0:L'\beta=0} l(\beta, y)$$

Sous l'hypothèse H_0 la statistique :

$$S = 2[l(b, y) - l(b*, y)]$$

suit approximativement une loi de Khi-deux à r degrés de liberté, où r est le rang de L.

1.2.3 - Les intervalles de confiance pour les coefficients du modèle

Il existe deux méthodes qui permettent d'obtenir un intervalle de confiance pour les coefficients , à savoir la méthode de Wald et la méthode du rapport de vraisemblance.

La méthode de Wald produit l'intervalle de confiance suivant :

$$\left[b_j - z_{1-\alpha/2}s_j; b_j + z_{1-\alpha/2}s_j\right]$$

Quant à la méthode du rapport de vraisemblance, elle établit l'intervalle de confiance suivant :

$$\left\{\beta_j: \ell^*(\beta_j, y) \geq \ell(b, y) - \frac{1}{2}\chi^2_{1-\alpha}(1)\right\}$$

1.2.4 - L'intervalle de confiance de la moyenne

La moyenne s'écrivant $\mu_i = g^{-1}(X_i'b)$, l'intervalle de confiance de μ_i s'obtient en utilisant l'intervalle de confiance de $X_i'b$. De plus, la variance de $X_i'b$ est $X_i'J^{-1}X_i$. Il est ainsi possible de déduire l'intervalle de confiance de μ_i d'ordre $1-\alpha$ comme étant

$$g^{-1}\left(X_i'b \pm z_{1-\frac{\alpha}{2}}\sqrt{X_i'J^{-1}X_i}\right)$$

Si l'argument de l'inverse de la fonction lien n'est pas intérieur au domaine valide alors il n'est pas possible de construire un intervalle de confiance.

1.2.5 - Les résidus

Pour l'observation i, le résidu empirique $r_i = y_i - \hat{\mu}_i$ ne permet pas d'obtenir d'information spécifique.

En effet si l'on prend l'exemple d'un modèle de Poisson, l'écart-type d'un effectif est $\sqrt{\hat{\mu}_i}$, des écarts non négligeables tendent à apparaître si μ_i prend une valeur élevée.

Les résidus de Pearson sont des résidus standardisés définis de la façon suivante :

$$r_{P_i} = \frac{y_i - \sqrt{\hat{\mu}_i}}{\sqrt{V(\hat{\mu}_i)}} = \frac{r_i}{\sqrt{V(\hat{\mu}_i)}}$$

Les résidus de déviance s'expriment de la manière suivante :

$$r_{D_i} = \sqrt{d_i}signe(y_i - \hat{\mu}_i) = \sqrt{d_i}signe(r_i)$$

Avec :

- d_i ; représentant la contribution de l'individu i à la déviance D. Il confère une vision individuelle de la linéarité du modèle.
- $D = \sum_{i=1}^{n} d_i$