INSTITUT DES ACTUAIRES

Lyon 1

ISFA GRANDE ÉCOLE D'ACTUARIAT ET DE GESTION DES RISQUES
INSTITUT DE SCIENCE FINANCIÈRE ET D'ASSURANCES

Mémoire présenté le :

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

Par : Gaël Gibaud

Titre Revue des provisions dossier/dossier avec des méthodes de machine learning

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

L. Eckert
B. Pierson
O. Renaudin

Membres présents du jury de l'ISFA

A. Couloumy

Entreprise :

Nom : Actuaris

Signature :

Directeur de mémoire en entreprise :

Nom : Marie Doitteau

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 d'Actuariat

Revue des provisions dossier/dossier avec des méthodes de Machine Learning

Gaël Gibaud

Tuteur Académique : M. Stéphane Loisel
Tuteur en Entreprise : Mme. Marie Doitteau

Promotion 2017

Remerciements

Je souhaite adresser mes remerciements à toutes les personnes qui ont participé à la construction de ce mémoire.

Je tiens tout d'abord à remercier mes tuteurs, Mme Marie Doitteau et M Stéphane Loisel, sans qui toute cette aventure n'aurait pas pu être possible. C'est grâce à leur encadrement et leur suivi que ce projet a pu aboutir.

Je remercie également tous les membres du pôle IARD d'Actuaris qui ont su apporter leur réflexion, leur aide et leur soutien lorsque j'en avais besoin. Je pense plus particulièrement à Khady pour son aide dans le processus de traitement des données, Joachim pour mes questions tardives et Nabil pour les réflexions sur l'application des algorithmes de machine learning.

Je tiens aussi à remercier tous ceux qui m'ont apporté de la réflexion par leurs questionnements et la mise en perspective de mon travail. Merci à Fanny, Kevin, Michaël & Sandrine.

Plus généralement, j'adresse mes remerciements à toutes les personnes d'Actuaris et de l'ISFA qui ont su prendre du temps pour s'intéresser à mes travaux.

Résumé

Mots Clefs : Provisionnement dossier/dossier, gestionnaire sinistre, IARD, Arbre de décision, Réseau de Neurones, Fonction de survie, Kaplan-Meier, Fonction Actuarielle.

La norme prudentielle Solvabilité II en vigueur depuis le 1er janvier 2016 modifie l'approche avec laquelle les assureurs doivent mesurer et gérer leurs risques. La notion de solvabilité est définie au travers de l'exigence réglementaire de fonds propres dit SCR [1] (*Solvency Capital Required*). Celui-ci assure une survie économique de l'entreprise sur l'année à venir avec une probabilité de 99,5%.

Solvabilité II définit aussi quatre fonctions clefs, dont celle qui va nous intéresser ici : la fonction actuarielle. Elle a notamment pour rôle d'encadrer le calcul des provisions et d'établir un rapport donnant un avis et formulant des recommandations sur l'ensemble des points fixés par la réglementation (la souscription, la réassurance, la qualité des données ou encore les provisions). Le responsable de la fonction actuarielle doit entre autre s'exprimer sur le calcul des provisions et la robustesse des modèles utilisés. Pour cela, il doit veiller à ce que le risque soit estimé de la manière la plus juste possible tout en tenant compte des incertitudes pour garantir des provisions adéquates.

L'un des enjeux majeurs d'une compagnie d'assurance est donc de pouvoir honorer ses engagements malgré l'inversion du cycle de production [2] et de faire face alors à l'aléa en anticipant les risques et en les évaluant.

L'estimation et la prévision des sinistres et de leur coût sont ainsi d'une importance cruciale. Ces sinistres, lorsqu'ils surviennent, sont étudiés, dès leur ouverture et tout au long de leur durée de vie, par des gestionnaires de sinistres chargés notamment d'évaluer le plus précisément possible le coût final de chacun des sinistres. Chacun étant amené à suivre les sinistres qui lui sont affectés pour l'entreprise d'assurance.

Ces évaluations sont prises en compte sous forme de provisions, ici les provisions du gestionnaire de sinistres, aussi appelées provision pour sinistres à payer dossier/dossier ($PSAP^{dossier/dossier}$).

Les provisions formant généralement la partie la plus représentative des passifs du bilan comptable d'une compagnie d'assurance, la fiabilité de leur évaluation est donc indispensable pour permettre à l'entreprise de tenir ses engagements.

L'idée ici est d'expérimenter un nouvel outil avec l'objectif ambitieux d'apporter, au responsable de la fonction actuarielle, un éclairage sur le provisionnement dossier/dossier. Le but est également de voir dans quelles situations l'outil créé sera performant ainsi que les situations dans lequel il ne le sera pas.

Chaque entreprise possède un processus d'évaluation de $PSAP^{dossier/dossier}$ qui lui est propre. Celui-ci peut aussi dépendre de l'expérience du gestionnaire de sinistres et des experts (avocats, experts sinistres). Les méthodes mises en place ici auront pour but de s'extraire de cette évaluation en ne se basant que sur le portefeuille de l'entreprise.

Les différentes méthodes mises en place ici sont issues du domaine de la *Data Science* et plus précisément du *Machine Learning*.

1. Le calcul du SCR est fondé sur une approche basée sur les risques
2. Pour fixer les primes, les assureurs doivent évaluer les risques, leurs coûts et leurs probabilités avant que ceux-ci n'arrivent

Nous partirons de la méthode des arbres de décision en utilisant les arbres de classification et de régression (*CART*), nous construirons des forêts aléatoires (*Random Forest*) et finirons avec l'implémentation du Gradient Boosting Machine (*GBM*). Nous mettrons aussi en avant les réseaux de neurones.

Ces méthodes commencent à émerger en actuariat (notamment les arbres de décisions et les forêts aléatoires) et sont principalement utilisées en tarification incendie, accident et risques divers (*IARD*) dans le but de créer des tarifs plus précis que ceux obtenus par les modèles linéaires généralisés (*Generalized Linear Models, GLM*). Nous allons tenter de déterminer ici si l'utilisation de ces techniques pourraient avoir un intérêt probant en provisionnement IARD.

Nous évaluerons les charges ultimes des sinistres clos afin de voir quelle marge de progression est possible dans l'appréciation des provisions. Ensuite nous appliquerons nos modèles sur les sinistres ouverts afin d'anticiper la charge de chaque sinistre en cours et comparerons les résultats obtenus par nos algorithmes aux $PSAP^{dossier/dossier}$ apposées à ces sinistres par le gestionnaire. Cette étude sera faite à deux étapes de la vie du sinistre : à son ouverture et à sa dernière évolution.

L'intérêt recherché étant d'aider la fonction actuarielle à se faire sa propre opinion sur le provisionnement dossier/dossier pour qu'elle puisse ensuite émettre des recommandations destinées à améliorer ou à parfaire la politique de provisionnement et les procédures internes d'évaluation des dossiers sinistres.

Abstract

Keywords : P&C, Individual Claim Reserving, Claim Manager, Decision Tree, Neural Network, Survival Function.

The Solvency II norm, applied as of the beginning of january 2016, changes the way in which insurers have to mesure and manage risks. The idea of solvency is defined through a risk based approach with the Solvency Capital Required (SCR). This capital allows for the economic survival of the company with a probability of 0.995 for the ongoing year.

This norm also sets up four key functions such as the one we are specificaly interested in : the actuarial function. Its role is to regulate the calculation of the reserves and hand in a report containing an opinion and recommandations on all the different points set by the reglementation. The person responsible of the actuarial function has to express himself on the calculation of the reserves but also on the robustness of the models that are used. He must thus check that the risk is estimated as precisely as possible while taking in account incertainty to allow for correct reserves.

Indeed, one of the major issues at stake for an insurance company is to be able meet its commitments despite the inversion of the production cycle [3] and to face the risk while correctly evaluating it.

Predicting and estimating claims and their cost is thus crucial. Once these claims are reported, they are studied throughout their entire lifespan by claim managers whose roles are to evaluate as precisely as possible the cost of each claim. Each one of them following the claims handed to him by the insurance company.

These estimates are taken into account under reserves, the claim manager reserves, also called the claim by claim reserves for claims oustanding.

As these reserves represent the largest part of the liability balance sheet of an insurance company, the reliability of their estimates is of the upmost importance to allow for the company to uphold its commitments.

The idea here is to create a new tool with the ambitious goal of bringing to the actuarial function manager an insight on the claim by claim reserving. We also want to be able to evaluate when the tool will give us good or bad performances.

Indeed, each company had its own reserving and evaluation process for the claim by claim approach. It can depend on the experience of the claim manager or the different experts involved. The methods we will be setting up will not involve these process' as they will only be based on the company's portfolio.

The different algorithms that we are going to use come from the *Machine Learning* part of *Data Science*.

We will first of be implementing decision trees *CART*, then build *Random Forest* and set up Gradient Boosting Machine. We will also be using artificial neural networks.

3. insurers have to price the risk before it happens, it is thus an estimate of the cost and not the actual cost of the risk that the premiums cover

These methods are starting to be used in actuarial science. They are especially used in P&C pricing in order to obtain more precise premiums than the ones given by the *Generalized Linear Models, GLM*. We will try to see if the use of these methods could be interesting in P&C reserving.

We will be evaluating the final cost of closed claims in order to see what progression margin is possible when calculating reserves. Then, the models will be used on the open claims (the ones that are not yet settled) in order to anticipate their final cost and compare the results with the ones given by the claim manager. This study will be made when the claim is reported but also at its last financial mouvement.

Our motivation is to help the actuarial function to make its mind up regarding the claim by claim reserves so that it can then give recommandations in order to improve the reserving policies and the internal procedures.

Sommaire

Remerciements iii

Résumé v

Abstract viii

Sommaire xii

Introduction 1

A Contexte 3

I La Fonction Actuarielle 5

II Les provisions techniques 9
II.1 Le gestionnaire sinistre 13
II.2 Challenge de la provision dossier/dossier 15

B Présentation des méthodes 19

I L'arbre de décision CART 20
I.1 Construction d'un arbre maximal 20
I.2 Récapitulatif 22
I.3 Exemple de construction d'arbre maximal 23
I.4 Elagage 25
I.4.1 Evaluation de l'erreur R(t) d'un arbre 26
I.5 Exemple d'élagage 26
I.6 Avantages et Inconvénients 28

II Les méthodes d'aggrégation 29
II.1 Le Bootstrap 29
II.1.1 Avantages et inconvénients 30
II.2 Les familles de modèles aléatoires 30
II.2.1 Le Bagging 31
II.2.2 Forêt Aléatoire 33
II.2.3 Avantages et Inconvénients 36
II.3 Famille de modèles adaptatifs 36
II.3.1 Gradient Boosting Machine 36
II.3.2 Version Aléatoire 37
II.3.3 Cas de la régression 37
II.3.4 Avantages et Inconvénients 38

III Les réseaux de neurones 39
III.1 Un réseau neuronal 39
III.2 Le réseau formel 39
III.3 Perceptron Multicouche (PMC) 41
III.3.1 Architecture 41
III.4 Apprentissage du réseau 43

III.5 Rétro-propagation de l'erreur . . . 43
III.6 Algorithme d'optimisation . . . 44
III.7 Paramètres et complexité . . . 45
III.8 Avantages et Inconvénients . . . 45

IV Le compromis biais-variance **47**

V Prise en compte de la censure **49**
V.1 L'estimateur de Kaplan-Meier . . . 49
V.2 Observations Censurées . . . 50
V.3 Exemple . . . 51

C Etudes et résultats **55**

I Description de la base **57**
I.1 Prise en compte de l'inflation . . . 58
I.2 Sinistralité du portefeuille . . . 59

II Etude à l'ouverture du dossier **63**
II.1 Données à l'ouverture et sélection de variables . . . 63
II.2 Postulats . . . 67
II.3 Modèles . . . 68
II.4 Résultats sur les sinistres ouverts . . . 79
II.5 Conclusion . . . 81

III Etude à date d'extraction **82**
III.1 Données à la date d'extraction . . . 83
III.2 Postulats . . . 84
III.3 Modèles . . . 84
III.4 Résultats sur les sinistres ouverts . . . 90
III.5 Conclusion . . . 94

Conclusion **97**

Liste des tables **103**

Liste des figures **105**

Références **106**

Introduction

La mutualisation du risque issue de l'émergence de l'assurance permet à l'assuré, moyennant le paiement d'une prime, de se protéger financièrement contre ce risque qu'il ne pourrait supporter seul. Dans le cadre de ce projet, nous tacherons d'évaluer le coût final des sinistres ouverts d'un portefeuille dont le risque étudié est celui de l'incendie domestique.

Le cycle inversé de l'assurance oblige les assureurs à anticiper du mieux possible le coût des sinistres futurs afin d'honorer leurs engagements, rester solvables et compétitifs. Leur but est donc d'anticiper au mieux la sinistralité future. Cette prévision de coût se traduit sous forme de provision. Cette provision, due par l'assureur, est le montant estimé du coût de tous les sinistres, ceux qui sont survenus et ceux qui ne sont pas encore survenus mais rattachés à des contrats en cours.

La provision pour sinistres à payer représente le coût estimé à la fin de l'année afin que l'assureur puisse se prémunir des pertes des sinistres déjà survenus, mais non encore réglés.

Les pertes liées à une année de survenance ne sont pas nécessairement déclarées lors de celle-ci. En effet, les paiements de certains sinistres peuvent dépasser l'année comptable et s'étaler sur deux ou trois ans, voire beaucoup plus dans certaines situations. C'est le cas lorsqu'il est nécessaire de faire appel à des experts (pour des actions en justice, pour déterminer les responsabilités par exemple). Le règlement du sinistre se fait donc des années après sa déclaration à l'assureur. Un autre exemple pourrait être lorsque l'assureur doit prendre en charge les soins et l'accompagnement d'une personne sinistrée pendant une période longue ou jusqu'au terme de sa vie (assistance d'une tierce personne au quotidien).

Anticiper le coût de cette sinistralité future est un pré-requis majeur pour l'assureur car les montants mis en jeu sont généralement conséquents. Le but recherché par l'assureur est donc de calculer ses provisions le plus précisément possible tout en tenant compte des incertitudes liées à l'estimation.

Dans ce projet nous nous attacherons à créer des modèles susceptibles d'aider la fonction actuarielle dans ses travaux, l'objectif ambitieux étant d'estimer le coût final des sinistres en cours déjà déclarés avec des méthodes issues du *Machine Learning*. L'estimation de ce coût sera faite à l'ouverture du dossier ainsi qu'au dernier mouvement fait.

La fonction actuarielle ayant la responsabilité d'encadrer la bonne gestion des provisions, l'idée ici est de mettre en place des modèles qui pourraient contribuer à challenger les provisions faites par les gestionnaires de sinistres.

Première partie
Contexte

Avec la mise en place de la directive Solvabilité II le 1er janvier 2016, le secteur de l'assurance a du revoir sa manière d'évaluer et de gérer les risques. Ce nouveau régime européen a pour but d'attendre une certaine consistance dans la gestion du risque et de la gestion du capital. Nous allons revenir très succinctement sur ses grands principes.

Le cadre imposé par Solvabilité II est constitué de 3 piliers :

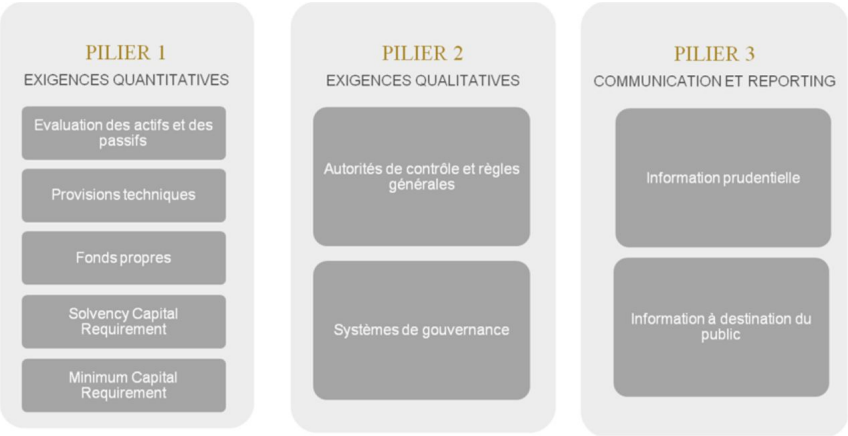

FIGURE 1 – Les 3 piliers de Solvabilité 2

Pilier 1 : Exigence quantitative

Le pilier 1 considère les pré-requis quantitatifs du système, en incluant le calcul des provisions techniques, les règles liées au calcul du capital de solvabilité et de la gestion des investissements.

Ce pilier expose la valuation standard du passif ainsi que le seuil de capital requis à atteindre.

Deux seuils de solvabilité sont requis :

— Le SCR (*Solvency Capital Required*), capital de solvabilité requis
— le MCR (*Minimum Capital Required*), capital minimum requis

Ces deux seuils sont des indicateurs de la santé financière d'une entreprise.

Le SCR peut être calculé via une formule standard déterminée par le régulateur ou l'entreprise peut développer son propre modèle interne pour refléter les risques spécifiques qui lui font face. Si cette seconde approche est adoptée elle doit être approuvée par l'autorité de supervision.

Pilier 2 : Exigence qualitative

Le pilier 2 encadre les aspects qualitatifs des contrôles internes de l'entreprise, de ses risques, de ses processus de management ainsi que le processus de production de rapports à l'autorité de supervision.

Le pilier 2 inclus aussi l'ORSA (*Own Risk ans Solvency Assessment*), processus selon lequel l'entreprise cartographie ses risques et les évalue.

Quel que soit le choix fait par l'entreprise pour le pilier 1 (formule standard ou modèle interne), celle-ci doit produire un rapport ORSA.

Si l'autorité de supervision n'est pas satisfaite avec l'évaluation de la partie risquée du capital ou avec la qualité de la gestion du risque, celle-ci peut imposer un capital requis supérieur.

Le Pilier 3 : Rapports et communication

Le pilier 3 a pour but d'augmenter la transparence des entreprises et donc du marché.

Les entreprises doivent interpréter les exigences de transparence, développer des stratégies de transparence et éduquer les parties prenantes.

Les entreprises auront la responsabilité d'organiser l'information, au travers d'une transparence publique, pour que celle-ci soit accessible aux régulateurs, aux analystes, aux agences de notations et aux investisseurs.

De plus, chaque organisation devra développer un processus interne ainsi que des systèmes de production de ces rapports.

Un point particulier de solvabilité II qui va nous intéresser est spécifiquement celui qui définit la création de 4 fonctions clefs :

— article 44 : la fonction de gestion des risques
— article 46 : la fonction de conformité
— article 47 : la fonction d'audit interne
— article 48 : la fonction actuarielle

Ces quatre fonctions porteuses de responsabilités importantes étaient auparavant entre les mains de la Direction Générale et doivent maintenant être mises en place. L'organisme d'assurance peut donc construire un système de gestion des risques plus adapté aux situations auxquelles il fait face.

La dernière fonction, la fonction actuarielle, est celle qui nous intéresse.

En effet, elle fait désormais partie de la structure de tout organisme d'assurance et occupe un poste clef car elle participe aux travaux techniques de la société et doit détecter les risques inhérents aux activités de la société. Son rôle est détaillé davantage dans la partie suivante.

I La Fonction Actuarielle

Le rôle de la fonction actuarielle est définie comme suit dans l'article 48 de la directive de Solvabilité II, elle doit :

1. coordonner le calcul des provisions techniques
2. garantir le caractère approprié des méthodologies, des modèles sous-jacents et des hypothèses utilisés pour le calcul des provisions techniques
3. apprécier la suffisance et la qualité des données utilisées dans le calcul des provisions techniques
4. comparer les meilleures estimations aux observations empiriques
5. informer l'organe d'administration, de gestion ou de contrôle de la fiabilité et du caractère adéquat du calcul des provisions techniques
6. superviser le calcul des provisions techniques dans les cas visés à l'article 82 [4]
7. émettre un avis sur la politique globale de souscription
8. émettre un avis sur l'adéquation des dispositions prises en matière de réassurance
9. contribuer à la mise en œuvre effective du système de gestion des risques visés à l'article 44, en particulier pour ce qui concerne la modélisation des risques sous-tendant le calcul des exigences de capital prévu au chapitre VI, sections 4 et 5, et pour ce qui concerne l'évaluation visée à l'article 45 [5]

Il est ensuite ajouté que les personnes faisant partie de la fonction actuarielle doivent comprendre intrinsèquement le risque étudié :

"*La fonction actuarielle est exercée par des personnes qui ont une connaissance des mathématiques actuarielles et financières à la mesure de la nature, de l'ampleur et de la complexité des risques inhérents à l'activité de l'entreprise d'assurance ou de réassurance et qui peuvent démontrer une expérience pertinente à la lumière des normes professionnelles et autres normes applicables.*"

La fonction actuarielle doit, comme le précise l'article 272 du règlement délégué, produire un rapport qui "*rend compte de tous les travaux conduits par la fonction actuarielle et de leurs résultats, il indique clairement toute défaillance et il émet des recommandations pour y remédier.*"

Ce rapport actuariel doit être produit au moins une fois par an et doit être validé par l'AMSB (*administrative, management or supervisory body*, c'est-à-dire le conseil d'administration) et tenu a disposition de celui-ci.

Ce rapport doit suivre quelques principes clefs :

— La transparence, avec la synthèse des hypothèses retenues et utilisées
— La complétude, qui prend en compte l'intégration de toutes les déficiences matérielles rencontrées, la cohérence entre les différentes parties du rapport et les différents rapports du Groupe le cas échéant et la possibilité de définir un seuil de matérialité
— L'adaptation du contenu du rapport à l'AMSB, mentionnant les définitions, les informations clés et toute information nécessaire à la compréhension

4. L'article 82 reprend les taux des niveaux de capital et leurs limites quantitatives
5. L'article 45 définit l'ORSA, c'est l'ensemble des processus qui constituent l'outils d'analyse décisionnelle et stratégiques qui évaluent en continu les besoins de solvabilité de l'entreprise

— La pertinence avec un niveau d'information suffisant pour que l'AMSB puisse juger la pertinence de l'opinion et l'adaptation des différentes dates d'évaluation aux données
— Une communication active avec l'AMSB pour l'identification des conflits d'intérêts relatifs à la fonction actuarielle, la nécessité de formation de l'AMSB aux différentes tâches de la fonction actuarielle ainsi que la prise en considération des attentes et retours de l'AMSB sur le contenu du rapport

La fonction actuarielle doit couvrir quatre sujets principaux : les provisions techniques, la politique de souscription, la politique de réassurance et la gestion des risques.

Les provisions techniques : En ce qui concerne les provisions techniques, ce rapport devrait couvrir les éléments suivants :

— Les processus encadrant l'estimation des provisions techniques
— La suffisance et qualité des données
— Les méthodes et modèles retenus
— Les hypothèses retenues et leurs justifications
— L'analyse de l'évolution des Best Estimates par rapport à l'expérience
— L'analyse de sensibilité aux facteurs de risque

Les provisions techniques doivent couvrir les champs que sont à la fois le Best Estimate et la marge pour risque.

La qualité des données est une information centrale, une synthèse doit donc être faite sur cette partie là. De plus, l'ACPR a mentionné à différentes compagnies que son intégration à la gouvernance d'ensemble était généralement trop faible. De manière générale, la sensibilisation de l'AMSB à ce sujet était insuffisante.

Dans le cas où les données sont insuffisantes pour utiliser des méthodes standards, il faut clairement identifier la situation et justifier l'utilisation de données marchés ainsi que de modèles non standards.

Tous les facteurs pouvant avoir un impact important sur les provisions techniques doivent être évalués ainsi les comparaisons de résultats avec l'expérience renforcées afin de donner à l'AMSB une vision globale du risque sur les provisions techniques.

Ce document doit aussi couvrir la politique de souscription, la politique de réassurance ainsi que la gestion des risques mais ces points nous intéressent moins.

La politique de souscription : Le rapport actuariel doit faire état de l'opinion du responsable de la fonction actuarielle sur la politique de souscription en développant certains points.

Tout d'abord, synthétiser les points essentiels de la politique de souscription, évaluer les responsabilités clefs et faire un point sur la gestion des conflits d'intérêts. Il doit ensuite faire état de la suffisance des primes au regard des futures prestations et frais tout en considérant les risques sous-jacents. La pertinence de la segmentation tarifaire doit aussi être mentionnée ainsi que le niveau de granularité qui a été retenue et ce en validation avec la qualité des données justifiant les paramètres et hypothèses de tarification.

Les autres points suivants doivent être au centre des considérations :

— Identifier et décrire les facteurs environnementaux externes qui peuvent avoir un impact sur la rentabilité de la société
— Quantifier l'impact des évolutions potentielles de ces facteurs avec une étude de sensibilité du portefeuille en faisant des considérations relatives à l'inflation, au risque juridique, au changement dans la composition du portefeuille, à l'effet des systèmes bonus-malus ou des systèmes semblables, aux risques affectant la souscription future
— Considérer l'évolution des primes ainsi que le contexte des décisions afin d'émettre un avis sur la future tarification : la modifier ou la maintenir constante
— L'analyse des mesures tarifaires passées et de leur pertinence et suffisance : étude de backtesting
— Analyser la tendance d'un portefeuille à attirer ou conserver des assurés avec des profils plus risqués. Ceci se fait par l'analyse des risques d'anti-sélection et par la manière dont ceux-ci sont gérés et devraient être gérés
— La cohérence de la politique de souscription avec les autres politiques actuarielles comme la politique de provisionnement ou la politique de réassurance.

Ce rapport apporte un contexte nouveau qui demande de normer les travaux à mener lorsque de nouveaux produits seront commercialisés et de plus d'exiger que les hypothèses du business plan soient justifiées. Il faut y ajouter les informations sur des sujets qui jusqu'à présent sont restés confidentiels, notamment des problématiques opérationnelles et stratégiques. Finalement il demande d'analyser la pertinence de la politique de souscription.

La politique de réassurance : La fonction actuarielle doit aussi émettre un avis sur la politique de réassurance de l'organisme.

Cette partie du rapport peut inclure la synthèse de :

— Le programme de réassurance choisi
— La procédure qui a permis de choisir le programme de réassurance ainsi que les difficultés soulevées par le choix de celui-ci
— La procédure de suivi de gestion du programme de réassurance
— Les différents impacts du choix de ce programme sur le bilan comptable et la solvabilité de l'organisme
— L'intérêt du programme de réassurance choisi justifié par l'efficacité de celui-ci dans des scenarii de *stress-test* et ce en lien avec la politique de souscription
— La cohérence du programme de réassurance avec les provisions techniques
— La cohérence entre le profil de risque et la politique de souscription choisie

Le rapport de la fonction actuarielle donne une opinion sur la politique de réassurance mais celle-ci peut différer entre différentes compagnies d'assurance, en fonction de leur taille, de leur complexité, de leurs implantations et du type de réassurance. Il faut donc être en mesure d'expliquer les changements de couverture s'il y en a eu, prendre en compte l'appétence au risque et considérer le risque de crédit.

La gestion des risques : Cette section doit viser à synthétiser comment la fonction actuarielle contribue au Risk Management System. Le but est de développer une collaboration étroite entre les deux

fonctions que sont la fonction actuarielle et la fonction gestion des risques même si une opinion n'est pas explicitement demandée dans la directive.

Les responsabilités de la fonction gestion des risques et les relations établies entre les deux fonctions (que sont la fonction actuarielle et la fonction gestion des risques) déterminent directement l'apport que peut avoir la fonction actuarielle pour la gestion des risques.

Celui-ci peut inclure certains éléments comme :

— Le calcul des mesures de risque et du besoin en capital
— Le calibrage du modèle de risque
— Une contribution au processus de production de l'ORSA
— L'aide à l'analyse du profil de risque
— La définition de l'appétence aux risques et ses limites
— L'attribution des profits et des pertes
— Rédiger les différentes politiques de l'organisme
— Une contribution aux rapports Solvabilité 2
— La gestion de l'actif et du passif

Nous comprenons donc que la fonction actuarielle doit maîtriser le risque, les données et les modèles qui entourent les provisions techniques, la souscription, la réassurance et intervient en tant que partenaire ou support de la fonction de gestion des risques.

Le point qui nous intéresse ici est celui des provisions techniques. Il convient donc de définir le périmètre de ces dites provisions techniques.

II Les provisions techniques

Les provisions constituent un poste du bilan du comptable, au passif, situées en dessous des capitaux propres. Elles correspondent à des charges probables qu'une entreprise aura à supporter à l'avenir et dont le montant n'est qu'estimable, pas connu de manière définitive. La représentation du bilan comptable sous Solvabilité II est le suivant [6] :

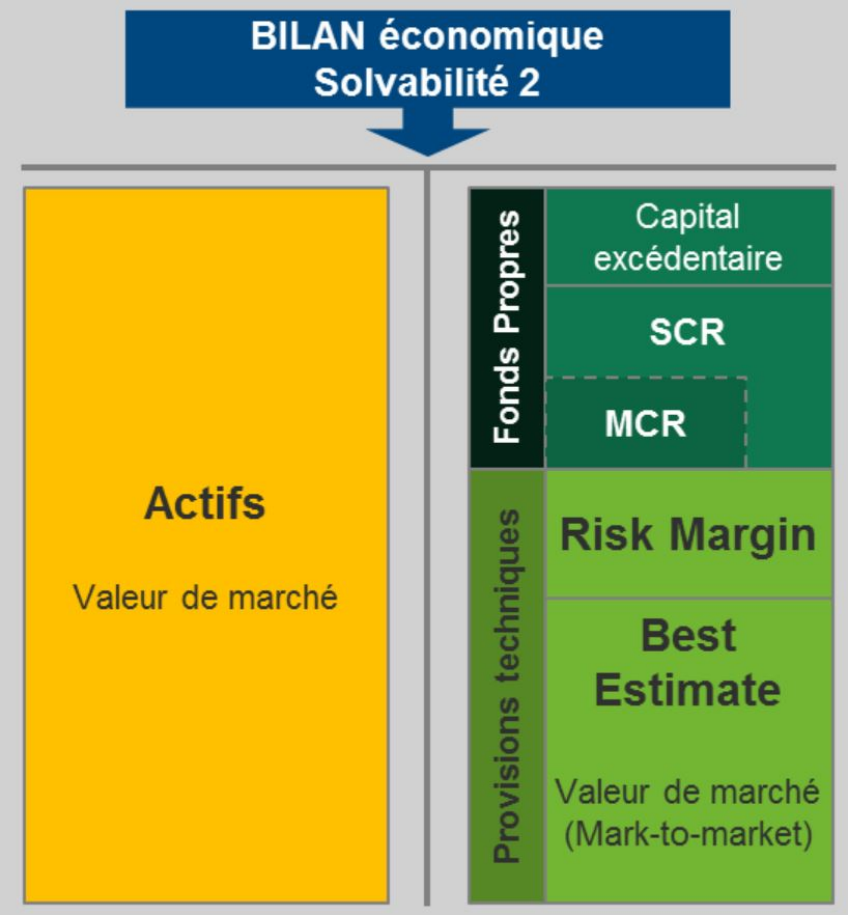

FIGURE 2 – Bilan sous Solvabilité 2

Le qualificatif technique sert à faire la distinction avec toutes les autres provisions. Il s'agit de toutes les provisions qui sont évaluées par les entreprises soit d'assurance ou de réassurance.

Dans une compagnie d'assurance, la notion de couverture des passifs par les actifs est imposée par le principe des engagements réglementés qui disposent que les provisions techniques doivent toujours être représentées par des actifs équivalents.

Afin de garantir les pertes liées aux assurés, l'organisme d'assurance doit constituer des réserves, les provisions techniques. Celles-ci représentent un pourcentage des primes reçues par l'assureur. Ces provisions ont pour but de permettre de régler les sinistres. Elles représentent une dette vis-à-vis de l'assuré. L'assureur pourrait avoir tendance à diminuer le montant de ses réserves et ce pour dégager du résultat, ne se préoccupant des sinistres qu'une fois qu'ils se produisent. Mettre une telle politique en place

6. https://www.insurancespeaker-wavestone.com/wp-content/uploads/2014/02/Solva2-Bilan.png

serait nocive pour la pérennité de la société ainsi que pour les intérêts des assurés. Il faut donc, dans l'intérêt des assurés, que l'organisme d'assurance ait la capacité à faire face à ses engagements à tout instant.

Le but des provisions techniques est de faire face à la sinistralité future pour des risques où la prime de l'assuré à déjà été encaissée. Le législateur demande à ce qu'à tout instant, les engagements soient représentés par des actifs équivalents.

Ce compte du passif est le plus important, pouvant en représenter jusqu'à 80%. C'est le plus spécifique du domaine de l'assurance car c'est le plus délicat à établir. Il représente la clef de voute du bilan car une variation des provisions impose une variation des actifs supposés les compenser. Les sommes mises en jeu étant généralement importantes, leurs évaluations doivent être les plus fines possible.

Ce montant de provisions techniques est calculé en fonction de la sinistralité future de l'entreprise pour que celle-ci soit en mesure de la régler.

Ces engagements font face à des placements mobiliers et immobiliers, il convient donc de comprendre que l'évaluation des provisions techniques n'est pas qu'une fonction technique, c'est aussi une fonction financière qui joue un rôle dans la vie de l'entreprise, que se soit dans sont équilibre ou sa croissance.

De par l'essence de son rôle, l'assureur détient les sommes qui lui sont remises par ses assurés. Ces sommes sont contractuelles, liées à la durée de vie du contrat mais aussi au temps du règlement des sinistres.

Pendant cette période, l'assureur se doit de justifier, soit :

— De l'existence de ces sommes en attente de la survenance des sinistres ou bien de leurs règlements
— De leur utilisation pour des prestations liées contractuellement aux bénéficiaires des contrats

Cette durée peut se révéler longue et ce à cause de certains sinistres dont les causes ou les responsables peuvent être difficiles à trouver ou simplement car le règlement va prendre du temps.

L'assureur se doit de conserver pendant ce temps les sommes qui lui ont été versées et de les gérer. Elles constituent une épargne créée par les primes perçues et versées par les assurés. Ces sommes, les provisions techniques doivent être distinctes de toute autre réserve, comme celle servant à rémunérer les actionnaires par exemple.

Ces sommes devant régler de manière intégrale les engagements de l'assureur sont donc de ce fait particulièrement surveillées par les autorités de contrôle.

Elles sont tout d'abord calculées brutes de réassurance, c'est-à-dire sans considération de la partie cédée du risque.

Ces dettes de l'assureur proviennent de la naissance des sinistres, qu'ils soient en cours de règlement ou même inconnus.

Cette évaluation des provisions techniques est destinée à la fin de l'exercice en cours et doit prendre en compte toute la sinistralité survenue liée à cette année là.

Avoir des provisions techniques correctes permet d'être à la fois au courant de la situation financière de la société ainsi que le coût réel du risque assuré.

Les provisions techniques sont la garantie que l'assureur tiendra ses engagements vis-à-vis des sinistres que subiront les assurés et des prestations qu'il aura à faire.

De part ce rôle majeur, elles font l'objet d'un contrôle rigoureux fait par l'autorité de contrôle que ce soit lors de leur constitution mais aussi de leur gestion.

Il existe un certain nombre de provisions techniques et elles sont divisées en deux catégories, générées par les deux grandes branches d'assurance :

— Les provisions techniques vie
— Les provisions techniques non vie

Provisions techniques	
Vie	Non Vie
Provision mathématique Provision pour aléas financier Provision pour frais de gestion Provision pour participation aux bénéfices Provision d'égalisation Provision pour frais d'acquisition reportés	Provision pour sinistres à payer Provision pour primes non acquises Provision pour risque en cours Provision pour risque croissant Provision d'égalisation
Provision d'exigibilité Réserve de capitalisation	

— Provision pour risque d'exigibilité : il s'agit de la différence, calculée pour les placements, entre le montant global de la valeur de marché et la valeur comptable nette des placements concernés quand cette différence est négative.
— Réserve de capitalisation : il s'agit des plus values réalisées lors de la cession d'obligations

Les provisions techniques vie

— Provisions mathématiques : il s'agit de la différence à la date d'inventaire entre les valeurs actuelles des engagements respectivement pris par l'assureur et les assurés. Les notes aux états financiers doivent fournir le détail des provisions mathématiques vie.
— Provision pour aléas financier : il s'agit de la compensation de la baisse de rendement de l'actif
— Provision de gestion : il s'agit d'une provision constituée pour couvrir les charges de gestion futures des contrats qui ne sont pas couvertes par ailleurs. Cette provision vise à couvrir les charges de gestion engendrées par les contrats en portefeuille dès lors qu'elles ne sont pas couvertes par des revenus futurs.
— Provision pour participation aux bénéfices : il s'agit d'une provision constituée pour enregistrer l'engagement de l'entreprise d'assurance envers les bénéficiaires des contrats lorsque les montants dus au titre des bénéfices n'ont pas encore été versés ou crédités au compte de l'assuré.
— Provision d'égalisation : il s'agit des montants provisionnés pour permettre d'égaliser les fluctuations des taux de sinistres pour les années à venir dans le cadre des opérations d'assurance de groupe contre le risque décès.
— Provision pour frais d'acquisition reportés : il s'agit de couvrir la charge résultant du report des frais d'acquisition

Les provisions techniques non vie

- Provision pour sinistres à payer : il s'agit de l'évaluation du montant qui sera versé postérieurement à la clôture de l'exercice au titre d'événements qui se sont réalisés antérieurement à la clôture de l'exercice. Les provisions comportent trois types de sinistres restant à payer :
  - Les sinistres dont l'évaluation est définitive, connue et pour lesquels il ne demeure que le mouvement de trésorerie à générer
  - Les sinistres pour lesquels l'évaluation n'est pas définitive et ayant fait ou non l'objet de règlements partiels,
  - Les sinistres survenus antérieurement à la clôture mais dont la survenance n'a pas été portée, à cette date, à la connaissance de l'entreprise. Il s'agit des sinistres tardifs.
- Provisions pour primes non acquises : il s'agit de la provision destinée à constater, pour l'ensemble des contrats en cours, la part des primes émises et des primes restant à émettre se rapportant à la période comprise entre la date d'inventaire et la date de la prochaine échéance de prime ou, à défaut, du terme du contrat.
- Provision pour risques en cours : il s'agit du montant à provisionner en supplément des primes non acquises pour couvrir les risques à assumer et destinée à faire face à toutes les demandes d'indemnisation et à tous les frais (y compris les frais d'administration) liés aux contrats d'assurance en cours excédant le montant des primes non acquises et des primes exigibles relatives aux dits contrats.
- Provision pour risque croissant : il s'agit de la différence des valeurs actuelles des engagements respectifs pris par l'assureur et les assurés.
- Provision d'égalisation (ou d'équilibrage) : il s'agit des montants provisionnés conformément aux dispositions légales et réglementaires permettant d'égaliser les fluctuations des taux de sinistres pour les années à venir ou de couvrir des risques spéciaux.

Celle qui nous intéresse ici est la provision pour sinistres à payer (PSAP).

Cette provision pour sinistres à payer regroupe plusieurs provisions :

- La provision dossier/dossier : évaluation pour chaque sinistre par un gestionnaire des règlements à venir en fonctions des informations disponibles. Celle-ci se doit d'être réévaluée à chaque fois qu'une nouvelle information sur le sinistre est disponible.
- Les IBNR (Incurred But Not Reported), survenu mais non déclarés. Ils sont composés des :
  - IBNyR (Incurred But Not yet Reported) : provision concernant des sinistres survenus mais non déclarés, aussi appelés tardifs. Elle est évaluée à chaque fin d'inventaire par segment
  - IBNeR (Incurred But Not enough Reported) : provision concernant des sinistres dont le coût est réévalué à la hausse. Les sinistres ont été insuffisamment provisionnés. Ce sont souvent des complétements au provision dossier/dossier qui peuvent être négatives en cas de surprovisionnement. Elle est évaluée à chaque fin d'inventaire par segment
- La Provision pour Frais de Gestion de Sinistres (PFGS) : elle doit couvrir les frais liés aux sinistres survenus
- La Provision pour Recours à Encaisser (PRAE) : c'est une prévision des recours futurs

Nous pouvons donc écrire la relation suivante :

$$PSAP = Provision^{dossier/dossier} + IBNR + PFGS - PRAE$$

L'élément qui va nous intéresser ici, c'est la $Provision^{dossier/dossier}$ et le rôle du gestionnaire des sinistres dans l'évaluation de cette provision.

II.1 Le gestionnaire sinistre

Le gestionnaire de sinistres intervient après la survenance d'un sinistre et lorsque le client transmet une déclaration de la réalisation de ce sinistre afin d'être indemnisé dans le cadre de son contrat.

Son rôle est majeur car en plus de mettre en œuvre les différentes garanties prévues par le contrat du client il doit identifier les circonstances du sinistre afin d'évaluer les responsabilités des personnes concernées par celui-ci. Son but est de comprendre au mieux le sinistre afin que personne ne soit lésé lorsque les dédommagements seront faits par l'organisme d'assurance pour lequel il exerce son métier.

Au départ généraliste, la pratique dans les organismes d'assurance veut que le gestionnaire tende à se spécialiser sur une catégorie de sinistres précise (la responsabilité civile, les sinistres automobiles, etc.). Cette profession est aujourd'hui subdivisée en différentes branches qui vont dépendre de la clientèle souscriptrice du contrat : les particuliers, les professionnels ou les entreprises par exemple, ou alors du risque (agricole, industriel, habitation, etc.).

Le travail du gestionnaire sinistre est donc impacté par le risque qu'il encadre. C'est lui l'interlocuteur privilégié des sinistrés. En effet, c'est son rôle de les informer sur la progression de l'instruction du dossier sinistre, de les orienter à toutes les étapes du dossier. Il s'occupe aussi de déterminer les responsabilités et le montant des dommages subis par les personnes impliquées dans le sinistre (qu'elles soient ses assurés ou non).

Le gestionnaire de sinistres doit prendre en charge l'instruction des dossiers, s'occuper des règlements amiables dans le cadre de déclarations de sinistres liés à des contentieux ou étant complexes. Il doit faire preuve de rigueur et d'analyse pour déterminer les responsabilités et mener les négociations dans le respect des procédures.

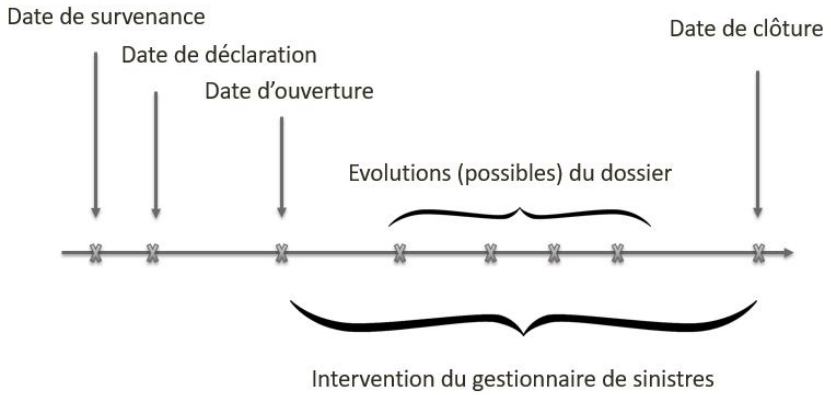

FIGURE 3 – Evolution de la vie d'un sinistre

Les évolutions ici peuvent être de nouvelles informations sur le sinistre : un expert qui livre son avis, une clôture suivi d'une réouverture ou différents paiements par exemple.

Ses principales tâches sont les suivantes :

— Enregistrer le sinistre : c'est lui qui va être contacté lors de la déclaration d'un sinistre. Il recueille toutes les informations qui lui sont nécessaires afin de constituer un dossier propre au sinistre

— Évaluer les responsabilités : avec les éléments et les témoignages qu'il a recueilli il va juger du degré de responsabilité des personnes concernés, notamment celle de son client
— Déterminer le coût du préjudice : après avoir pris connaissance des éléments du sinistre il doit chiffrer le montant du préjudice et se mettre en lien avec le responsable de l'indemnisation
— Statuer sur les dossiers : tâche importante et lourde. Il doit vérifier que le sinistre rentre bien dans le cadre du contrat. C'est lui qui va rendre un avis sur le versement ou non des montants qu'il a déterminé.
— Négocier avec les autres compagnies d'assurance : ceci peut arriver si un tiers est impliqué dans un sinistre. Le gestionnaire va donc devoir traiter avec l'organisme assureur. Il peut soit négocier une gestion du dossier à l'amiable soit défendre les intérêts de son entreprise et tenter de mettre en place la meilleure solution possible

De manière plus synthétique, le gestionnaire de sinistres est au cœur du processus d'indemnisation des sinistres, il doit vérifier les applications du contrat, déterminer les responsabilités et les indemnités.

Il lui faut donc être rigoureux, organisé, être capable d'arbitrer des situations complexes et donc d'avoir un jugement fiable en toute circonstance tout en sachant être ferme et objectif.

Par professionnalisme, une autre obligation du gestionnaire est de se tenir informé de l'actualité qui touche son secteur ainsi que les nouveaux types de contrats qui apparaissent.

Le gestionnaire, bien qu'il soit spécialisé la plupart du temps ne peut se permettre, de par la nature de son travail à exceller dans un seul domaine. Sa polyvalence est de mise, que ce soit sa performance dans le droit général, mais aussi dans le droit des affaires. Une formation juridique est indispensable car lorsqu'il gère les dossiers de sinistres, seuls le droit civil et le code des assurances lui apporteront de la légitimité dans sa prise de décision.

Il n'est pas nécessaire pour lui d'être un juriste expert car dans beaucoup d'organismes il existe un pôle de juristes pouvant répondre à ces problématiques et pouvant leur venir en aide. Mais des connaissances solides lui apportent une meilleure performance.

En revanche, une base solide du monde de l'assurance lui est indispensable voire même primordiale.

Il doit en effet bien savoir de quoi il est question lorsque l'on lui déclare un sinistre.

Son rôle est simplifié lorsque le sinistre n'est pas complexe. Mais lorsqu'il l'est, il doit être en mesure d'intégrer, d'interpréter, comprendre et analyser les différents avis d'expert qui lui sont remis.

Le rôle du gestionnaire sinistre est important car les incendies, accidents et risques divers incluent l'assurance de dommages et l'assurance de responsabilité. Ce sont des assurances coûteuses pour les assureurs qui se doivent de rembourser les sinistrés autant pour des dommages matériels que corporels.

Si c'est l'assuré qui est responsable de l'accident, l'assureur doit de plus indemniser les victimes dans le cadre de la responsabilité civile.

Le gestionnaire de sinistres gère les plus grosses dépenses de l'organisme assureur pour lequel il travaille. Il se doit donc d'être efficace dans sa compréhension du contexte assurantiel ainsi que dans des situations de sinistres complexes.

Il se doit d'appliquer la politique de gestion des sinistres les procédures propres à l'organisme.

La précision de son évaluation, la provision$^{\text{dossier/dossier}}$, est déterminante pour la stabilité et la santé de l'organisme pour lequel il travaille. Généralement, à l'ouverture d'un dossier sinistre, la provision$^{\text{dossier/dossier}}$ constituée est égale au coût moyen statistique des sinistres de même catégorie. Le but de la prochaine partie va être de mettre en avant un nouveau pan de la statistiques qui jusqu'à présent ne s'était pas faite beaucoup entendre en actuariat : le *Machine Learning*, méthodes statistiques d'apprentissage issues du *Big Data*.

Ces méthodes sont aujourd'hui déjà en implantation dans le domaine de la tarification notamment. Des travaux de recherche en provisionnement sont aussi en train d'émerger.

Le but sera, via une application numérique faite en dernière partie, d'utiliser ces méthodes afin de prévoir, d'estimer, le coût final d'un sinistre a partir des données contractuelles de celui-ci, que ce soit à sont ouverture ou à l'instant d'extraction de la base. Ce afin de faire une comparaison de l'estimation faite par le gestionnaire dossier/dossier. La finalité de cet outils serait de potentiellement implémenter ces algorithmes afin d'aider les gestionnaires dans la gestion de leurs sinistres, de justifier les plafonds de forfaits ou alors de faire réagir le gestionnaire lorsque l'estimation algorithmique est très différente de celle qu'il a faite lui-même.

II.2 Challenge de la provision dossier/dossier

Ce qui nous intéresse ici est le coût d'un sinistre, son coût à l'ultime. Le but sera donc de mettre en place des algorithmes qui, avec un certain nombre de données en entrée (des données liées au sinistre, au bien et au sinistré), pourront prédire le coût ultime dudit sinistre.

Nous regarderons à l'ouverture d'un sinistre les caractéristiques de celui-ci afin d'en prédire le coût à l'ultime. Cette prédiction sera ensuite comparée à la PSAP$^{\text{dossier/dossier}}$ faite par le gestionnaire sinistre.

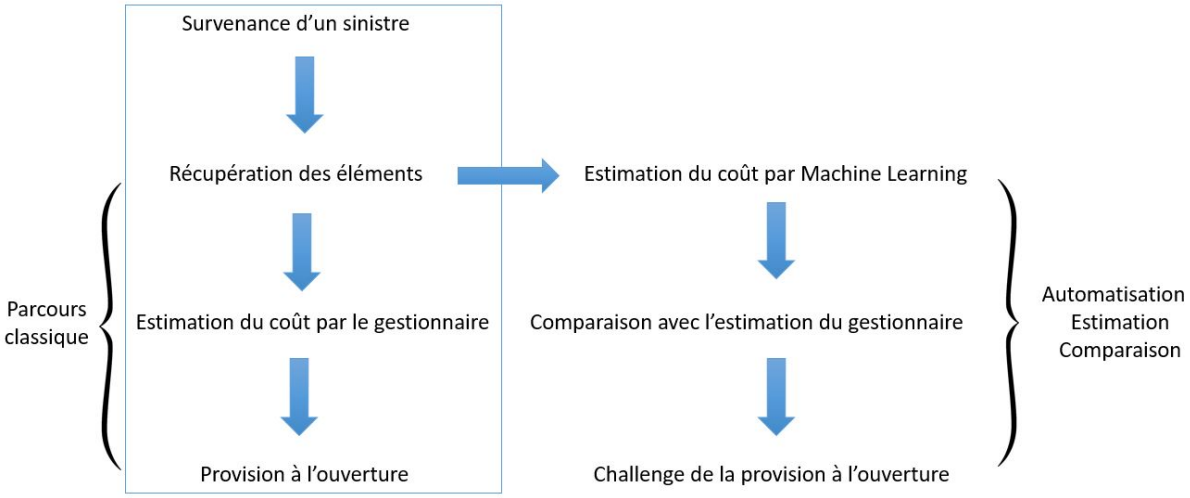

FIGURE 4 – Challenge de la PSAP à l'ouverture

Cette comparaison est faite à l'ouverture mais sera mise en œuvre à l'instant d'extraction de la base, des informations complémentaires étant récoltées au fur et à mesure de la vie des sinistres.

Ces algorithmes font partie du domaine du Machine Learning et travailleront en apprentissage supervisé. Nous nous pencherons sur les arbres de décision, les forêts aléatoires, le gradient boosting ainsi que sur les réseaux de neurones.

Ce sont des méthodes qui n'ont pas pour but d'expliquer la fonctions de lien entre les entrées (variables explicatives) et la sortie (variable à expliquer), ici, le coût à l'ultime d'un sinistre. C'est une approche non paramétrique qui nous permet d'analyser nos données ainsi que de faire une prédiction.

Les assureurs possèdent une certaine quantité d'informations sur leurs clients, leurs biens et leurs contrats mais ne fusionnent pas toujours ces informations afin d'en tirer partie. Ce sera fait ici.

Nous allons donc regarder les sinistres à l'ouverture et faire une prédiction du coût total du sinistre avec les éléments disponible à cet instant. Puis nous ferons cette étude à l'instant d'extraction de la base afin d'utiliser le maximum d'informations possibles.

Ces prédictions, si les résultats s'avèrent cohérents, pourront être utilisée de manière opérationnelle comme un outils d'aide à la gestion pour les gestionnaire de sinistres. Elles apporteront un point de comparaison avec l'évaluation faite par le gestionnaire. Elles pourront aussi servir à la valorisation ou revalorisation des forfaits souvent établis à l'ouverture des sinistres. Finalement, elles pourront interpeler le gestionnaire si ses prévisions et les prédictions algorithmiques sont significativement différentes.

Il est important de noter que ces algorithmes travaillent sur chaque sinistre de manière individuelle. L'analyse est faite ligne à ligne et non pas de manière agrégée. Chaque prédiction sera directement reliée au sinistre qu'elle concerne. Nous aurons donc autant de prédictions que de sinistres évalués.

L'algorithme pourra se baser sur des données à l'instant du sinistre mais pas seulement. Il pourra aussi intégrer des données antérieures, comme le nombre de sinistres précédemment subis par l'assurés, ou le nombre d'avenants à son contrat.

Plus le profil du sinistré sera complet (et la base fournie) plus la prédiction sur le coût de son sinistre pourra être pertinente. C'est l'un des points importants qu'il est nécessaire de maîtriser avec le machine learning : plus l'information est complète, plus l'algorithme pourra amener une meilleure prédiction.

L'intérêt de la mise en place de ces algorithmes est de pouvoir aiguiller la fonction actuarielle dans le processus de provisionnement dossier/dossier. Ils pourraient aussi permettre de pouvoir faire gagner du temps aux gestionnaires de sinistres. En effet, ils pourront grâce à des règles de décision obtenir directement le coût final probable du sinistre qu'ils gèrent. Ceci leur fera économiser du temps, notamment de manière administrative. Ils pourront donc se concentrer sur les sinistres qui nécessitent une attention plus particulière (sinistres graves) et ne pas perdre de temps avec des sinistres attritionnels.

Ce gain de temps permettrait aux gestionnaires de s'investir sur les dossiers complexes mais il permet aussi de réévaluer rapidement le coût d'un sinistre, d'actualiser les dossiers de manière plus régulière.

Ces méthodes permettent aux gestionnaires de tirer une conclusion directe sur le coût potentiel d'un sinistre car elles mettent des règles de décision desquelles découlent un coût. Un coût basé sur des sinistres du portefeuille aux caractéristiques similaires et donc forcément cohérent avec les pratiques du gestionnaire.

La seule modification a apporter lors de la seconde partie de l'étude est qu'il est nécessaire d'introduire une notion de censure à droite lorsque les algorithmes sont sur les sinistres ouverts à une date ultérieure à celle de l'ouverture du dossier. En effet, à une date ultérieure de la date d'ouverture, la variable représentant la durée d'ouverture du dossier (delta de temps entre l'ouverture du dossier et sa fermeture) prend une importance considérable dans le caractère explicatif du coût du sinistre. Il est donc fondamental de la prendre en compte.

Notre but est donc double : d'un côté aiguiller le gestionnaire de sinistres et l'aider à optimiser son temps et de l'autre offrir des modèles cohérents et robustes à la fonction actuarielle afin qu'elle puisse rendre compte des provisions dossier/dossier, de leur cohérence et leur précision.

Deuxième partie
Présentation des méthodes

Cette partie a pour objectif de présenter les méthodes de machine learning qui seront utilisées lors de l'application numérique qui est faite dans la dernière partie.

Ces algorithmes peuvent être catégorisés comme étant les plus connus dans le domaine.

Rappelons que l'intérêt dans notre étude va être de donner des variables explicatives à ces algorithmes pour qu'ils puissent faire une prédiction sur le coût ultime de notre sinistre.

Il faut aussi ajouter que les sinistres ouverts sont censurés à droite à tout instant (car les sinistres ne sont pas clos) et qu'il faut considérer cette censure, censure qui concerne la durée d'ouverture du dossier. Ceci est vrai, sauf à l'instant d'ouverture du dossier car cette variable n'intervient pas.

Tout d'abord nous présenterons les arbres de décisions CART (*Classification and Regression Tree*), pierre angulaire en Data Science qui permet l'exploration et l'exploitation des données.

Cette méthode manquant de robustesse, nous nous tournerons vers les forêts aléatoires. Cette méthode génère des arbres de décisions (des estimateurs donc) et va les agréger tout en permettant à toutes les variables de pouvoir s'exprimer (ce qui n'était pas le cas dans les CART si des variables étaient beaucoup plus influentes que les autres).

Les forêts aléatoires comblent les défauts de robustesse des arbres qui vont avoir tendance à sur-apprendre mais ne permettent pas un paramétrage profond. Elles permettent toutefois une généralisation bien plus fiable que celle des arbres.

Le gradient boosting machine offre une solution à cela optimisant la fonction d'estimation à chaque étape (en permettant à l'utilisateur de choisir une fonction de perte de son choix).

Il permet donc une complexité plus grande que les forêts aléatoires mais peut induire un sur-apprentissage qui biaiserait les résultats. Celui-ci se contourne par exemple en utilisant des méthodes de *k-fold*, méthode qui permet de faire permuter la base d'apprentissage et donc de tester k fois le modèle paramétré et ainsi de prendre celui qui aura les meilleurs résultat sur toute la base d'apprentissage et de test et non plus seulement sur la base de test.

Ensuite nous mettrons en avant les réseaux de neurones. Ce modèle a une architecture calquée sur celle des neurones humains et vise à répliquer ces mêmes interactions. Il peut permettre une prédiction fine mais sa paramétrisation afin d'obtenir le modèle le plus précis est chronophage.

I L'arbre de décision CART

La procédure CART [3] crée un arbre exclusivement binaire. Nous entendons par arbre, un ensemble de questions binaires qui vont scinder les données en fonction de leurs caractéristiques. Celui-ci peut être considéré comme un ensemble de nœuds. Chaque nœud peut se scinder en deux branches ou non dépendant de la séparation qui est faite. Le mot feuille correspond au nom donné à un noeud duquel n'émergent pas de branches. Une feuille peut aussi être considérée comme un nœud terminal

La valeur que l'on cherche à prédire, nommons la Y, qui est obtenue grâce aux variables d'entrée, se lit au niveau de la feuille. Cette variable Y, actuariellement parlant pourrait être un tarif automobile ou dans le cas qui nous intéresse le coût d'un sinistre. Pour cela il est nécessaire de lire l'arbre : il faut partir du commencement de l'arbre, sa racine, de poursuivre le chemin créé au fur et à mesure des critères de segmentations au niveau des nœuds qui amènent à la feuille correspondante.

Cette représentation possède un intérêt majeur : elle est facile à lire et à comprendre, notamment en ce qui concerne la hiérarchisation des variables aléatoires d'entréeOn comprend le poids de leur impact, leur importance et donc leur influence sur la variable aléatoire que l'on cherche à prédire. De la première question on tire la variable la plus importante, de la seconde celle qui l'est un peu moins et ainsi de suite.

Un arbre n'est donc qu'une fonction de $(X_1, \ldots, X_m)$ constante par morceaux sur des rectangles du type :

$$\{a_1 \leq X_1 \leq b_1, \ldots, a_m \leq X_m \leq b_m\}$$

On notera $\mathscr{T}$ l'ensemble des arbres binaires.

Les variables $(X_1, \ldots, X_m)$ peuvent par exemple reprendre les caractéristiques du contrat du client (âge, code postal de résidence, situation familiale), ainsi que du bien qu'il assure (caractéristiques de son véhicule, de son appartement).

La procédure CART est l'algorithme qui vise à rechercher le meilleur prédicteur qui soit : l'arbre $\hat{T} \in \mathscr{T}$. Une fois l'algorithme terminé, nous obtenons un arbre qui prédit notre variable d'intérêt inscrite au niveau des feuilles en fonction des variables explicatives X_i.

I.1 Construction d'un arbre maximal

L'algorithme CART débute par la construction de l'arbre maximal T_{Max}. Le point le plus important dans cette construction est de déterminer quel sera le critère de séparation à placer au niveau de chaque noeud. Au début de la procédure, CART ne connaît ni leur nombre, ni leur emplacement.

Le premier noeud, ou racine, est celui qui est le plus haut dans l'arbre. CART va poser toutes les questions binaires possibles afin de savoir quelle question placer sur ce noeud et ce en utilisant toutes les variables explicatives d'entrée. Prenons un exemple, pour déterminer le coût d'un incendie sur un contrat particulier : "X_1 = Quel est le type de bien ?" ; "X_2 = Le client est-il célibataire ?", etc. CART va poser autant de questions qu'il y a de modalités de variables d'entrée.

L'arbre doit tout d'abord choisir la variable sur laquelle il va effectuer la segmentation et ensuite comment séparer la base de données en deux. Après le processus de sélection de variable, le *split* va

dépendre de la variable.

Si celle-ci est quantitative alors il sera considéré comme un seuil, il faudra donc regarder si la modalité de la variable est au dessus ou au dessous de ce seuil.

Si la variable est qualitative (avec n modalités, disons 4 que nous nommerons A, B, C et D), alors l'algorithme va choisir la séparation qui minimise la variance dans les deux parties. Ceci peut être fait de manière exclusive : si la modalité vaut A alors je fais mon choix. Mais elle ne l'est pas forcément, la base peut être séparée entre deux groupements de modalités : est-ce que la modalité vaut A ou B ? La base est donc scindée avec les modalités A et B d'un côté avec C et D de l'autre. Si nous avions eu plus de modalités sur la variable alors toutes les combinaisons de *split* auraient été analysées.

Une fois que chaque question est posée, la base de données est scindée en deux parties distinctes : la partie pour lequelle la réponse à la question posée est oui, et la partie pour lequelle la réponse est non. Ceci crée deux nouvelles sous-parties qui vont à leur tour créer un noeud. Après chaque question, CART calcule la partition associée ce qui lui permet d'obtenir toutes les différentes partitions de la base de données. Une fois ceci fait, il choisi la partition la plus optimale.

A chaque question est associée une partition. CART va ensuite évaluer l'hétérogénéité produite par la sous-partie par rapport à la variable Y à expliquer, et ce dans les deux noeuds inférieurs obtenus. Si t est un noeud qui correspond à une partie de l'ensemble $\{1, \ldots, n\}$, on note :

$$R(t) = \sum_{e \in t} (Y_e - \mu_t)^2$$

avec e qui désigne un élément de t (c'est à dire un numéro quelconque de ligne de la base de données qui appartient à t) et Y_e qui est la valeur de Y lue dans la base de données, et

$$\mu_t = \frac{\sum_{e \in t} Y_e}{|t|}$$

la moyenne des valeurs de $(Y_e)_{e \in t}$ dans le noeud t.

Ce qui fait donc que $R(t)$ est la variance empirique des éléments appartenant au nœud t, multipliée par le cardinal de t.

Afin de quantifier l'hétérogénéité qui émerge d'une subdivision, CART va calculer le risque associé chaque noeud et va effectuer la somme suivante :

$$\text{Hétérogénéité de la Subdivision } = R(t_g) + R(t_d)$$

Les deux termes du membre de droite représentant respectivement la variance des éléments dans le nœud de gauche et dans le nœud de droite. L'algorithme va ensuite identifier la partition d'hétérogénéité minimale dans les deux nœuds. C'est donc cette partition qui classe le mieux les données car celles-ci sont regroupées de la manière la plus homogène possible dans chaque nœud. C'est donc la partition que CART retient en définitive (avec la question associée).

La subdivision diminue effectivement l'hétérogénéité : si t est la racine et t_g et t_d les nœuds associés à une question donnée, on a $R(t) \geq R(t_g) + R(t_d)$.

Une fois l'opération effectué pour le premier nœud, il faut recommencer avec les deux nouveaux nœuds obtenus. Chaque nœud correspond à un nouvel ensemble de données qu'il convient de classer. CART va ainsi obtenir par récurrence un autre arbre exactement du même type que ceux qui ont été décrits précédemment. Il est à noter que le processus se termine toujours de manière effective car l'arbre est fini. En effet, le nombre de questions possibles est fini car le nombre de variables d'entrée est fini. Il n'a donc qu'à affecter une valeur à chaque nœud terminal. Pour chaque nœud terminal t est affecté la moyenne empirique μ_t de ses données.

Une fois ce processus terminé, CART a obtenu un arbre maximal, que l'on note T_{Max}. Qui plus est, cet arbre est un candidat intéressant comme un "bon" estimateur. C'est le cas car, à chaque étape, nous avons essayé de trier les données de la meilleure manière possible. Nous pouvons donc nous autoriser à penser que que T_{Max} prédit bien notre variable à expliquer Y. Ce n'est pas si simple.

Pour la suite, une simple remarque sur la notation employée jusqu'ici : vu a façon dont nous avons construit l'arbre T_{Max}, il faut constater qu'à un noeud t correspond une partie de l'ensemble $\{1, \ldots, n\}$: t va donc naturellement être associé aux indices des éléments qu'il contient. Prenons un exemple, la racine est associée à l'ensemble $\{1, \ldots, n\}$ tout entier ; la question qui est posée au niveau de la racine est sous la forme "$X_j < a$?" et va donner deux noeuds, celui de gauche associé à la partie $\{i | X_{i,j} < a\}$ et celui de droite associé à $\{i | X_{i,j} > a\}$.

Nous identifierons systématiquement le noeud t et la partie de l'ensemble $\{1, \ldots, n\}$ associée.

I.2 Récapitulatif

Voici un récapitulatif des étapes de construction d'un arbre maximal ainsi de comment la prédiction est effectuée.

Algorithm 1 Création d'un arbre maximal

1: Sélection de la variable de segmentation
2: Sélection du critère de segmentation (*split*) suivant le type de la variable
3: Scinder la base données en deux avec le *split* qui minimise la variance intra groupe et maximise la variance inter groupe
4: Recommencer la première étape dans chacun des deux nouveaux nœuds obtenus

Les critères d'arrêt sont les suivants :

L'arbre a atteint une profondeur fixée
Le nombre de feuilles maximum fixé est atteint
L'effectif du nœud est inférieur au seuil fixé (arrêt de la procédure pour ce nœud)
La qualité de l'arbre n'augmente plus au dessus du seuil fixé

Une fois cet arbre obtenu, son but est de servir à prédire la variable qui nous intéresse sur des données nouvelles.

Dans chaque feuille (ou nœud terminal), pour une régression, la valeur attribuée correspond à la moyenne des observations.

Pour une nouvelle observation, nous lirons et descendrons l'arbre en fonction des critères de segmentation et des modalités de l'observation afin d'arriver à une feuille. La prédiction associée à la nouvelle observation sera donc la moyenne des valeurs observées dans la feuille lors de la création de l'arbre.

I.3 Exemple de construction d'arbre maximal

Prenons l'univers représenté par la figure 5 [7]. Il représente les données connues, c'est la base d'apprentissage. Cet univers ne possède que les deux variables que sont X_1 et X_2.

Supposons que celui-ci représente un zonier de coût de sinistres quelconque et que nous décidons d'utiliser un arbre de décision afin de lier un nouvel arrivant (et ses coordonnées) à une zone. Il faut établir un certain nombre de règles qui en fonction de X_1 et X_2 nous permettrons d'anticiper le coût d'un nouveau sinistre se produisant dans le portefeuille.

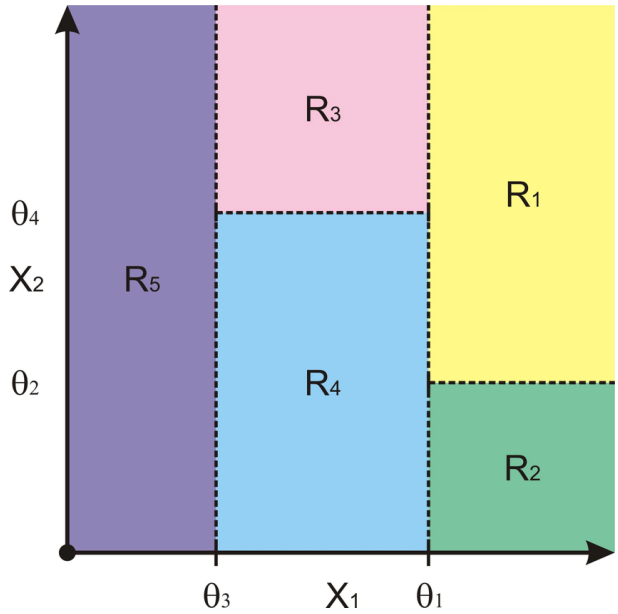

FIGURE 5 – Exemple 1 : segmentation d'un univers bivarié

Le but de cet exemple simpliste va être d'établir un ensemble de règles permettant de déterminer la couleur d'un point (ou son appartenance à un groupe $R_i, i \in [\![1 : 5]\!]$) ayant des coordonnées données en apprenant sur l'univers bivarié représenté précédemment.

L'arbre a pour but de segmenter l'univers afin d'obtenir des zones homogènes.

Supposons que sa première "question" concerne X_1, sa première règle pourrait être :

Si $X_1 < \theta_3$ alors l'individu est de couleur violette (R_5). La base est donc scindée suite à cette décision (que nous nommerons décision 1)

7. https://www.researchgate.net/figure/281602747_fig7_FIGURE-310-Exemple-de-partitionnement-recursif-avec-deux-variables-d%27entree-X-1-et-X-2

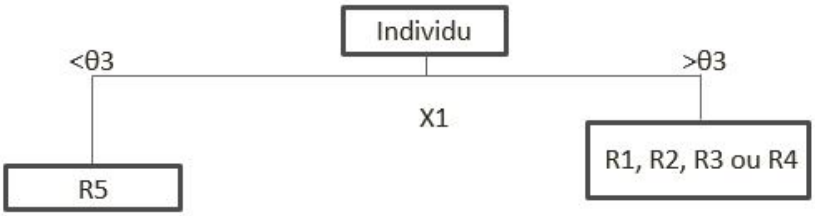

FIGURE 6 – Exemple 1 : Décision 1

Il n'y a plus de questions à se poser concernant la population à gauche de l'arbre, elle est parfaitement classifiée. Nous nous tournons donc vers la branche de droite.

Nous classons donc tous individus qui ne rentrent pas dans la catégorie $X_1 < \theta_3$. La prochaine question concernant X_1 est donc de savoir s'il est supérieur ou non à θ_1, c'est la décision 2.

Si oui alors l'individu est soit jaune, soit vert (dans R_1 ou dans R_2), si non l'individu est soit rose, soit bleu (dans R_3 ou dans R_4).

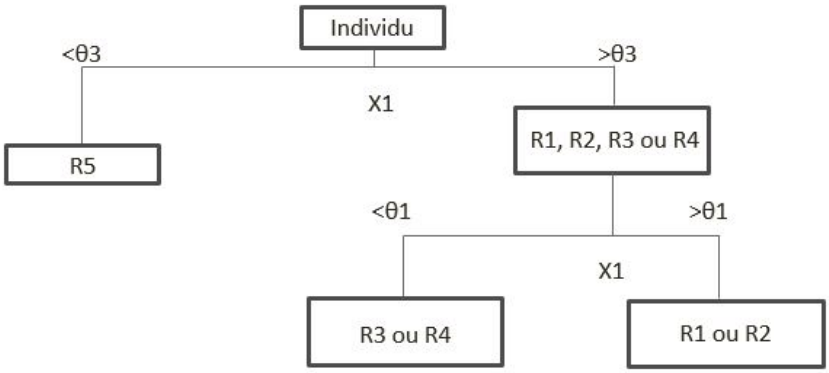

FIGURE 7 – Exemple 1 : Décision 2

Les feuilles de l'arbre ne sont toujours pas homogènes, pas pures, comme celle de gauche (pour $X_1 < \theta_3$). Il faut donc continuer à se poser des questions sur les variables.

Tous les choix possibles ont été faits sur la première variable, on se tourne donc vers la seconde. Par exemple, regardons dans notre branche de gauche restante si $X_2 < \theta_2$. Si c'est le cas alors l'individu est bleu (R_4) et lorsque ce n'est pas le cas, l'individu est bleu ou rose (dans R_3 ou dans R_4). Mais ce choix n'est pas optimal.

En effet, la question de savoir si $X_2 < \theta_4$ dans notre branche de gauche scinde directement l'univers en deux parties homogènes, elle devient la décision 3. Si c'est vrai alors l'individu est bleu (dans R_4) ou sinon il est rose (dans R_3). C'est donc ce choix qui est privilégié.

Il ne nous reste plus qu'à nous interroger sur le membre de droite de notre arbre. La question ici étant de savoir si $X_2 < \theta_2$, la décision 4. Si c'est le cas, l'individu est vert (dans R_2), sinon il est jaune (dans R_1).

Nous pouvons donc observer, après cette décision 4, notre arbre optimal. Les choix ont tous été faits des manière optimale en fonction des séparations de l'espace de départ et dans les feuilles (nœuds ter-

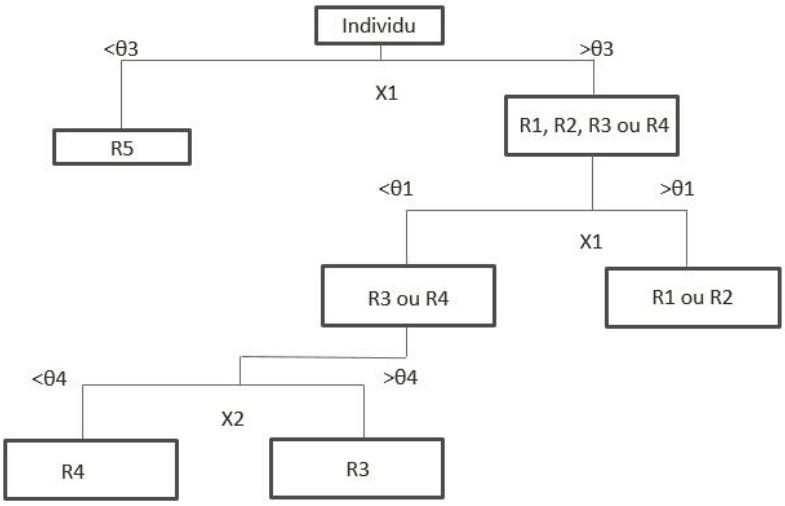

FIGURE 8 – Exemple 1 : Décision 3

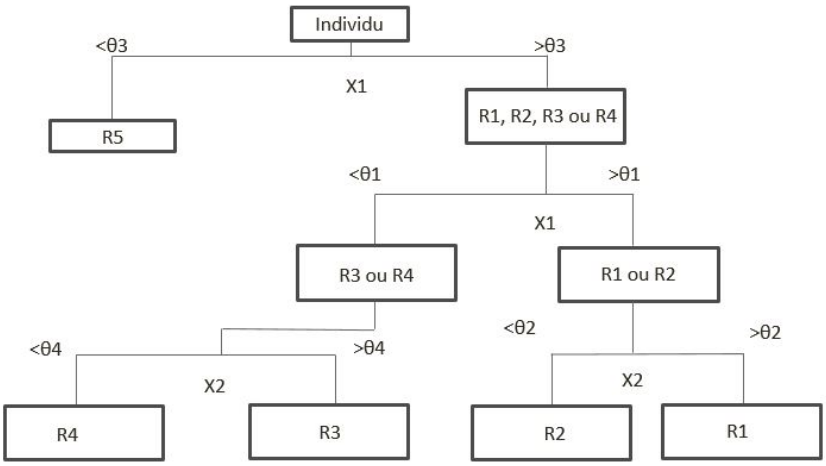

FIGURE 9 – Exemple 1 : Décision 4

minaux), chaque classe possible est unique. On dit que la pureté de la feuille est de 100%. Ici, ceci a été possible car l'exemple était simpliste.

Évidemment, le même procédé peut s'utiliser en dimension supérieure.

I.4 Elagage

Le but de l'élagage va être de permettre à l'arbre construit de se détacher de la base d'apprentissage à partir de laquelle il a été construit.

Il est nécessaire, afin d'augmenter le pouvoir de généralisation de l'arbre, de supprimer certaines parties de l'arbre (branches) qui soit sont peu représentatives soit qui sont redondantes. Faire un élagage permet d'éviter le sur-apprentissage.

Le choix des critères d'élagages se font relativement au souhait de prédiction de l'arbre. Ceci est

abordé dans la section "Le compromis biais-variance".

Prenons T et T' deux arbres, nous noterons $T \preceq T'$ si T est un sous-arbre de T'. Cela veut dire que tous les nœuds et toutes les branches de T sont des nœuds et branches de T'. Par conséquent T et T' ont la même racine. De plus, si $T \preceq T_{Max}$, nous associerons à une feuille t de T la valeur moyenne μ_t des données qu'elle contient.

I.4.1 Evaluation de l'erreur R(t) d'un arbre

Soit $T \in \mathscr{T}$ un arbre binaire.

Nous souhaitons disposer d'un critère qui permet de déterminer si T est un bon prédicteur, c'est-à-dire si la quantité

$$\mathbb{E}[(Y - T(X))^2]$$

est petite.

Idéalement, nous rechercherions même à minimiser cette quantité sur T tout entier, mais nous allons en fait nous contenter d'effectuer la minimisation parmi quelques arbres bien choisis.

Ne connaissant pas la loi du couple (X, Y) et n'ayant que des données en nombre fini à disposition nous en sommes réduit à évaluer une estimation de l'erreur théorique $\mathbb{E}[(Y - T(X))^2]$. Appelée taux d'erreur de l'arbre T, cette erreur empirique est définie comme suit : $R(T) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - T(X_i))^2$ (nous faisons la moyenne des résidus). D'après la loi des grand nombres, l'erreur empirique tend vers l'erreur théorique lorsque le nombre d'observations n tend vers l'infini, le taux d'erreur est donc consistant.

En regroupant les termes pour chaque feuille, nous voyons que le taux d'erreur $R(T)$, en désignant par $\tilde{T}$ l'ensemble des feuilles de l'arbre T, vaut également $R(T) = \sum_{t \in \tilde{T}} R(t)$.

Pour obtenir le meilleur arbre possible, nous sommes donc tenté de prendre :

$$\hat{T} = \underset{T \preceq T_{Max}}{\arg\min}\, R(T)$$

Une fois notre arbre obtenu sur des données dont nous connaissons la réponse Y, il suffit de prendre les individus dont cette réponse est inconnue et de parcourir l'arbre en suivant les règles de décision produite par celui-ci. Nous obtenons donc, en arrivant au nœud terminal une variable de sortie pour des données d'entrée qui n'en n'avaient pas.

I.5 Exemple d'élagage

Si nous reprenons l'exemple précédent, l'arbre ne nécessite aucun élagage car il est impossible de généraliser l'univers plus que ce qu'il n'a été fait. Afin de comprendre l'élagage, il est nécessaire de prendre un exemple plus complexe.

Prenons comme univers la figure 10. Considérons que tous les individus observés aient été sinistrés et qu'il n'existe que deux coûts de sinistres possibles : les cercles correspondent à des individus ayant

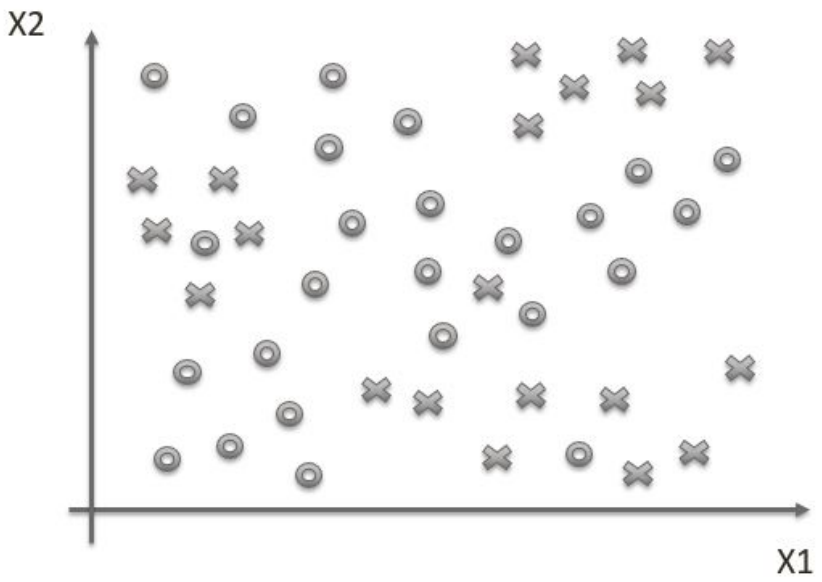

FIGURE 10 – Exemple 2 : segmentation d'un univers bivarié

subi un sinistre de coût α et les croix correspondent à des individus sinistrés ayant coûté β. Ici, nous souhaitons donc prédire le coût d'un nouvel individu sinistré en fonction de ses variables explicatives X_1 et X_2 (qui peuvent être par exemple son âge et le coût de sa cotisation).

De la même manière que sur l'exemple précédent, nous obtenons une segmentation de l'univers (segmentation faite en optimisant la variance inter et intra groupes), avec l'arbre maximal, représentée figure 11.

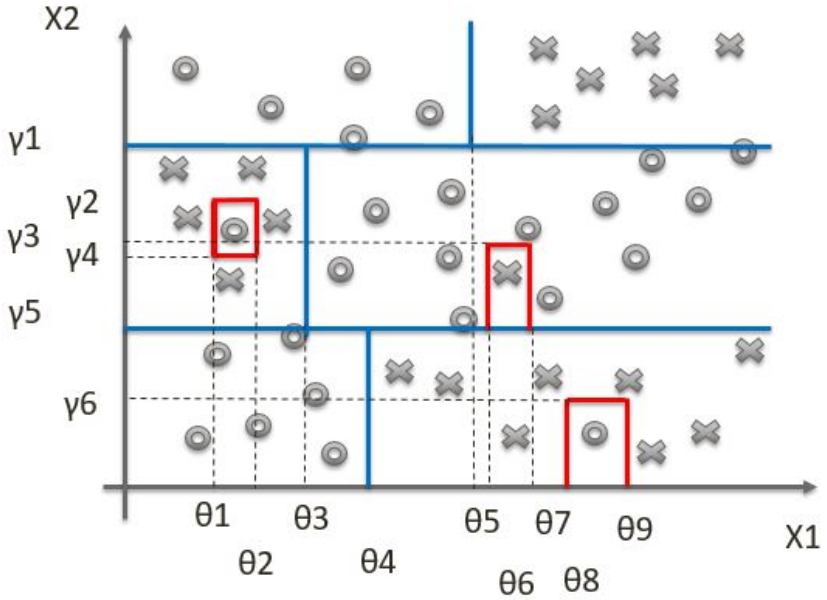

FIGURE 11 – Exemple 2 : segmentation maximale

Comme pour l'exemple précédent, il est possible de récréer un certain nombre de règles qui per-

mettraient de découper avec précision l'univers en question. Toutefois, ici, trois individus ont été isolés (partitions en rouge). En effet, ils sont fondus dans un groupe d'individus de la catégorie différente de la leur. L'arbre maximal, qui pousse l'information jusqu'au bout, a bien repéré ces individus et les a isolés.

Ceci s'appelle le sur-apprentissage. L'algorithme a analysé toutes les aspérités de la base d'étude. C'est à ce moment qu'intervient la compromis biais-variance.

Malheureusement, dans le cadre d'une généralisation, pour classer un individu quelconque intégrant la base d'étude, ces points son aberrants. L'élagage va donc consister en la suppression des segmentations entourant les individus isolés. La prédiction future de l'arbre sur des données inconnues sera meilleure grâce à cela.

Après cet élagage, nous obtenons la division de la figure 12.

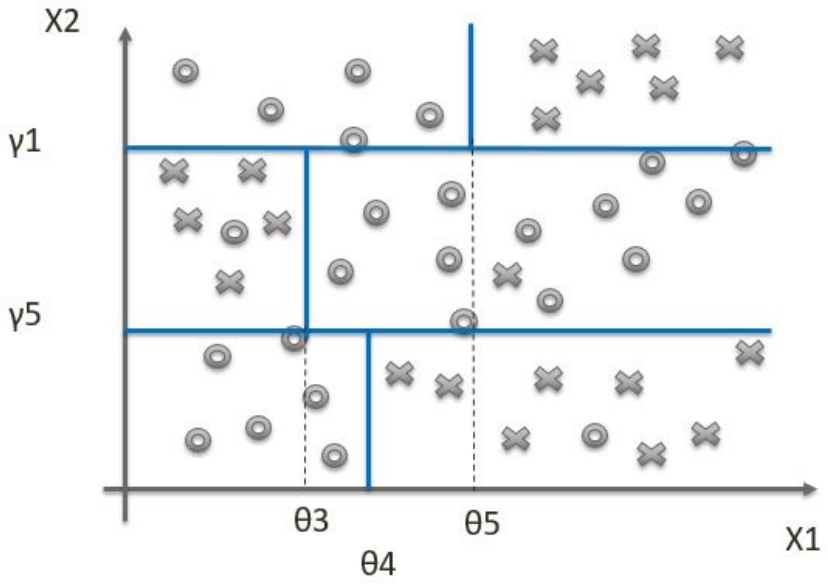

FIGURE 12 – Exemple 2 : segmentation après élagage

Ceci revient à simplifier l'arbre, à supprimer des règles de décision, à décomplexifier le modèle. Mais aussi à le rendre plus robuste, meilleur en généralisation. Il sera une meilleur prédicteur sur des données nouvelles. L'élagage sert à gommer les variations particulières de la base étudiée.

I.6 Avantages et Inconvénients

Les avantages et inconvénients de cette méthode sont les suivants :

Avantages	Inconvénients
Performant	Déconseillé si l'effectif est petit
Il n'y a pas de paramétrage complexe	La binarisation n'est pas toujours appropriée
Connaissances intelligibles	Recherche pas-à-pas
Robuste face aux données aberrantes	Performances dépendant de la taille de l'arbre
Rapidité de traitement	

II Les méthodes d'aggrégation

Les méthodes d'aggrégation sont des méthodes qui regroupent les résultats d'algorithmes pour obtenir une meilleure prédiction qu'un estimateur simple. Le Random Forest est une méthode d'ensemble.

Le Random Forest est une amélioration du Bootstrap Aggragation (Bagging), qui lui même vient de l'application du Bootstrap. Pour comprendre ce concept il est donc nécessaire de comprendre les concepts dont il découle.

II.1 Le Bootstrap

Le Bootstrap correspond à un échantillonnage de la base de données initiale.

Nous allons en effet créer des nouveaux échantillons à partir de notre base de données initiale en tirant aléatoirement avec remise n individus de notre base de données (n étant le nombre d'individus initial dans la base). Une représentation de ceci est fait Figure 13.

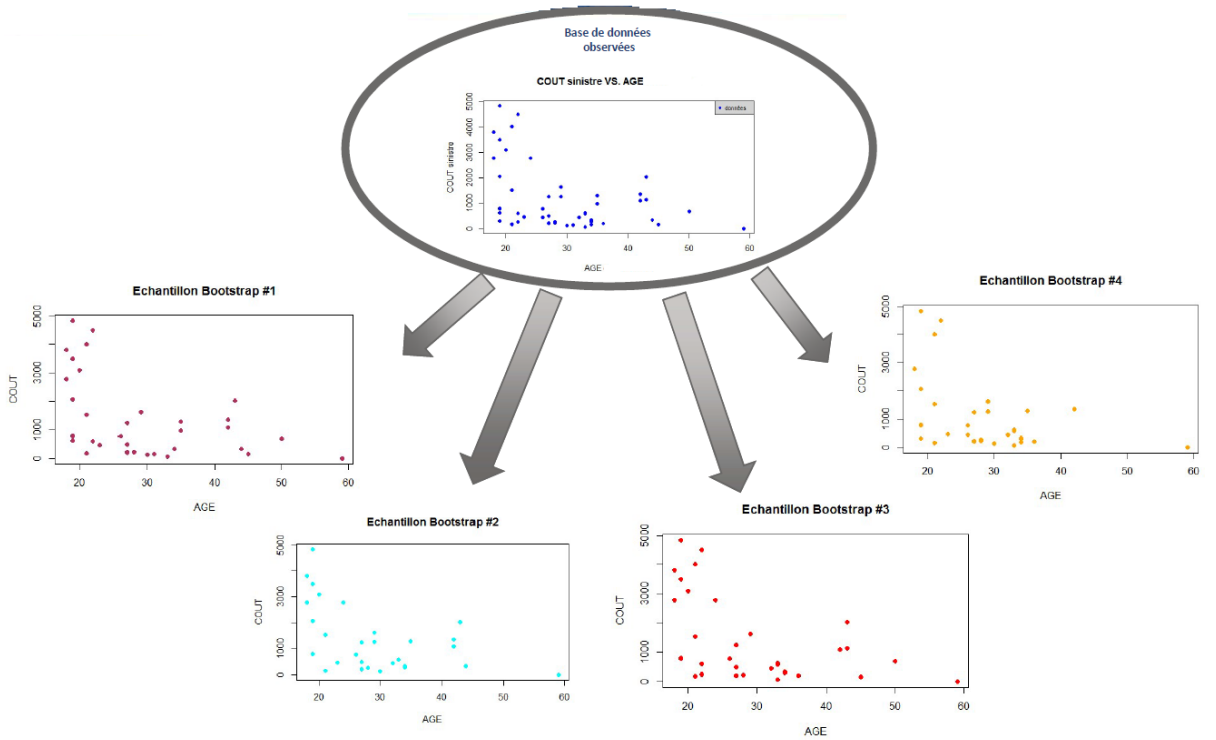

FIGURE 13 – Création d'échantillons Bootstrap

De manière purement mathématique on considère une variable aléatoire X dont la fonction de répartition F_x est inconnue.

Le but est de faire une estimation à partir d'un échantillon aléatoire $(X_1, \ldots, X_n)$ indépendant identiquement distribué.

On distingue :

— le bootstrap non paramétrique : la loi de F est inconnue

— le bootstrap paramétrique : la loi est connue mais le paramètre est inconnu

Le principe est le suivant :

Pour toute fonction de répartition de la forme

$$\theta(F) = \int h(x)dF(x)$$

On utilise l'approximation suivante qui correspond à un estimateur des moments :

$$\theta(\hat{F}) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

On détermine la loi de $\hat{F}$ via des simulations par Monte Carlo et en opérant des tirages avec remises dans $(X_1, \ldots, X_n)$.

Bootstrap et provisionnement

Dans notre cas, nos variables explicatives sont les différentes caractéristiques du sinistre, du client et du bien sinistré. On applique le bootstrap aux résidus d'un modèle pertinent déjà calibré.

En effet les $X_{i,j}$ sont indépendants mais pas identiquement distribués. On procède donc au rééchantillonage des résidus qui eux le sont.

La procédure bootstrap est la suivante :

— obtention d'un échantillon bootstrap de résidus de taille $b \in [\![1; B]\!]$ par rééchantillonage de l'ensemble des résidus initiaux,
— détermination de la provision bootstrappée,
— pour le modèle retenu et ces nouvelles données $\mathscr{T}$, prévisions des valeurs des différentes moyennes et calcul de la charge ultime,
— on stocke cette charge ultime à b donné et on recommence avec b incrémenté.

II.1.1 Avantages et inconvénients

Avantages	Inconvénients
Donne un intervalle de confiance Pas d'hypothèses sur la loi des variables	Ne fonctionne pas si les queues de distributions sont trop épaisses Ne fonctionne pas si les statistiques font intervenir des valeurs extrêmes

II.2 Les familles de modèles aléatoires

Le principe de base de ces méthodes réside dans l'agrégation de modèles [4][6][11]. Celles-ci peuvent être basées sur des stratégies de deux types :

— Aléatoire (*bagging*, *random forest*)
— Adaptative (*boosting*) : construction adaptative (déterministe ou aléatoire) d'une famille de modèles.

Ceci permet une amélioration de l'ajustement par cette combinaison (agrégation) d'un grand nombre de modèles tout en évitant le sur-ajustement (sur-apprentissage).

A noter que l'erreur déterminée est l'erreur du modèle plus l'erreur d'estimation. Ici nous sommes dans un cadre non paramétrique, il n'y a donc pas de formule explicite pour la formulation de l'erreur, il est donc nécessaire de recourir à la validation croisée.

Ces principes de *bagging* et de *boosting* s'appliquent à toutes les méthodes de modélisation mais n'ont d'intérêt réel que dans le cas de modèles instables (CART). En effet il y a un lien immédiat entre instabilité du modèle et variance de l'estimateur. L'agrégation permet de réduire cette variance.

II.2.1 Le Bagging

Définissons quelques notations :

— Y : variable à expliquer (qualitative ou quantitative)
— $X = (X^1, \dots, X^p)$ les covariables, avec $x = (x^1, \dots, x^p) \in \mathbb{R}^p$
— $\Phi(x)$ est le modèle
— $\hat{\Phi}(x)$ l'estimateur du modèle
— $z = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ l'échantillon observé

Dans le bootstrap, on considère b échantillons indépendants notés $z_b, b \in [1, B]$ et on construit un modèle agrégé :

— si Y est quantitative : $\hat{\Phi}_B() = \frac{1}{B} \sum_{b=1}^{B} \hat{\Phi}_b()$ (moyennisation)
— si Y est qualitative : $\hat{\Phi}_B() = \underset{j}{argmax}$ card $(b|\hat{\Phi}_b() = j)$ (notion de vote majoritaire). Pour j donné on regarde le plus grand nombre d'estimateurs qui donne j

Le principe consiste à moyenner les résultats (prévisions) pour réduire la variance et donc l'erreur des prévisions sur plusieurs modèles indépendants. En fait, on utilise le bootstrap car il n'est pas toujours raisonnable (par manque de données par exemple) de créer B échantillons indépendants (tout simplement en ne prenant en compte que des individus différents). On se sert donc de la mesure empirique ($\hat{F}$) pour créer des réplications bootstrap.

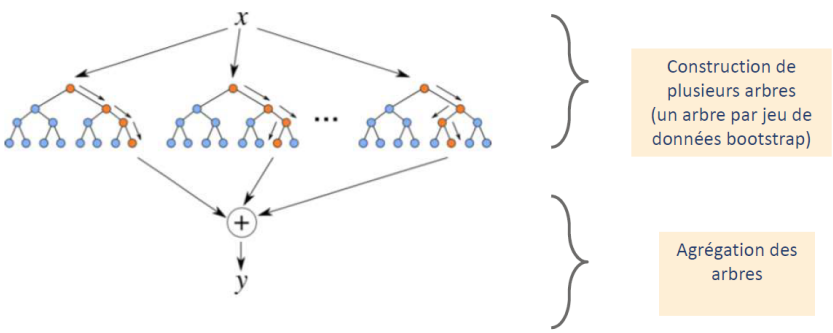

FIGURE 14 – Processus du Bagging

La méthode de Bootstrap Aggregation est une méthode simple mais dans l'ensemble très efficace.

Cette technique combine les prédictions de multiples algorithmes afin de donner une prédiction plus précise qu'un modèle individuel.

Le Bootstrap Aggregating est une procédure générale qui peut être utilisée pour réduire la variance de résultats produits par des algorithmes qui ont une grande variance, ce qui est le cas pour les arbres de décisions comme le CART.

Les arbres de décision sont sensibles aux données spécifiques sur lesquelles ils font leur base d'apprentissage. Si celle-ci est modifiée (si l'on fait l'apprentissage de l'arbre sur une sous base de la base d'apprentissage) alors les résultats inhérents à cet arbre seront modifiés.

Le Bagging est l'application d'une procédure de Bootstrap à un algorithme de machine learning de variance élevée, il est donc naturel de penser à appliquer cette méthode aux arbres de décision.

Lorsque l'on utilise une procédure de Bagging avec des arbres de décision le fait qu'un arbre fasse du sur-apprentissage sur sa base d'apprentissage est moins un soucis. Pour cette raison et pour plus d'efficacité, les arbres de décisions individuels sont "poussés" au maximum et ne sont pas élagués. Ces arbres auront donc une variance élevée mais un biais faible. Ce sont les caractéristiques les plus importantes des sous modèles lorsque l'on combine des prédictions utilisant le Bagging.

Lorsque l'on fait du Bagging sur des arbres de décision, les seuls paramètres sont le nombre d'échantillons et donc le nombre d'arbres à inclure. Ceci peut être déterminé en faisant augmenter le nombre de simulations jusqu'à ce que le nombre de simulations ne montre plus d'améliorations. Simuler un grand nombre de modèles peut prendre du temps mais il n'y aura pas de sur apprentissage des données.

L'erreur Out-of-Bag : Le principe consiste à tirer des individus aléatoirement et à effectuer un bootstrap à partir de ces individus la (les individus non choisis). Les individus choisis sont les individus *Out-of-Bag*.

L'échantillon est comme habituellement divisé en deux :

— base d'apprentissage pour la construction du modèle
— base de validation pour l'optimisation des paramètres de *tuning* (*cross-validation*)

Enfin, la performance du modèle optimisé est testée sur les données *OOB* via la *Mean Squared Error* si Y est quantitative ou par la probabilité de se tromper si Y est qualitative.

Utilisation du concept : En pratique ce sont souvent les arbres CART qui sont utilisés comme "briques de base" pour construire une famille de modèles (famille d'arbres binaires). Les forêts aléatoires sont généralement les plus robustes. En effet on fait une "moyenne" d'arbres, ce qui correspond à un lissage des passages de l'espace χ des covariables X.

Il existe trois stratégies d'élagage :

1. Laisser construire chaque arbre de l'estimateur agrégé (forêt), comme un arbre maximal

2. Construire un arbre d'au plus "q" feuilles
3. Laisser construire chaque arbre comme un arbre maximal puis faire de l'élagation par validation croisée

Le bagging possède l'avantage d'être simple et facile à programmer mais demande un temps de calcul important ainsi qu'un stockage en mémoire de chacun des modèles élémentaires et ce sans apporter une interprétation facile.

II.2.2 Forêt Aléatoire

La méthode de Forêt Aléatoire a pour but de faire pousser un certain nombre d'arbres (de classification ou de régression dépendant du problème à traiter). Pour classifier un nouvel objet venant d'un vecteur d'entrée il faut l'insérer dans chaque arbre de la forêt. Chaque arbre va classifier l'objet mis en entrée, il va en quelque sorte "voter" pour la classe qui est sensée lui revenir. La Forêt Aléatoire va ensuite agréger ces résultats pour donner la classification de l'objet. En ce qui concerne la régression, c'est une moyennisation du résultat qui est extraite.

Cette méthode peut être considérée comme une application double du Bagging : une première fois sur les données, puis ensuite sur les variables à choisir à chaque nœud.

Chaque arbre est construit de la manière suivante :

1. Si le nombre de cas dans la base de test est N, il faut échantillonner N cas au hasard (mais avec remplacement) des données d'origine. Cet échantillon sera la base de test pour construire l'arbre
2. S'il y a M variables d'entrée, un nombre $m << M$ est spécifié pour qu'à chaque noeud m variables soient sélectionnées au hasard (parmi les M) et que la meilleure séparation sur ces variables soit faite sur ce noeud. La valeur de m est constante lors de la création de l'arbre
3. Chaque arbre est poussé à son nombre de feuille maximal, il n'y a pas d'élagage

La construction d'une réponse d'une forêt aléatoire est montrée figure 15 [8]. Le nombres de variables m tirées est choisi par l'utilisateur, mais de manière empirique celui-ci est fixé à $\sqrt{p}$ ou $\frac{p}{3}$.

Il a été montré que le taux d'*erreur* de la forêt dépendait de deux facteurs :

1. La *corrélation* entre deux arbres de la forêt : plus la corrélation entre deux arbres est élevée, plus le taux d'erreur augmente
2. La *force* de chaque arbre individuel dans la forêt : un autre avec un taux d'erreur faible est un classificateur fort. L'augmentation de la force individuelle des arbres fait chuter le taux d'erreur

La réduction de m fait à la fois chuter la corrélation et la force des arbres. L'augmentation de m fait augmenter ces deux facteurs. Il faut donc trouver la valeur optimale de m, qui peut être contenue dans un intervalle large. On peut déterminer cette valeur en utilisant l'erreur "Out of Bag", l'erreur sur les individus qui ne font pas parti de la base de test (voir infra), elle nous donne rapidement une valeur de m proche. C'est le seul paramètre ajustable auquel la forêt aléatoire est quelque peu sensible.

8. https://www.researchgate.net/profile/Yohann_Mansiaux/publication/281184702/figure/fig7/AS:335467745693698@1456993151132/FIGURE-46-Construction-d%27une-foret-aleatoire.png

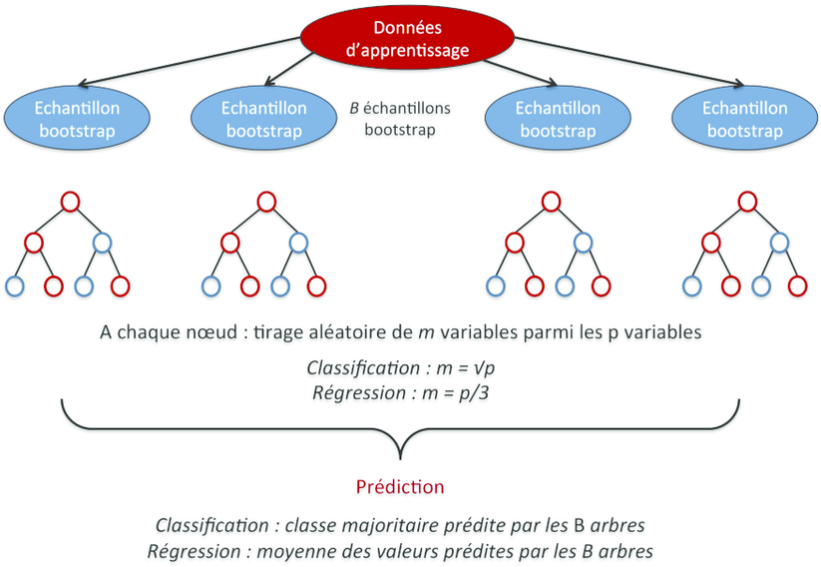

FIGURE 15 – Principes de construction d'une forêt aléatoire

Les intérêts de la forêt aléatoire sont les suivants :

— Sa précision, meilleure que d'autres algorithmes dans beaucoup de cas
— Fonctionne efficacement sur une base de données large
— Elle peut gérer un nombre conséquent de variables d'entrée en les conservant toutes
— Elle extrait les variables importante dans la classification ou la régression
— Elle génère un estimateur non biaisé de l'erreur de généralisation au fur et à mesure que la forêt progresse
— C'est une méthode efficace pour estimer des données manquantes en maintenant la précision lorsque une majeure partie de l'information est manquante
— Les modèles peuvent se réutiliser sur d'autres données
— Elle offre une méthode expérimentale d'interaction entre les variables

Les forêts aléatoires ne sont pas considérées comme du *Bagging* car toutes les variables ne sont pas sélectionnées lors de la séparation d'un noeud en deux feuilles.

Cette méthode n'est basée uniquement que sur des arbres CART et ajoute une composante aléatoire. L'objectif étant de rendre la prévision la plus robuste possible et plus indépendante de chaque arbre en ajoutant du hasard dans le choix des covariables qui interviennent dans chacun des arbres.

A noter que plus le nombre de covariables est grand, plus le modèle est efficace.

Importance des covariables Celle-ci peut être évaluée par deux mesures :

— *Mean Decrease Accuracy* : plus l'erreur de permutation est importante moins la variable l'est
— *Mean Decrease Gini*

Remarque : Les Forêts Aléatoires ne font pas de sur-apprentissage, on peut générer autant d'arbres qu'on le souhaite.

L'erreur "Out-of-Bag" (OOB) est utilisée pour obtenir un estimateur non biaisé de l'erreur de classification lorsque les arbres sont ajoutés à la forêt. Il est aussi utilisé pour obtenir des estimations de l'importance de la variance.

A chaque fois qu'un arbre est construit, toutes les données sont passées à travers la forêt et des proximités sont calculées pour chaque paire de cas. Si deux cas occupent la même feuille terminale, alors leur proximité est augmentée de 1. A la fin de la simulation, les proximités sont normalisées par le nombre d'arbre. Les proximités sont utilisées pour remplacer des données manquantes, localiser des données aberrantes et amener des statistiques sur les données.

L'estimation de l'erreur "Out-of-Bag" : Dans des forêts aléatoires il n'y a pas besoin de validation croisée ou d'autre test pour obtenir un estimateur non biaisé de l'erreur de la base de test. Il est calculé de manière interne de la manière suivante :

Chaque arbre est construit en utilisant un échantillon de bootstrap différent provenant des données originales. A peu près un tiers des cas sont mis de côté par le bootstrap et ne sont pas utilisés dans la construction du k^{è} arbre. On moyenne ensuite les cas non utilisés pour construire le k^{è} via le k^{è} arbre. Des cette manière, une moyennisation de la base de test est obtenue pour chaque cas dans un tiers des arbres. A la fin de la simulation, en considérant la classe j comme étant celle qui a eu le plus de votes à chaque fois que la case n était OOB. La proportion de fois que j n'est pas égal à sa vraie classe, moyenné sur n, sur tous les cas est l'estimation de l'erreur OOB. Ce qui a été prouvé comme étant sans biais.

Importance de variable : Dans chaque arbre construit par la forêt, il faut noter les cas d'OOB et compter le nombre de votes donnés pour la classe correcte. Les valeurs de la variable m sont permutées aléatoirement dans les cas d'OOB et on relance le processus dans l'arbre concerné. Ensuite, il est nécessaire de soustraire le nombre de votes pour la classe correcte dans les données de variables m-permutées du nombre de votes corrects dans les données non concernées par l'OOB. La moyenne de ce nombre sur tous les arbres dans la forêt est l'importance brute du score de m.

Si le nombre de variables est important, la forêt peut être simulée une fois avec toutes les variables et compilée une autre fois en utilisant seulement les variables les plus importantes issues de la première simulation.

Pour chaque cas, il faut considérer tous les arbres qui sont concernés par l'OOB. Il faut soustraire le pourcentage de votes pour la classe correcte dans les données OOB m-permutées du pourcentage de votes pour la bonne classe corrects dans les données non concernées par l'OOB. Ceci représente l'importance locale du score pour les variables m dans ce cas.

Interactions : La définition opérationnelle utilisée pour l'interaction est que les variables m et k interagissent si une séparation sur une de ces variables, disons m, dans l'arbre rend une séparation sur k qui soit systématiquement moindre (ou à l'inverse plus possible). L'implémentation utilisée est basée sur les valeurs de Gini pour chaque arbre dans la forêt. Elles sont classées pour chaque arbre et pour chaque variable deux à deux, la valeur absolue de la différence de leurs rangs étant moyennée sur tous les arbres.

En régression, le but est donc de rechercher des découpes qui tendent à diminuer la variance dans les noeuds obtenus : c'est-à-dire diminuer $\sum(Y_i - \overline{Y}_t)$ où Y_t est la moyenne des Y_i des observations présentes dans le noeud t.

II.2.3 Avantages et Inconvénients

Les avantages et inconvénients de cette méthode sont les suivants :

Avantages	Inconvénients
Bonnes performances en prédiction	Problème si nombre de variables pertinentes très faibles
Paramétrage simple	Le déploiement d'un tel modèle reste compliqué
Evaluation de l'erreur intégrée	
Mesure de l'importance des variables	

II.3 Famille de modèles adaptatifs

Le principe du boosting est d'apporter une amélioration séquentielle de la suite des estimateurs créés. De la même manière, on cherche à réduire la variance grâce à un estimateur agrégé. Mais là où on ne corrigeait pas le biais de l'espérance B des arbres biaisés, on le fait désormais.

Dans les modèles adaptatifs, une série de modèles sont construits de sorte qu'à chaque étape, chaque modèle ajouté à la combinaison, apparaît comme un pas vers une meilleure solution.

Remarque : Dans le cas des arbres CART "simples" par exemple, le biais peut être important. La méthode de boosting diffère fortement du bagging sur la construction de la famille de modèles qui est cette fois récurrente. Chaque modèle est une version adaptative du précédent, en donnant plus de poids aux observations mal ajustées.

Intuitivement, l'algorithme concentre ses efforts sur les observations les plus difficiles à ajuster tandis que l'agrégation limite le sur-apprentissage.

Il existe plusieurs algorithmes pour le boosting. Ils diffèrent par leurs caractéristiques :

- La façon de pondérer l'importance des observations mal estimées précédemment
- La façon de pondérer les modèles élémentaires lors de l'agrégation
- Leur objectif : prédire une variable Y qualitative ou quantitative
- Leur fonction perte pour mesurer l'erreur d'ajustement (plus ou moins sensible aux valeurs atypiques)

II.3.1 Gradient Boosting Machine

La version originale du *boosting* (ADABOOST, *Adaptative Boosting*) était proposée pour un problème de classification à deux classes : $Y \in \{-1; 1\}$.

Voyons l'algorithme d'origine.

On note δ la fonction de classification (ou discrimination).

Nous indicerons les modèles par la lettre m.

L'algorithme est le suivant :

Soit x_0 à prévoir et $z = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un échantillon.

On initialise les poids $\omega = \{\omega_i = \frac{1}{n}, i \in [\![1, n]\!]\}$ (équipondération).

Pour m allant de 1 à M :

— On estime δ_n sur l'échantillon pondéré par ω
— On calcule le taux d'erreur apparent de notre modèle : $\hat{\xi}_p = \frac{\sum_{i=1}^{m} \omega_i \mathbf{1}_{\delta_m(x_i) \neq y_i}}{\sum_{i=1}^{m} \omega_i}$
— On calcule les logit : $c_m = -\ln\left(\frac{1-\hat{\xi}_p}{\hat{\xi}_p}\right)$
— On met à jour les pondérations : $\omega_i \leftarrow \omega_i e^{c_m \mathbf{1}_{\delta_m(x_i) \neq y_i}}$
— On prédit le résultat (du vote) : $\hat{\Phi}_M(x) = \text{ sgn } \left(\sum_{m=1}^{M} c_m \delta_m(x)\right) = \pm 1$

En pratique, on ajoute à cet algorithme une étape de vérification du taux d'erreur du modèle de chaque modèle, où le taux d'erreur de prévision $\hat{\xi}_p$ doit être plus petit que 0.5 (modèle aléatoire).

II.3.2 Version Aléatoire

Au lieu de définir les poids de cette manière, il est possible d'influer sur l'échantillon en lui même en sur-représentant les observations mal prédites.

C'est la même idée que le bootstrap avec des probabilités de tirage non uniformes. En particulier, cette probabilité va dépendre de l'erreur sur l'observation. Quand l'erreur augmente, la probabilité de tirer l'observation augmente pour la construction de l'échantillon.

II.3.3 Cas de la régression

Y est ici quantitative.

L'algorithme est le suivant :

Soit x_0 à prévoir et $z = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un échantillon.

On initialise un vecteur de probabilité p avec une distribution uniforme $p = \{p_i = \frac{1}{n}\}$

Pour m allant de 1 à M :

— On tire avec remise dans z un échantillon z_m^* suivant p
— On estime $\hat{\Phi}_m$ sur z_m^*
— On calcule à partir de l'échantillon initial z :
  — $l_m(i) = Q(y_i, \hat{\Phi}(x_i))$ pour $i \in [\![1, m]\!]$
  — $\hat{\xi}_m = \sum_{i=1}^{m} p_i l_m(i)$
  — $\omega_i = g(l_m(i))$ avec g continue décroissante

— On calcule les nouvelles probabilités de tirage : $p_i = \frac{\omega_i}{\sum_{i=1}^{m} \omega_i}$

On calcule $\hat{\Phi}_M$ comme moyenne (ou médiane) des prévisions $\hat{\Phi}_M(x)$ pondéré par des coefficients $log(\frac{1}{\beta_m})$ où $\beta_m = \frac{\hat{\xi}_m}{L_m - \hat{\xi}_m}$ avec $L_m = \sup_{i \in [\![1,m]\!]} l_m(i)$ et $g(l_m(i)) = \beta_m^{\frac{1-l_m(i)}{L_m}}$

On doit ajouter une condition comme précédemment si l'erreur se dégrade trop : $\hat{\xi}_m < 0.5L_m$.

Concernant la censure : Il faut noter qu'encore une fois, lorsqu'une variable est censurée, il suffit de réutiliser le processus de censure sur les arbres générés afin d'intégrer la censure dans le Gradient Boosting Machine.

II.3.4 Avantages et Inconvénients

Les avantages et inconvénients de cette méthode sont les suivants :

Avantages	Inconvénients
Souplesse avec le choix des fonctions de coût	Modèle non explicite
S'adapte aux spécificités du problème	Paramètres nombreux
Evaluation de l'erreur intégrée	Danger de sur-apprentissage
Mesure de l'importance des variables	Lourdeur et intensité des calculs

III Les réseaux de neurones

Le terme *réseau de neurones* a évolué pour englober un certain nombre de modèles et de méthodes de machine learning [11]. Dans cette partie nous décrirons le réseau de neurone *vanilla* ou perceptron simple puis le perceptron multicouches. Cette méthode est un modèle statistique non linéaire.

Dans cette méthode, les poids sont déterminés par rétro-propagation, il n'est donc pas nécessaire d'en instaurer comme dans les méthodes avec les arbres.

Les réseaux de neurones sont aussi appelés "perceptron multicouches". Ils sont issus de la branche informatique et plus précisément de l'intelligence artificielle. Leur but est de simuler le comportement d'un cerveau humain.

Au départ, une approche "connexioniste" était développée (1940-1960) : couches (entrée, cœur, sortie). Mais la présence des ordinateurs était limitée.

Une autre approche a ensuite été développée : l'approche séquentielle, dont le but est d'automatiser le principe de l'expertise humaine via 3 concepts :

— une base de connaissance (composée de propositions logiques élémentaires)
— une base de faits (observation, données)
— un moteur d'inférence : applique des règles expertes sur la base de faits. en déduit de nouveaux fait jusqu'à la réalisation d'un objectif.

Ceci posait un problème, la modélisation des connaissances d'un expert humain et la complexité des algorithmes d'inférence ont freiné le développement des réseaux de neurones mais aujourd'hui la technique est de nouveau largement utilisée grâce à certains développements théoriques aussi :

— estimation d'un gradient par rétropropagation de l'erreur (Hopkins, 1982)
— analogie avec les modèles Markoviens en mécanique statistique (Hopfield, 1982)

Aujourd'hui, les réseaux de neurones ont une grande variété d'application (reconnaissance d'image) et sont complémentaires d'autres techniques plus classiques.

III.1 Un réseau neuronal

Association de ce que l'on appelle des neurones "formels" [11]. Cela crée un graph plus où moins complexe d'objets élémentaires.

Les différents réseaux de neurones se distinguent par l'organisation du graphe (couches), leur niveau de complexité (nombre de neurones), le type des neurones (activation ou transition) et l'objectif du réseau (apprentissage supervisé ou non)

III.2 Le réseau formel

Défini (comme un neurone biologique) par :

— un état interne $s \in S$
— des signaux d'entrée $x_1, \ldots, x_p$

— une fonction d'estimation : $s = h(x_1, \ldots, x_p) = f(\alpha_0 + \sum_{j=1}^{p} \alpha_j x_j)$
  — On appelle "poids" le vecteur $\alpha = (\alpha_0, \ldots, \alpha_p)$
  — "α_0" est le biais

Remarque : les poids α_j sont estimés durant la phase d'apprentissage : ils sont la mémoire ou "connaissance répartie" du réseau.

Les différents types de neurones se distinguent par leur fonction d'activation : "σ"

— type linéaire : $\sigma = Id$
— type sigmoïde : $\sigma(x) = \frac{1}{1+e^{-x}}$
— type seuil : $\sigma(x) = \mathbf{1}_{[0,\infty[}(x)$
— type radiale : $\sigma(x) = \frac{1}{\sqrt{(2\Pi)}} e^{-\frac{x^2}{2}}$
— type stochastique :

$$\sigma(x) = \begin{cases} 1 \text{ avec une certaine probabilité} \\ 0 \text{ sinon} \end{cases}$$

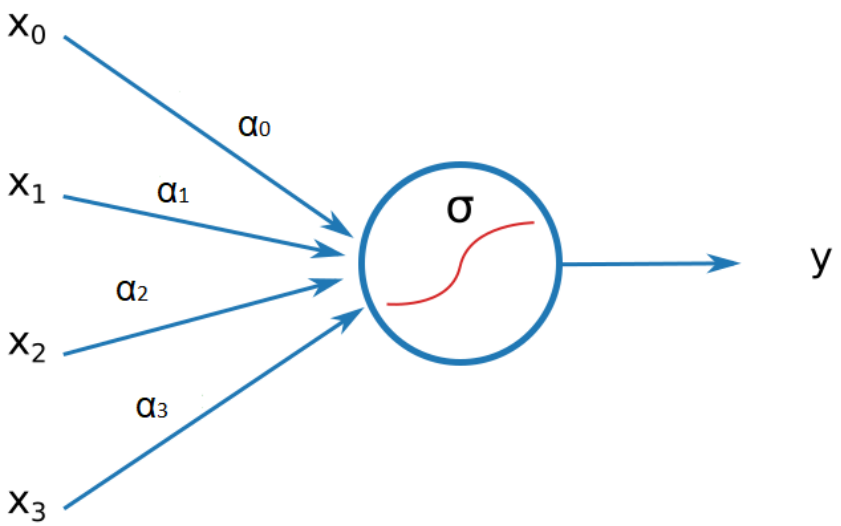

FIGURE 16 – Illustration d'un perceptron simple

Sur la figure 16[9], les x_i représentent nos variables explicatives, les α_i sont les poids qui leurs sont apposés. Toutes les variables sont reliées à un neurone, qui avec sa fonction de transfert σ a pour but de prédire la sortie y.

En suivant le fil conducteur de ce mémoire, les x_i représentent les informations du sinistre, du client, de son contrat et de son bien et la sortie représente la $\text{PSAP}_{\text{dossier/dossier}}$ à l'ouverture.

Nous aurons donc un lien entre les informations et la provision associée.

9. https://www.miximum.fr/blog/introduction-au-deep-learning-1/

III.3 Perceptron Multicouche (PMC)

Ce type d'algorithme peut s'utiliser dans un cadre d'apprentissage supervisé ou non, ici nous nous restreignons au cadre de l'apprentissage supervisé.

III.3.1 Architecture

Le PMC est composé de couches successives, où une couche est un ensemble de neurones "sans connexion" entre eux.

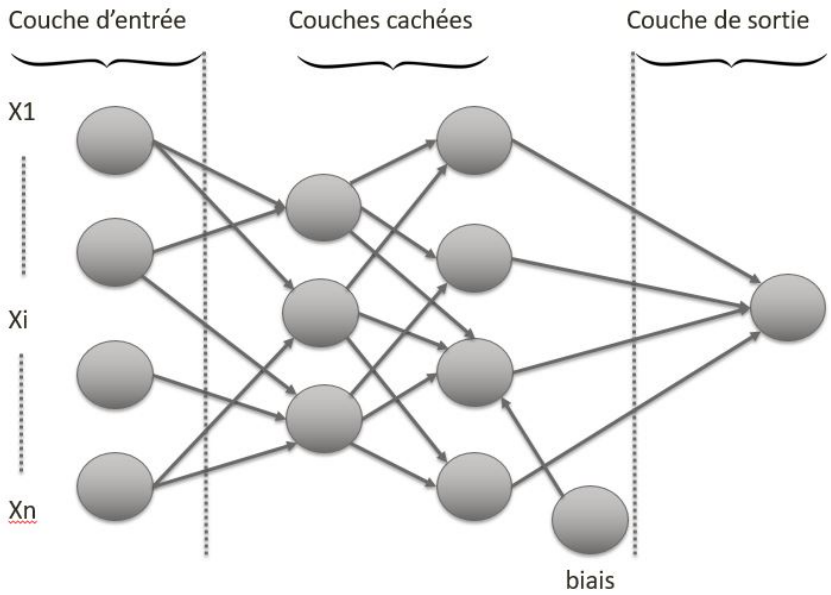

FIGURE 17 – Réseau de neurones multicouches

Voici un résumé intuitif du fonctionnement d'un perceptron multicouches.

Sur la figure 17, nous pouvons voir un réseau de neurones (3,4). La première couche cachée possède 3 neurones, la seconde en possède quatre. Nous voyons aussi qu'un neurone de biais a été introduit.

Le fait d'utiliser plusieurs neurones et plusieurs couches permet de trouver des solutions non linéaires au problème étudié.

Ici, chaque flèche représente une pondération d'une variable d'entrée vers un neurone (ou d'un neurone vers un neurone). Les pondérations α_i (ou θ_i sur la figure 16) n'ont pas été apposés pour ne pas surcharger la figure.

Le neurone excentré de la seconde couche cachée n'est pas liée directement à une pondération des entrée. C'est une pondération autre : un biais.

Considérons encore une fois le problème du provisionnement : nos entrées X_i représentent les variables rassemblées via le contrat du sinistré, la situation de son sinistré et les informations sur son bien. Le but est toujours d'estimer le coût final du sinistre dans la couche de sortie.

Toutes les informations reçues et possédées par le gestionnaire de sinistre peuvent être renseignées en variables d'entrée.

Au départ, ces poids sont tous initialisés de manière identique (ou aléatoirement) et l'algorithme est lancé une première fois. Celui-ci obtient donc une prédiction pour notre sortie (le coût d'un sinistre par exemple) qu'il va comparer au résultat attendu. L'erreur sera ensuite "remontée" aux neurones précédents la sortie avec une répondération des poids et ceux de couche en couche jusqu'à ce que les poids initiaux (ceux qui affectent les variables d'entrée) soient modifiés. Ceci s'appelle la rétropropagation et permet, une fois le passage fait, d'avoir des poids plus justes et une prédiction plus précise. L'algorithme vient de boucler sur lui-même une première fois.

Suite à cela, l'algorithme va reprendre les données d'entrée avec les nouveaux poids et recompiler jusqu'à obtenir une nouvelle fois une estimation de la sortie.

Le processus de rétropropagation est réitéré. Ceci jusqu'à ce que l'algorithme ait convergé ou que le nombre d'itérations fixées par l'utilisateur ait été dépassé. Le réseau apprend donc des données.

De manière purement mathématique, il a été montré qu'avec un nombre fini de neurones, le perceptron multiple était capable d'approximer des fonctions continues avec très peu d'hypothèses sur les fonctions d'activation, c'est le théorème d'approximation universelle.

Il a de plus été mis en avant le fait que c'est la structure du réseau de neurone et non pas le choix spécifique de la fonction d'activation qui permet au réseau de neurones d'être un approximateur universel.

Théorème "d'approximation universelle" : Toute fonction régulière peut être approchée uniformément avec une précision arbitraire et dans un domaine fini de l'espérance de ses variables par un réseau de neurones comportant une couche de neurones cachée, en nombre fini possédant tous la même fenêtre d'activation et une neurone de sortie linéaire.

Usuellement on a :

— Une régression (la variable de sortie Y est continue) avec une dernière couche composée d'un neurone, avec une fenêtre d'activation identité, tandis que les neurones cachés ont une fenêtre d'activation sigmoïde
— Une classification binaire ($Y \in \{0, 1\}$) : neurones de sortie munis de la sigmoïde
— En discrimination à m classes : m neurones de sortie avec activation sigmoïde

La généralisation est la suivante : en régression avec un perceptron à une couche cachée de p neurones et un neurone de sortie, la fonction de transfert s'écrit :

$$Y = \Phi(x_1, \ldots, x_p, \alpha) = \Phi(x_1, \ldots, x_p, \alpha, \beta) = \beta_0 + \beta^T z$$

où

$$z_k = f(\alpha_{k0} + \alpha_k^T x) \text{ pour } k \in [\![1, q]\!]$$

k étant l'identifiant du neurone dans la couche cachée, Φ la fonction de transfert et β le vecteur $(\beta_1, \ldots, \beta_q)$.

III.4 Apprentissage du réseau

Supposons avoir une base d'apprentissage de n observations $(x_i^1, \ldots, x_i^p, y_i)_{1 \leq i \leq n}$ de variables explicatives $X_1, \ldots, X_p$ et la variable à prévoir Y.

Dans le cas de la régression et avec un réseau de neurones composé d'une couche cachée à p neurones et d'une sortie linéaire, les paramètres (poids α, β) sont optimisés par moindres carrés. Ceci se généralise à toute fonction de perte dérivable.

"L'apprentissage" est l'estimation des paramètres (α, β) pour $j \in [\![1, p]\!]$, $k \in [\![1, q]\!]$ et (β_k) où $k \in [\![1, q]\!]$ par minimisation de la perte quadratique.

$$Q = \sum_{i=1}^{n} Q_i = \sum_{i=1}^{n} (Y_i - \phi(x_i, \alpha, \beta))^2$$

avec Φ fonction de transfert du réseau.

Différents algorithmes sont utilisés pour estimer (α, β), généralement basés sur une évaluation du gradient par auto-propagation.

III.5 Rétro-propagation de l'erreur

Ceci revient à évaluer la dérivée de la fonction de perte (ou de coût) en une observation et par rapport à l'ensemble des paramètres.

Sur l'observation i, $z_{ki} = f(\alpha_{k0} + \alpha_k^T x_i)$ (avec k le $k^{\text{è}}$ neurone de la couche cachée) et $z_i = (z_{1i}, \ldots, z_{qi})$ (z_i outputs de la couche cachée).

$$\begin{cases} \frac{\partial Q_i}{\partial \beta_k} = \frac{\partial (Y_i - \phi(x_i, \alpha, \beta))^2}{\partial \beta_k} = \frac{\partial (Y_i - (\beta_0 + \beta^T z_i))^2}{\partial \beta_k} = -2(Y_i - \phi(x_i))(\beta_{z_i}^T) z_{ki} \\ \frac{\partial Q_i}{\partial \alpha_{k_j}} = \frac{\partial (Y_i - (\beta_0 + \beta^T z_i))^2}{\partial \alpha_{k_j}} = \frac{\partial (Y_i - (\beta_0 + \beta^T (f(\alpha_{k0} + \alpha_k^T x_i))))^2}{\partial \alpha_{k_j}} = -2(Y_i - \phi(x_i)(\beta_{z_i}^T) \beta_k f'(\alpha_k^T x_i) x_{ip}) \\ \qquad = \delta_i \beta_k f'(\alpha_k^T x_i) x_{ip} = s_{ki} x_{ip} \end{cases}$$

Posons :

$$\delta_i = -2(Y_i - \phi(x_i))(\beta^T z_i)$$

Avec :

- δ_i : terme d'erreur du modèle courant à la sortie ("erreur de passage" de la couche cachée vers la sortie finale)
- S_{ki} : terme d'erreur du neurone k dans la couche cachée ("erreur commise" dans le passage des inputs x_i vers la couche cachée)

Ces deux termes vérifient les équations dites de "rétropropagation"

$$s_{ki} = f'(\alpha_k' x_i) \beta_k \delta_i$$

Le but étant d'estimer α et β.

Pour évaluer les gradients, on a donc besoin d'évaluer δ_i et S_{ki}. Cela se fait en deux étapes :

1. une passe "avant" : valeurs courantes des poids permet de déterminer $\hat{\phi}(x_i)$
2. une passe "retour" : $\hat{\phi}(x_i)$ $(= Y_i)$, on peut évaluer δ_i puis s_{ki} par "rétropropagation"

III.6 Algorithme d'optimisation

Nous savons évaluer les gradients, donc il suffit d'utiliser un algorithme adapté. Le plus simple et le plus court est l'utilisation itérative du gradient, par exemple l'algorithme du Newton-Raphson.

L'idée est qu'en tout point de l'espace des paramètres le vecteur gradient de Q pointe dans la direction de l'erreur croissante. Il suffit donc de se déplacer en un sens contraire que pour Q décroisse.

Avec un modèle GLM il faut maximiser la vraisemblance (max $\log L(\beta; Y)$) et on cherche $\hat{\beta}$.

On va chercher l'annulation de la dérivée de la log-vraisemblance.

La formule de Taylor-Lagrange nous apprend que pour toute fonction f (dérivée de la log vraisemblance) C^∞, au voisinage d'un point (x_0) :

$$\forall x, f(x) = f(x_0) + (\frac{x - x_0}{1!})f'(x_0) + \frac{(x - x_0)^2}{2!}f''(x_0) + o(x - x_0)^2$$

A l'étape ν :

$$\beta_k^{\nu+1} = \beta_k^{\nu} - \tau \sum_{i=1}^{n} \frac{\partial Q_i}{\partial \beta_k^{(\nu)}}$$
$$\alpha_k^{\nu+1} = \alpha_k^{\nu} - \tau \sum_{i=1}^{n} \frac{\partial Q_i}{\partial \alpha_k^{(\nu)}}$$

avec τ le taux d'apprentissage

Implémentation de la rétropropagation élémentaire du gradient : L'algorithme de rétropropagation est le suivant :

Initialisation : les poids sont tirés de manière aléatoire et uniforme sur [0,1] (en ayant normalisé les données d'apprentissage au préalable)

Boucle : Tant que Q > erreur-max ou nombre-itération < nombre-itération-max

Il est nécessaire de ranger la base d'apprentissage dans un nouvel ordre aléatoire :

Pour chaque individu $i \in [\![1, n]\!]$:

— Calculer $\xi(i) = (Y_i - \phi(x_i^1, \ldots, x_i^p))$ en propageant les entrées vers l'avant

— L'erreur est retropopagagée dans les différentes courbes pour affecter à chaque entrée une "responsabilité" dans l'erreur globale
— Mise à jour des poids $\alpha_{jkl}(i) = \alpha_{jkl}(i-1) + \Delta\alpha_{jkl}(i)$

Rappelons que τ est le taux d'apprentissage. Il peut être fixé par l'utilisateur (c'est un paramètre de tuning), voir varier en cours d'exécution. Si τ est grand l'algorithme converge plus vite vers la solution au prix d'une solution moins précise. Si τ est petit c'est l'inverse.

III.7 Paramètres et complexité

Le réseau de neurones possède certains paramètres et sa complexité dépend de la structure voulue par l'utilisateur. Il est aussi à noter que plus le volume des données en entrée est dense et plus le nombre de couches et de neurones est élevé, plus le réseau de neurones devra faire d'estimation. Ce qui fera exploser le temps de calcul.

Les variables d'entrée et de sortie sont évidemment des paramètres à posséder afin de pouvoir faire de l'apprentissage et générer des prédictions.

Il faut ensuite se questionner sur l'architecture du réseau et notamment choisir les deux paramètres suivants :

— le nombre de couches (cachées)
— le nombre de neurones par couche

Ces deux choix jouent sur la complexité du réseau, et donc sur la recherche d'un bon compromis biais/variance pour un bon équilibre entre qualité d'adéquation et prévision.

Il existe trois autres paramètres que sont :

— erreur maximale tolérée
— nombre maximum d'itérations
— un terme de régularisation éventuel

Il faut aussi prendre en compte le taux d'apprentissage RIDGE : $Q(\alpha_{jkl}) + \gamma\|\alpha_{jkl}\|_2$

γ, aussi appelé "decay", est donc un paramètre de contrôle de sur-apprentissage.

En pratique l'utilisateur ne règle pas simultanément tous ces paramètres. Il est confronté à des choix dont le but est essentiellement de contrôler le phénomène de sur-apprentissage, par une technique classique de l'erreur : les validations croisées (dans une moindre mesure, échantillon apprentissage-test, bootstrap).

III.8 Avantages et Inconvénients

Les avantages et inconvénients de cette méthode sont les suivants :

Avantages	Inconvénients
Classificateur très précis (fonction des paramètres)	Modèle "boîte noire"
Incrémental	Paramétrisation complexe
Utilisable sur de grandes bases	Danger de sur-apprentissage

IV Le compromis biais-variance

Cette section concerne tous les modèles que nous avons décrit précédemment. Le compromis biais-variance reflète le fait de vouloir être le plus proche des données tout en étant possible de généraliser le mieux possible. Malheureusement, il est souvent impossible de faire ces deux choses de manière simultanée.

Les modèles qui présentent un biais faible sont généralement plus complexes et permettent de représenter les données d'apprentissage avec une précision accrue. Malheureusement, ceci entraîne un sur-apprentissage et donc fait baisser la prédiction de ces modèles.

Les modèles qui eux présentent un biais élevé sont des modèles généralement plus simples. Ils ne captent pas toutes les complexités de la base d'apprentissage et font donc des erreurs du à leur sous-apprentissage.

Les différents cas sont représentés de manières schématiques sur la figure 18

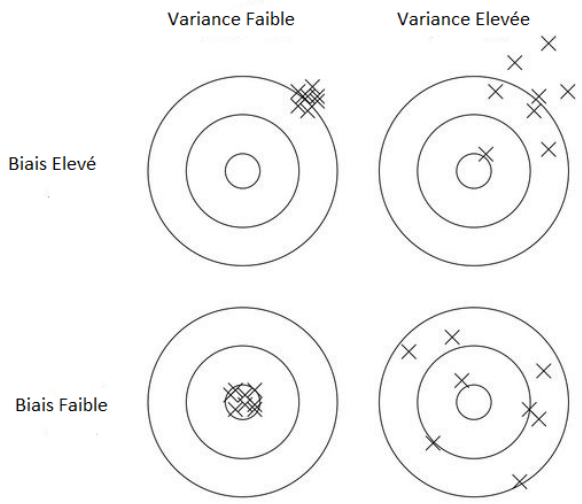

FIGURE 18 – Visualisation du biais et de la variance des modèles

Réussir à concilier ces deux complications est appelé le compromis biais/variance [10].

Ce compromis revient à minimiser deux sources d'erreurs qui ne permettent pas aux algorithmes de pouvoir généraliser au delà des données sur lesquelles il a appris.

Si nous prenons par exemple un modèle d'arbre CART, nous comprenons donc qu'il existe aussi un compromis à établir entre la précision du modèle et les fluctuations des données.

— D'une part, un arbre trop peu ramifié montre un biais très important. Dans le cas extrême d'un arbre ne possédant qu'une feuille, la valeur de la variable à expliquer est toujours la moyenne des données.
— D'autre part, un arbre trop profond perd en pertinence dans les détails de la base, il présente donc une variance importante. C'est le cas en particulier de notre arbre maximal T_{Max}.

10. https://www.quantmetry.com/single-post/2015/03/27/Initiation-au-Machine-Learning

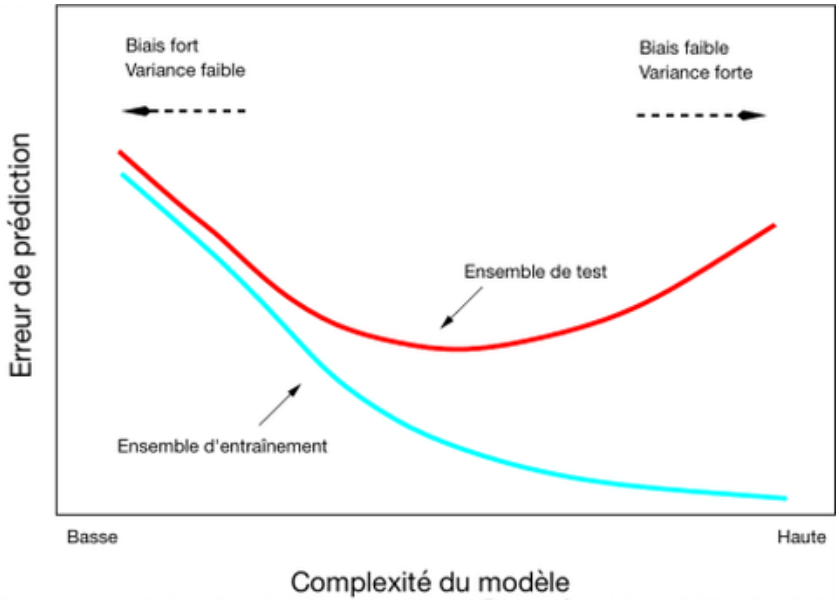

FIGURE 19 – Le compromis biais-variance

Il faut donc élaguer l'arbre pour résoudre ce problème.

On peut influer sur le compromis-biais variance des arbres CART en modifiant les critères d'élagage. C'est évidemment le cas pour tous les autres méthodes basées sur ces arbres.

Pour les réseaux de neurones ce système d'élagage n'existe pas, seule une modification des paramètres est possible.

V Prise en compte de la censure

Dans toute la partie précédente nous avons présenté des méthodes qui permettent d'obtenir une réponse (ici la provision du gestionnaire) concernant des données dont nous ignorons la variable de sortie. Toutes les données explicatives sont brutes à la date d'extraction et ne nécessitent aucune modélisation lorsque l'étude est faite à l'ouverture du dossier.

Lorsque l'on souhaite faire l'étude à la date d'extraction, une variable n'est pas complètement visualisée : la durée de vie du dossier sinistre. Si nous la considérons telle qu'elle, nous n'observons qu'un minimum (si le dossier ferme le jour de la date d'extraction alors nous avons la durée exacte, mais s'il ferme plus tard alors ce n'est plus le cas). Cette variable est censurée à droite.

Cette variable étant d'une importance cruciale, il est nécessaire de prendre en compte cette censure pour ne pas avoir des prédictions qui sous estiment la réalité.

Nous allons donc opérer sur cette variable une modification : nous allons, pour les dossiers ouverts à la date d'extraction, lui rajouter sa durée de vie résiduelle pour obtenir une estimation de la durée totale d'ouverture du dossier sinistre.

Cette durée de vie résiduelle sera obtenue grâce à l'estimateur de Kaplan-Meier qui permet d'obtenir une fonction de survie de nos données.

Il convient donc de présenter cet estimateur.

V.1 L'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier [12] (Kaplan et Meier [1958]) peut être introduit via les processus ponctuels, en remarquant que la fonction de survie de base du modèle est l'unique solution de l'équation intégrale suivante :

$$S(t) = 1 - \int_0^t S(u^-)h(u)du$$

L'équation ci-dessus exprime simplement le fait que la somme des survivants en t et des individus sortis avant t est constante. Lorsque la fonction de survie est continue, la démonstration est immédiate en effectuant le changement de variable $v = lnS(u)$, $dv = -h(u)du$.

On a :

$$h(u)du = \mathbb{P}(\text{sortie en u et u+du}|\text{en vie en u})$$

Un estimateur de cette quantité est

$$\frac{\overline{N}^1(u+du) - \overline{N}^1(u)}{\overline{R}(u)} = \frac{d\overline{N}^1(u)}{\overline{R}(u)} \text{ si } \overline{R}(u) > 0$$

Où

$$\overline{N}^1(t) = \sum_{i=1}^{n} N_i^1(t) \text{ où } N_i^1(t) = 1_{\{T_i \leq t, D_i = 1\}}$$

$$\overline{R}(t)) = \sum_{i=1}^{n} R_i(t) \text{ où } R_i(t) = 1_{\{T_i \geq t\}}$$

En remplaçant $h(u)du$ par son estimateur $\frac{d\overline{N}^1(u)}{\overline{R}(u)}$, on peut proposer un estimateur de la fonction de survie en cherchant une solution à l'équation suivante :

$$\hat{S}(t) = 1 - \int_0^t \hat{S}(u^-) \frac{d\overline{N}^1(u)}{\overline{R}(u)}$$

On peut montrer qu'il existe une unique solution à cette équation et on obtient alors l'estimateur de Kaplan-Meier de la fonction de survie.

La construction heuristique de l'estimateur de Kaplan-Meier s'appuie sur la remarque suivante qui est que la probabilité de survivre au delà de $t \geq s$ peut s'écrire :

$$S(t) = \mathbb{P}(T > t | T > s)\mathbb{P}(T > s)S(s)$$

On peut renouveler l'opération, ce qui fait apparaître des produits de termes en $\mathbb{P}(T > t | T > s)$; si l'on choisit comme instants de conditionnement les instants où se produit un évènement (sortie ou censure), on se ramène à estimer des probabilités de la forme :

$$p_i = \mathbb{P}(T > T_{(i)} | T > T_{(i-1)})$$

p_i est la probabilité de survivre sur l'intervalle $]T_{(i-1)}; T_{(i)}]$, sachant qu'on était vivant à l'instant $T_{(i-1)}$. Un estimateur naturel de $q_i = 1 - p_i$ est $\hat{q}_i = \frac{d_i}{r_i} = \frac{d_i}{n-i+1}$. On observe alors qu'à l'instant $T_{(i)}$, et en l'absence d'ex aequo, si $D_i = 1$ alors il y a sortie par décès donc $d_i = 1$, et dans le cas contraire l'observation et censurée et $d_i = 0$. L'estimateur de Kaplan-Meier s'écrit donc finalement :

$$\hat{S}(t) = \prod_{T_{(i)} \leq t} (1 - \frac{1}{n - i + 1})^{D_{(i)}}$$

En pratique cependant on est confronté à la présence d'ex aequo ; on suppose alors par convention que les observations non censurées précèdent toujours les observations censurées. On obtient l'expression suivante de l'estimateur :

$$\hat{S}(t) = \prod_{T_{(i)} \leq t} (1 - \frac{d_i}{r_i})$$

V.2 Observations Censurées

Dans la suite [9], nous nous intéressons au vecteur aléatoire (M, T, X) où $M^{\in}\mathbb{R}^p$, $T \in \mathbb{R}^+$ la variable de durée, et $X \in \mathscr{X} \subset \mathbb{R}^p$ un ensemble de covariables aléatoires qui pourraient avoir un impact sur T et/ou M. La présence de censure nous empêche l'observation directe de (M, T), alors que X est toujours observé. On introduit la variable de censure $C \in \mathbb{R}^+$. Elle représente ici la durée d'ouverture de notre dossier sinistre.

Par soucis de simplicité, nous considérons que T et C sont des variables continues. Nous faisons aussi l'hypothèse, sans perte de généralité, que les éléments de M sont tous strictement positifs. Les variables observées, au lieu de (M, T) sont :

$$\begin{cases} Y = inf(T, C) \\ \delta = 1_{T \leq C} \\ N = \delta M \end{cases}$$

Les données sont composées de $(N_i, Y_i, \delta_i, X_i)_{1 \leq i \leq n}$ i.i.d. Comparativement à une régression classique, les variables M_i ne sont des quantités observées que lorsque l'individu i n'est pas censuré.

Dans notre étude, M représente la charge ultime d'un sinistre, T représente la durée d'ouverture du dossier sinistre et X les informations concernant le bien, le contrat et le sinistre.

V.3 Exemple

La censure impacte-t-elle vraiment de manière significative la manière dont nous pouvons faire une estimation ?

Considérons le tableau suivant :

Charge ultime	Survenance	Durée d'ouverture	variable 4	variable 5	variable 6	variable 7	variable 8
176,4	2015	312	1	1	4	1	2
418	2015	32	1	2	6	1	2
36343,26	2015	431	1	2	6	4	1
830,9	2015	93	2	1	2	4	1
2194,66	2014	374	1	2	3	4	1
0	2015	452	1	2	6	4	1
959,05	2015	19	1	2	7	4	1
294,8	2015	31	1	2	7	4	1
8257,98	2015	19	1	2	6	1	2
1340	2015	-	1	2	6	4	1
525,9	2015	46	1	2	5	4	1
0	2015	232	1	2	5	1	2
1160,6	2015	51	1	2	8	4	1
2901,25	2015	-	2	2	7	1	2
855	2015	384	1	2	3	1	2
857,97	2015	-	2	2	2	4	1
11550	2015	-	1	2	3	4	1

Table 1 – Exemple d'un extrait de base potentielle

Nous voyons ici des sinistres ouverts et clos avec les variables caractéristiques de ces sinistres (celles-ci pouvant être l'âge du souscripteur, la zone du sinistre, le type de bien assuré ou les caractéristiques de ce bien).

Le but est de prédire la charge ultime d'un sinistre (colonne de gauche) en fonction de toutes les autres variables. Le problème vient du fait qu'une variable n'est pas observée complètement.

La variable qui nous intéresse ici est la variable durée d'ouverture. Celle-ci représente la durée d'ouverture du dossier, c'est à dire l'intervalle de temps durant lequel le sinistre a été géré par un gestionnaire, de son ouverture à sa clôture.

Cette variable est généralement la plus explicative dans les modèles mis en place pour expliquer le coût final d'un sinistre. En effet, plus cette période est longue, plus le sinistre a été difficile à gérer, plus il a été complexe et donc générateur de coût.

Les sinistres possédant un tiret (case vide jaune) dans la figure sont toujours ouverts, en cours. Ils sont visualisés à la date t d'extraction de la base figure 20.

Pour le sinistre clos, il n'y a pas de problème, sa durée d'ouverture n reflète exactement sa durée d'ouverture réelle. L'utiliser à des fins de prédictions ne pose pas de problèmes.

Pour les sinistres 1 et 2, les durées n_1 et n_2 ne sont pas le reflet de ce qu'est la durée d'ouverture réelle du sinistre. Considérer n_1 et n_2 en tant que telles voudrait dire que l'on considère que les sinistres ont pour date de clôture définitive la date d'extraction de la base.

La prise en compte de la censure (à droite ici) revient à dire que ces durées n_1 et n_2 ne sont qu'un minimum et qu'il faut considérer la probabilité que ces dates évoluent dans le temps et pendant combien de temps.

Il n'est pas impossible que le date d'extraction soit effectivement la date de clôture d'un sinistre ouvert mais c'est relativement rare ou complètement fortuit et ne reflète pas la grande majorité des cas.

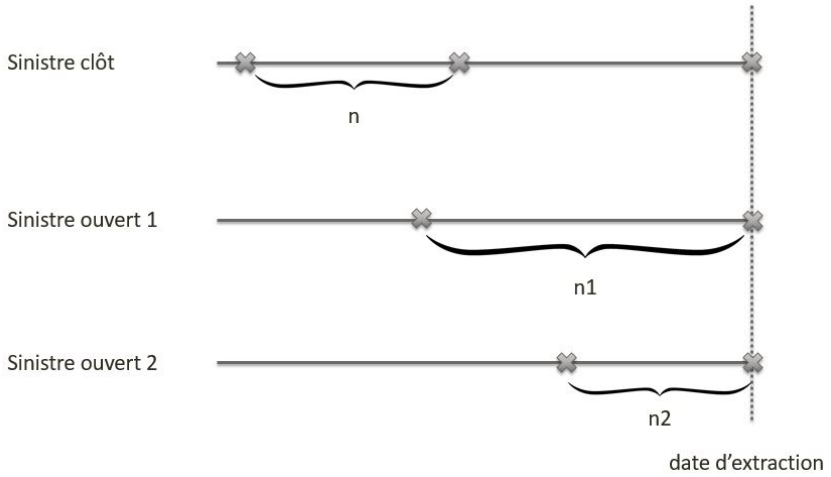

FIGURE 20 – Visualisation à date d'extraction

Considérons les figures 21 et 22.

Notre premier sinistre est clôt, il n'y a donc aucune différence entre sa charge ultime censurée et non censurée, sa durée d'ouverture est inchangée.

Le second sinistre a, de manière toute à fait exceptionnelle, une date de clôture qui correspond à la date d'extraction. Celui-ci est toujours considéré comme ouvert malgré le fait que sa date de clôture est égale à la date d'extraction, ceci peut être du à une latence dans la gestion des sinistres et dans le

traitement des dossiers. Lui aussi a une charge ultime inchangée et une durée d'ouverture identique, que nous soyons dans le cas censurée ou non censurée. Mais rappelons que ce cas est exceptionnel.

Le cas le plus fréquent concerne notre dernier sinistre. Extrait en t lors de l'extraction, sa date de clôture ne sera que plus tard, lui octroyant une durée d'ouverture de m_2, supérieure évidemment à la durée n_2.

Les versements liés à ce sinistre (brut de recours) ne peuvent qu'augmenter (au sens large, ils peuvent ne pas évoluer). Nous avons donc coût(m_2) $\geq$ coût(n_2) ainsi que la durée d'ouverture non censurée m_2 supérieure à la durée d'ouverture censurée n_2.

Ne pas prendre en compte cette évolution possible du dossier revient à sous-estimer les coûts des sinistres lors de la prédiction. Il faut en effet modéliser la probabilité de survie du dossier. Elle s'obtient en apprenant de la base de sinistres clos. Il faut estimer la survie d'un dossier.

Modéliser le fait qu'il reste ouvert ou non est crucial ici car c'est un facteur capital dans l'explication de la charge finale du sinistre. Ne pas prendre en compte cette censure revient à biaiser les estimations en les sous-estimant, chose très peu prudente en provisionnement.

En effet, les actuaires préfèrent prendre une marge de sécurité en ayant une estimation à la charge ultime légèrement supérieure à celle attendue plutôt que d'en avoir une inférieure.

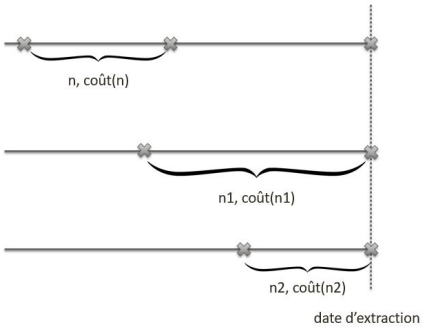

FIGURE 21 – Censure non prise en compte

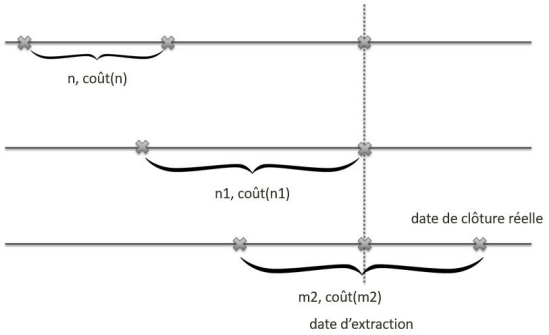

FIGURE 22 – Censure prise en compte

Afin de considérer la censure pour la variable représentant la durée d'ouverture du dossier sinistre, censurée à droite lors de l'extraction de la base, nous allons la modéliser en l'ajustant avec la durée de vie résiduelle du dossier.

Troisième partie
Etudes et résultats

L'étude menée dans ce mémoire résulte du souhait d'utiliser les méthodes citées précédemment sur des données réelles afin d'en tester la précision de prédiction et donc d'en mesure l'intérêt pratique. Le but est de savoir si ces méthodes peuvent avoir un intérêt à être mises en place de manière opérationnelle afin de contribuer à aiguiller la fonction actuarielle dans la compréhension des provisions dossier/dossier ou même améliorer ou parfaire les règles internes d'évaluation des dossiers sinistres propres à chaque assureur.

Le but est de mettre en place un outil permettant à la fonction actuarielle de pouvoir challenger les provisions dossier/dossier de manière efficace. Bien évidemment, l'idée de cette approche n'est pas de se substituer aux autres bonnes pratiques existantes au sein des services d'indemnisation qui permettent également d'apprécier le provisionnement dossier/dossier telles que le suivi de la liquidation des sinistres avec la surveillance des boni et des mali dossier/dossier ou le sondage direct d'un échantillon de dossier sinistres dans les services d'indemnisation afin de valider ou non la cohérence des provisions dossier/dossier estimées. L'outil que nous allons mettre en place à pour but d'aider la fonction actuarielle à pondérer son avis et à enrichir son analyse.

Aujourd'hui la fonction actuarielle n'est pas dotée d'algorithmes afin de challenger le coût moyen prédéfini à l'ouverture du dossier sinistre pour la $\text{PSAP}^{\text{dossier/dossier}}$ à l'ouverture. Généralement, seul ce coût moyen est attribué sauf éléments complémentaires importants. Ceci évolue lorsque de nouvelles informations sont disponibles lors du développement de la situation du sinistre. De nouvelles informations permettent une modulation de la provision et ce jusqu'à la clôture de celui-ci. La fonction actuarielle peut, comme précisé précédemment, à ce moment-là s'intéresser aux boni et aux mali associés aux dossiers et donc à cet instant déterminer si les provisions associées aux sinistres sont correctes ou si elles soulèvent des dysfonctionnements.

Les méthodes mises en place dans cette partie ont pour objectif d'être un support à la fonction actuarielle pour challenger les provisions dossier/dossier faites par les gestionnaires et éventuellement dans le futur un outil d'aide à la gestion pouvant permettre de revoir les forfaits associés aux sinistres (généralement des coûts moyens) en prenant en compte plus d'information afin de donner une indication plus précise. La mise en place des ces algorithmes au sein de la fonction actuarielle permettrait d'apporter un autre point de vue statistique qu'il conviendra de confronter aux autres analyses classiques de provisionnement dossier/dossier.

L'étude sera menée en deux étapes. La première sera une prédiction du coût final du sinistre à l'ouverture du dossier. La seconde sera à la dernière vue du sinistre, c'est-à-dire au moment de l'extraction de la base. Cette seconde étape prendra en compte la censure à droite [11] due à cette extraction (qui ne sera pas présente dans la première étape).

De manière schématique, nous allons comparer les résultats obtenus par le modèle des gestionnaires aux résultats obtenus par les modèles de Machine Learning, comme le résume la figure 23.

11. La censure à droite est induite via l'observation de la durée de vie du dossier sinistre. Pour les sinistres ouverts, cette durée n'est pas entièrement connue à la date d'extraction du sinistre. Nous n'observons que la durée d'ouverture jusqu'à extraction et non pas jusqu'à clôture. Il faut intégrer le fait que cette durée n'est pas entièrement observée, c'est ce que fait la censure. Ne pas prendre en compte viendrait à biaiser le modèle et sous estimer les provisions.

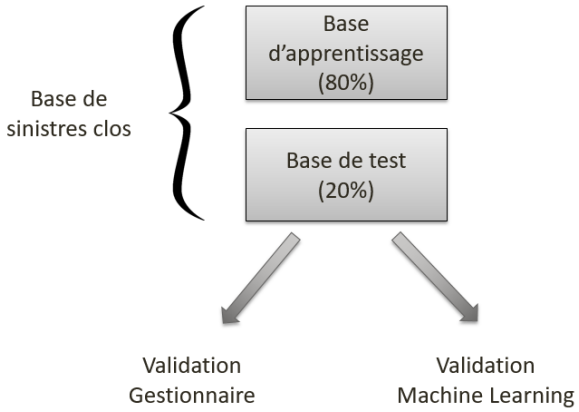

FIGURE 23 – Démarche de validation et de comparaison

Le risque étudié ici est une garantie incendie sur un portefeuille de particuliers. C'est une branche qui présente l'intérêt d'avoir aussi bien des sinistres attritionnels des sinistres gérés dans un temps relativement court mais aussi des sinistres plus graves qui eux s'étendent sur une période plus longue.

Généralement, les sinistres que nous allons étudier sont soumis à des contrats qui encadrent le risque d'incendie et d'évènements assimilés.

Voici une rapide description des termes que ces contrats peuvent contenir [12] : ces contrats garantissent les dommages matériels du bien assuré, les dommages corporels ainsi que la responsabilité civile (matérielle et corporelle). La garantie peut s'appliquer au contenu, au contenant ou aux deux simultanément. Ces dommages pouvant être causés par :

— un incendie
— une explosion ou une implosion
— des fumées
— les conséquences de la chute de la foudre
— l'action de l'électricité due à des perturbations ou surcharges sur le réseau
— le choc d'un véhicule terrestre à moteur
— le choc de la chute d'objets ou appareils de navigation aérienne
— les mesures de sauvetage et l'intervention des secours suite à un sinistre garanti
— des frais annexes (honoraires, pertes indirectes justifiées)

Ces conditions peuvent être soumises au bon entretien des évacuations et des conduits ainsi qu'éventuellement des conditions de débroussaillement.

Toutes les garanties portant sur ce qui est à l'extérieur du bien sont généralement des conditions particulières optionnelles .

Les remboursements peuvent intervenir éventuellement sous forme de franchises.

12. Ces informations sont génériques et non spécifiques aux contrats délivrés par la société pour laquelle cette étude a été menée.

I Description de la base

La base étudiée est en fait la fusion de quatre bases provenant de quatre services différents de l'entreprise qui nous permet d'utiliser ses données.

Ces bases sont les suivantes :

— une base client : celle-ci reprend les informations et caractéristiques de l'assuré
— une base bien : cette base reprend les informations concernant le bien assuré (zone, type)
— une base contrat : cette base reprend des caractéristiques propres au contrat (garantie, type de produit, mode de paiement, avenants)
— une base sinistre : elle reprend les informations liées au sinistre qui a eu lieu (cause, dates, coût)

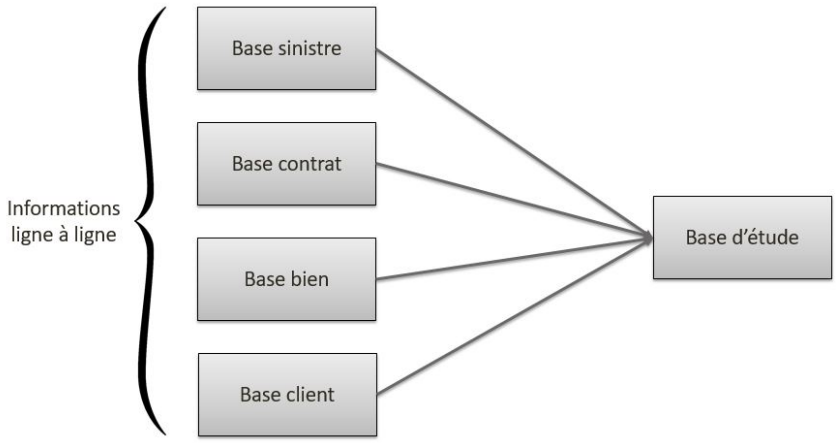

FIGURE 24 – Base d'étude

L'utilisation des algorithmes de machine learning nécessite une qualité de la donnée optimale : des données exhaustives, exactes et appropriées. Il a donc fallu vérifier la qualité de la donnée tout au long de la chaîne de traitement des bases sources jusqu'à la base finale.

Une fois ceci fait, il a fallu vérifier la complétude et la cohérence de la donnée.

Les sinistres suivants n'ont pas été pris en compte pour l'étude :

— Tout sinistre ayant une date de survenance antérieure à celle de la vérification de la base par l'entreprise
— Tout sinistre ayant une date de déclaration antérieure à sa date de survenance
— Tout sinistre ayant une date d'ouverture antérieure à sa date de déclaration
— Tout sinistre ayant une date de clôture antérieure à sa date d'ouverture
— Tout sinistre ayant une charge ultime négative
— Tout sinistre impactant un contrat d'un souscripteur considéré à la fois comme professionnel et un particulier
— Les sinistres survenus lors de périodes non couvertes par le contrat
— Les sinistres ayant une estimation à l'ouverture négative

Concernant le traitement des variables, ont été écartées les variables :

— à une seule modalité
— renseignées à moins de 75%
— reprenant une caractéristique personnelle du souscripteur
— redondantes

Nous sommes conscients que ces exclusions sont susceptibles de pouvoir peser sur la prédiction potentielle des algorithmes. En effet, elles pèsent pour 2% en nombre et un peu plus d'un pourcent en charge. Cette perte a toutefois été jugée relativement acceptable du fait qu'il restait un nombre pertinent de données à utiliser pour faire l'étude. Dans l'idéal, il aurait fallu mener des tests de sensibilité afin de vérifier l'absence de biais matériel introduits dans les résultats. Une autre solution aurait été de corriger les données aberrantes matérielles en échangeant avec le service d'indemnisation. Ceci n'a pas été fait ici et constitue un axe d'amélioration de l'étude.

La qualité des données reste aujourd'hui un des axes de progrès des assureurs.

Pour utiliser ces algorithmes, il est nécessaire d'avoir les données les plus fiables possibles et il n'est pas envisageable d'utiliser des variables insuffisamment renseignées, mal renseignées ou incohérentes, à moins de les corriger. La nécessité d'utiliser des données ligne à ligne nous oblige à être intransigeant sur leur qualité et leur complétude.

Notre base de sinistres d'étude est constituée de 5% de sinistres ouverts.

I.1 Prise en compte de l'inflation

Les sinistres sur lesquels nous allons faire porter notre étude peuvent s'écouler sur plusieurs années ou s'être déroulés il y a plus d'une dizaine d'années. Afin de garder une cohérence en terme de montants ils ont tous été inflatés afin d'avoir une vision en euros au 31 décembre 2016, date d'extraction de la base.

Une caractéristique importante à considérer est le type de dommage subi lors du sinistre. En effet, un sinistre impactant un bien pur ne sera pas valorisé de la même manière qu'un sinistre impactant une personne. Il a donc fallu indexer la base sur deux inflations différentes : l'une pour les sinistres matériels, l'autre pour les sinistres corporels.

Dans cette étude nous ne prenons en compte que l'inflation passée afin d'indexer tous les montants à date de vue égale à la date d'extraction, soit le 31 décembre 2016. Quelques sinistres pouvant être remboursés quelques années après la date de vue (dans le cas d'un accident grave par exemple), nous aurions pu considérer l'inflation future ainsi que les cadences de règlements possibles et donc annexer nos coûts avec une inflation future. Ceci n'a pas été fait ici et représente un axe d'amélioration de l'étude.

Sinistre Matériel Comme précisé précédemment, ces sinistres impactent un bien. Nous avons donc pris en compte le taux d'inflation de la FFB [13] (Fédération Française du Bâtiment). Le taux étant trimestriel il a été choisi, afin de simplifier les calculs, d'inflater à partir des taux du second trimestre de chaque année jusqu'au taux du dernier trimestre de l'année 2016.

13. http://www.ffbatiment.fr/federation-francaise-du-batiment/le-batiment-et-vous/en_chiffres/indices-index/Chiffres_Index_FFB_Construction.html

La prise en compte de l'inflation a été faite de la manière suivante :

$$\text{Charge Ultime}^{\text{inflaté}}_{\text{matériel}} = \text{Charge Ultime}_{\text{matériel}} \times \frac{\text{Taux FFB 2016}_{\text{T4}}}{\text{Taux FFB Année de Clôture}_{\text{T2}}}$$

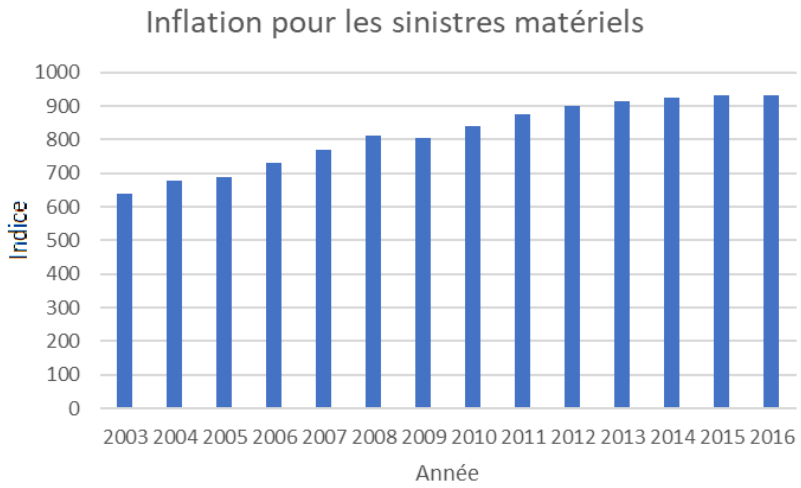

FIGURE 25 – Indice d'inflation FFB pour les sinistres matériels

Sinistre Corporel Pour ce qui concerne les sinistres corporels, cela n'avait aucun sens de les indexer sur l'évolution du coût des bâtiments. Nous avons donc pris le taux d'évolution de la SCOR qui a publié des indices d'évolution pour les sinistres corporels graves.

Ces indices étant réévalués annuellement, le calcul est le même fait que pour l'inflation des sinistres matériels :

$$\text{Charge Ultime}^{\text{inflaté}}_{\text{corporel}} = \text{Charge Ultime}_{\text{corporel}} \times \frac{\text{Taux SCOR 2016}}{\text{Taux SCOR Année de Clôture}}$$

A l'instar de l'inflation matérielle qui a stagné sur les 3 dernières années (+17 points avec une baisse entre 2015 et 2016, soit 2%), l'inflation pour les sinistres corporels ne cesse de grimper (+ 36 points, soit 10%). Sur les dix dernières années le taux matériel a quant à lui progressé de 29% tandis que le taux corporel a plus que doublé (augmentation de 108%).

La distinction entre sinistre matériel et corporel est donc fondamentale car les évolutions inflationnistes au cours du temps sont très différentes.

I.2 Sinistralité du portefeuille

Intéressons nous désormais à la sinistralité de notre portefeuille. Nous allons regarder l'évolution du nombre de sinistres et de leur coût au fur et à mesure du temps.

Le nombre de sinistres par année, présenté sur la figure 27, représente le nombre de sinistres déclarés sur l'année considérée. Il inclut donc les dossiers qui s'avèrent sans suite plus tard ainsi que les dossiers

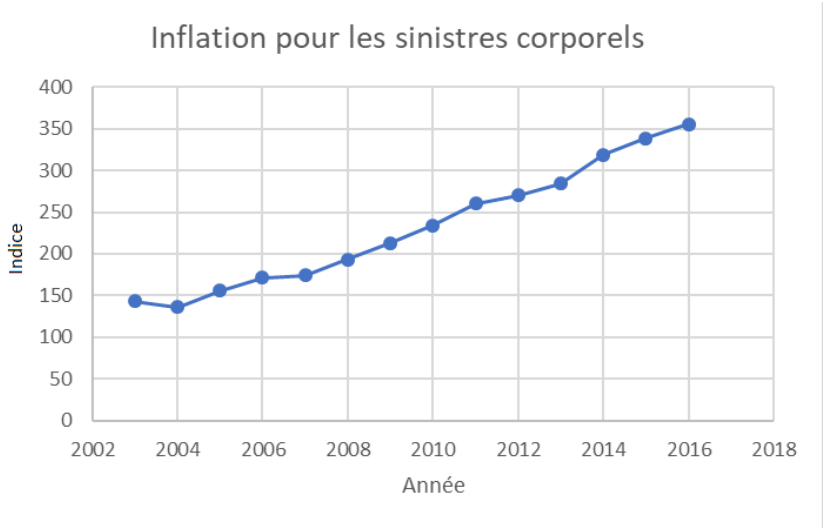

FIGURE 26 – Indice d'inflation SCOR pour les sinistres corporels

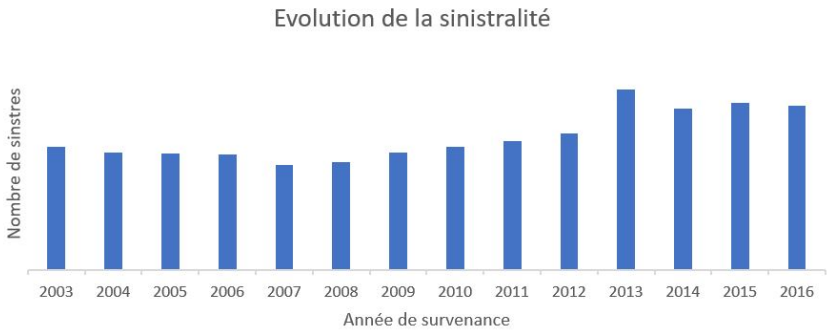

FIGURE 27 – Nombre de sinistres déclarés par année de survenance

qui donneront lieu à une indemnisation. Ce graphique ne prend pas en compte les tardifs qui auraient pu être estimés à la date d'extraction. Il est à noter que ce nombre de sinistres est brut, il n'est pas indexé sur les fluctuations de la taille du portefeuille (évolution de l'exposition au risque).

Le nombre de sinistres reste relativement constant de 2003 à 2012, avec une année plus faible en 2007. La sinistralité grimpe de +31% entre 2012 et 2013 (année la plus catastrophique) avant de se stabiliser sur les trois dernières années. L'année 2016 représentant une sinistralité 8,5% moins importante qu'en 2013.

Ce sursaut de sinistralité entre 2012 et 2013 ne saurait s'expliquer qu'avec les données de la base finale. En effet, cette année ne correspond pas à un grand évènement identifié par l'entreprise. Il pourrait s'agir du rachat d'un portefeuille ou alors d'une augmentation drastique du nombre d'assurés.

La figure 28 considère la charge ultime pour les sinistres clos réglés en fonction de leur année de survenance. La charge correspond à la somme de la provision du dossier et des règlements faits pour le sinistre. Hors lorsqu'un sinistre est clos, sa provision est nulle. Sa charge ultime correspond donc à la somme des règlements effectués pour ce sinistre. C'est ce qui est représenté ici. Celle-ci est brute de recours mais inflatée à date de vue, en *As if*, au 31 décembre 2016.

Le coût par an est bien plus variable en revanche.

Une étude plus poussée entre le coût de la sinistralité figure 28 et le nombre de sinistres clos tous les ans 29 serait une piste d'amélioration afin de mieux comprendre la gestion et l'évolution du portefeuille de sinistres.

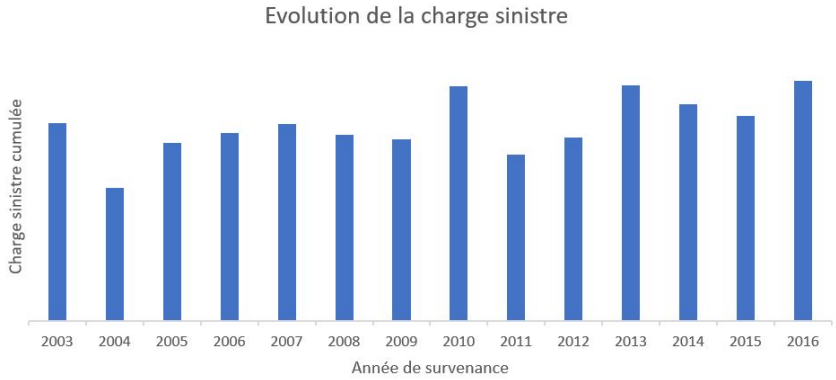

FIGURE 28 – Représentation de la charge sinistre par année de survenance

Nous voyons bien que malgré une sinistralité relativement constante entre 2003 et 2012, le coût cumulé des sinistres entre ces années est bien plus variable : 76% de différence entre 2004 et 2010.

Le coût de l'année 2013 est quant à elle quasiment identique à l'année 2010 (+0.3% de coût supplémentaire) avec une sinistralité en fréquence bien plus importante.

Bien évidemment, ces coûts ne sont pas définitifs car ils ne prennent pas en compte les potentiels coûts futurs rattachés aux sinistres toujours ouverts qui pourraient se répercuter sur ces années de survenance. Mais nous voyons donc que la charge cumulée par année de survenance n'a pas de relation claire avec le nombre de polices sinistrées (exemple entre 2005 et 2009 avec les courbes de nombre et de coûts qui sont inversées).

On remarque tout de même que les fluctuations du nombre de sinistres et de leur charge totale a tendance à être plus stable sur les trois dernières années.

Ceci peut être surprenant à la vue de la gestion des dossiers.

En effet, le nombre de dossier clos ne suit pas la tendance de survenance des sinistres. Il n'est pas surprenant de voir que la survenance, la déclaration et l'ouverture des dossiers soient proches. La clôture des dossiers est assez naturellement décalée de ces dates-là. Nous remarquons tout de même que sur notre portefeuille nous avons eu une période de 2006 à 2013 où toutes ces évolutions sont restées très proches.

Cette scission à partir de 2013 s'explique car ces courbes ne prennent en compte que les dossiers clos. La période de gestion moyenne d'un dossier étant inférieure à un an sur ce portefeuille, nous pouvons donc intuiter que dernièrement, malgré le nombre d'affaires croissant depuis 2013, la société semble être dans une gestion plus dynamique des dossiers.

Les différents coûts des sinistres sont eux aussi importants. Leur visualisation permet d'établir un

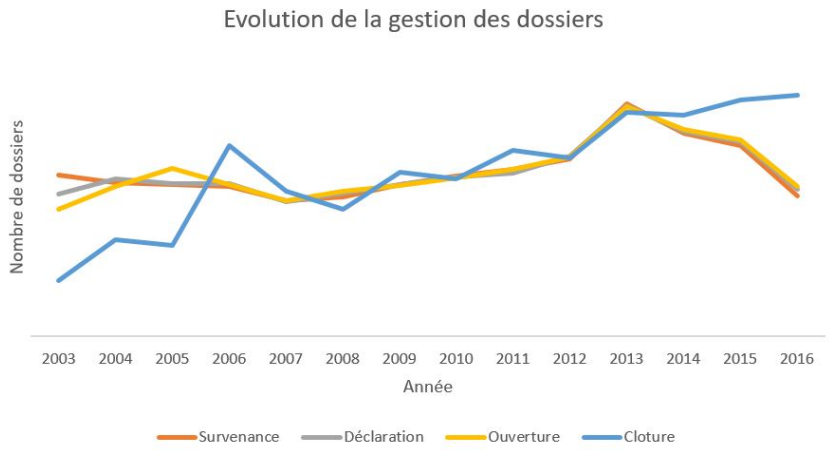

FIGURE 29 – Evolution des dossiers clos

seuil entre les sinistres attritionnels et graves.

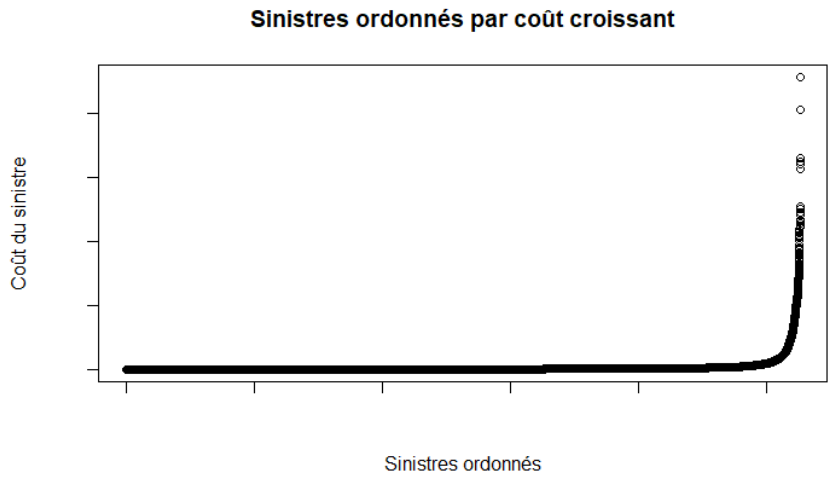

FIGURE 30 – Représentation de l'envergure des coûts des sinistres du portefeuille

Nous voyons clairement qu'une grande majorité des sinistres est en dessous d'un certain seuil.

Nous voyons ce saut en coût de manière plus précise en regardant les différents quantiles.

Il serait possible de la déterminer de manière plus précise (avec un la théorie des valeurs extrêmes via un mean excess plot ou avec l'estimateur de Hill) mais ceci ne nous intéresse pas ici car les algorithmes fonctionnent sur la base entière.

Nous apprenons tout de même que nous avons une ampleur de coût de sinistre assez importante. Le fait que notre base de données possède tout type de sinistres (attritionnel et grave) est un atout pour la prédiction. En effet, le machine learning ne pourra prédire des sinistres graves que s'il y en a dans la base d'apprentissage.

Un axe d'amélioration serait de considérer la sensibilité des résultats aux deux dernières années en

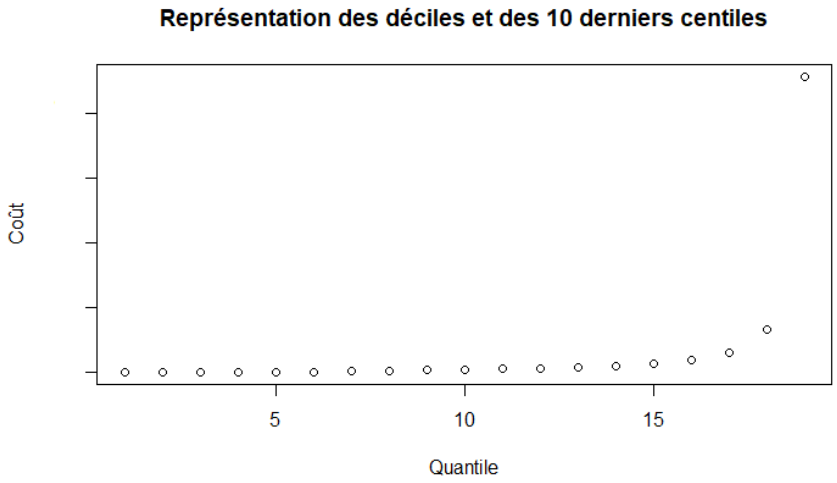

FIGURE 31 – Visualisation du saut de coût

cours. En retirant les deux dernières années de survenance observées, nous pourrions comparer les résultats obtenus sur la prédiction et voir si cela impacte les résultats.

Par la suite, le terme gestionnaire désignera soit le gestionnaire sinistre, soit le service indemnisation ou bien le système de gestion.

II Etude à l'ouverture du dossier

Lorsqu'un dossier sinistre est ouvert, le gestionnaire reçoit les informations relatives à celui-ci et doit faire une estimation de son coût. C'est la PSAP$^{\text{dossier/dossier}}$ à l'ouverture.

Nous allons donc, avec les informations contenues dans les différentes bases recueillies, comparer les estimations à l'ouverture faites par les gestionnaires sur les sinistres avec les estimations faites par les algorithmes de machine learning étant donné l'information connue à l'ouverture.

II.1 Données à l'ouverture et sélection de variables

L'ensemble des données à l'ouverture d'un dossier reprend un certain nombre d'informations. Afin de mener l'étude la plus complète possible il a été nécessaire de faire fusionner des bases d'informations provenant de différents services.

Suite au tri des variables, il ne reste dans la base que celles figurant dans le tableau suivant.

Nous nous retrouvons donc avec 37 variables, 5 de la base client, 14 de la base contrat, 7 de la base de bien et 11 de la base sinistre. De ces données, 5 autres en ont été extraites.

Par la suite nous nommerons Delta 1 la variable *Temps entre la survenance et la déclaration du sinistre* et Delta 2 la variable *Temps entre la déclaration et l'ouverture du sinistre*.

Données à l'ouverture			
Base Client	Base Contrat	Base Bien	Base Sinistre
Sexe	Cotisation annuelle	Qualité de l'occupant	Charge Ultime
Situation familiale	Capacité mobilière	Code résidence	Garantie sinistrée
Situation professionnelle	Capacité de vol	Pièces principales	Date de survenance
Code postal	Fractionnement du contrat	Surface de dépendance	Date de déclaration
Catégorie d'adresse	Taux dégressif	Surface de développement	Date d'ouverture
	Mode de paiement	Type de bien	Evènement
	Options du contrat	Zone tarifaire	Fait générateur
	Avenants		Cause
	Sinistres antérieurs		Type de dommage
	Conditions générales		Taux de responsabilité
	Type de prix		Situation du sinistre
	Produit		
	Code technique		
	Type de réassurance		

TABLE 2 – Information contenues dans la base à l'ouverture du dossier

Données extraites
Année de survenance
Année de déclaration
Année d'ouverture
Temps entre la survenance et la déclaration
Temps entre la déclaration et l'ouverture

TABLE 3 – Variables déduites de la base à l'ouverture

Nous possédons de plus un indicateur précisant si le sinistre est ouvert ou non (mais il ne fait pas parti de nos variables explicatives).

Plus nos données sont corrélées moins la régression faite est fiable. Nous calculons le V de Cramer afin d'avoir une représentation des liens entre les variables. Celui-ci a été obtenu à l'aide du logiciel *Addactis Pricing®* [14].

A la vue des corrélations entre nos variables (un certain nombre au dessus de 30% et d'autres extrêmes au dessus de 90%), nous allons faire une sélection à l'aide d'un modèle linéaire complet et d'une forêt aléatoire .

Ces méthodes de sélection de variables seront faites sur toutes les variables mentionnées précédemment sauf sur les dates (seules les années sont conservées). L'indicateur de clôture du sinistre sera lui aussi mis de côté.

Les variables retenues par les différents modèles sont listées dans la table 4.

Le modèle AIC sélectionne 15 variables, 9 sont sélectionnées par le modèle BIC et 18 par la forêt

14. Logiciel de tarification de la gamme de logiciels Addactis, propriété d'Addactis Group

Variables sélectionnées		
Modèle linéaire (AIC)	Modèle linéaire (BIC)	Forêt Aléatoire
Delta 1	Delta 1	Delta 1
Delta 2	Delta 2	Delta 2
Cause du sinistre	Cause du sinistre	Cause du sinistre
Type de dommage	Type de dommage	Fait générateur
Taux de responsabilité	Taux de responsabilité	Code Postal
Cotisation annuelle	Nombre de sinistres	Cotisation annuelle
Capacité vol	Type de prix	Capacité Vol
Nombre de pièces principales	Nombre de pièces principales	Nombre de pièces principales
Nombre de sinistres	Type de bien	Garantie sinistrée
Type de prix		Fractionnement du contrat
Code Résidence		Code Réseau
Nombre d'avenants au contrat		Nombre d'avenants au contrat
Type de Bien		Conditions générales
Qualité de l'occupant		Qualité de l'occupant
Produit		Situation professionnelle
		Surface de dépendance
		Situation familiale
		Zone tarifaire

TABLE 4 – Variables retenues par les différentes méthodes de sélection à l'ouverture

aléatoire. A noter que la forêt aléatoire classe toutes les variables par ordre d'importance, elle ne sélectionne pas explicitement de variable. Mais il est possible de représenter graphiquement l'importance des variables et un saut notoire était visible entre les 18 premières variables et celles restantes.

Pour la sélection de modèles linéaires via l'AIC et BIC, les méthodes *forward*, *backward* et *stepwise* ont été utilisées. Seules les variables du meilleur modèle pour chaque méthode ont été conservées.

Concernant la méthode de la forêt aléatoire, elle a été utilisée afin de voir quelles variables étaient les plus significatives lors de la création d'arbres, celles qui scindaient le mieux le risque et donc les plus importantes.

Il est a noter que les dates de survenance, de déclaration et d'ouverture des sinistres apparaissent aussi dans les modèles mais elles sont conservées automatiquement à des fins opérationnelles.

Les variables triées par nombre d'apparitions dans les modèles se trouvent dans la table 5.

Le nombre de variables qui apparaissent dans tous les modèles est au nombre de 4, 10 variables apparaissent dans deux des modèles alors que 10 apparaissent dans un seul modèle.

Nous décidons de ne garder que les variables apparaissant dans au moins deux modèles.

Nous faisons aussi le choix de conserver la variable Code postal ainsi que la variable Zone tarifaire car celles-ci sont classées respectivement 3^{è} et 7^{è} dans l'ordre d'apparition des segmentations faites par la forêt aléatoire.

Apparition dans les modèles		
Tous les modèles	2 modèles	1 seul modèle
Delta 1	Type de dommage	Produit
Delta 2	Taux de responsabilité	Fait générateur
Cause du sinistre	Capacité de vol	Code postal
Nombre de pièces principales	Cotisation annuelle	Garantie sinistrée
	Nombre d'avenants	Fractionnement du contrat
	Qualité de l'occupant	Conditions générales
	Code Résidence	Situation professionnelle
	Nombre de sinistres	Surface de dépendance
	Type de prix	Situation familiale
	Type de bien	Zone tarifaire

TABLE 5 – Apparition des variables dans les modèles de sélection à l'ouverture

Nous nous retrouvons donc avec 19 variables : les quatorze présentes dans au moins deux modèles, auxquelles s'ajoutent les deux variables récupérées par la forêt aléatoire ainsi que les trois dates.

Les variables et leurs types sont référencés dans la table 6.

Variables conservées		
Variable	Type	Modalités
Date de survenance	Entier	
Date de déclaration	Entier	
Date d'ouverture	Entier	
Cause du sinistre	Catégorielle	23
Delta 1	Continue	
Delta 2	Continue	
Nombre de pièces principales	Entier	
Type de dommage	Catégorielle	2
Taux de responsabilité	Continue	
Capacité de vol	Continue	
Cotisation annuelle	Continue	
Nombre d'avenants	Entier	
Qualité de l'occupant	Catégorielle	17
Code Résidence	Catégorielle	4
Nombre de sinistres	Entier	
Type de prix	Catégorielle	2
Type de bien	Catégorielle	5
Code Postal	Entier	
Zone Tarifaire	Entier	

TABLE 6 – Résumé des variables conservées à l'ouverture

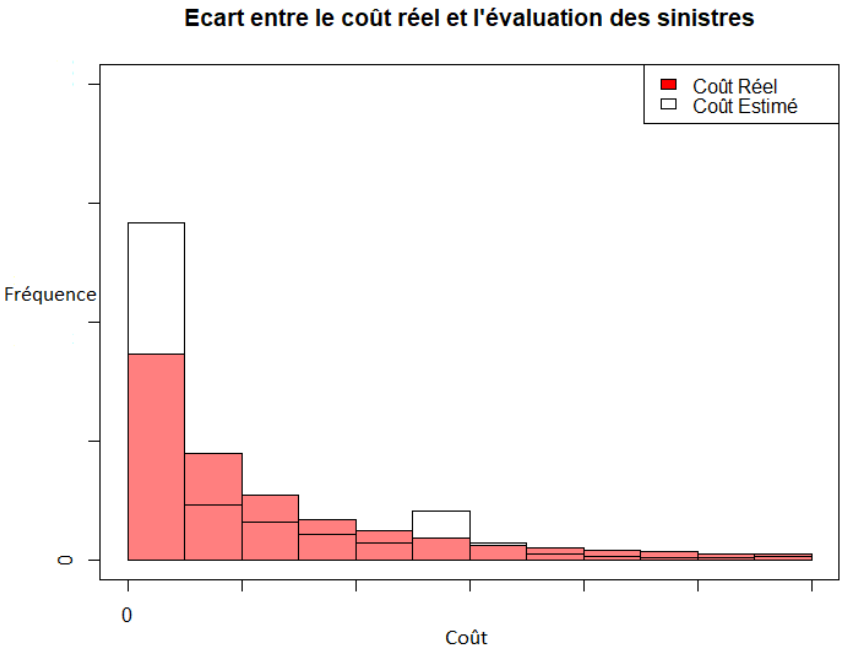

FIGURE 32 – Visualisation de l'écart entre l'estimation des coûts à l'ouverture et des coûts réels à l'ouverture

II.2 Postulats

Les gestionnaires de sinistres font des postulats sur le coût final des dossiers, à leur ouverture, desquels ils ont reçu les divers éléments.

Leur but est d'approcher de la manière la plus précise possible le coût final du sinistre afin que la somme provisionnée, la PSAP$^{\text{dossier/dossier}}$ à l'ouverture, soit la plus proche possible de la charge ultime, c'est-à-dire le coût effectivement versé par l'organisme d'assurance pour ces différents sinistres.

Dans notre portefeuille, l'estimation n'est pas aussi précise que nous pourrions le souhaiter. Sur la figure 32, nous pouvons visualiser les écarts entre les estimations faites des sinistres à l'ouverture et les coûts réels des sinistres. Nous voyons une surestimation des coûts les moins élevés et une sous estimation quasiment systématique des coûts plus élevées.

Ces deux pics, sur l'histogramme 32, pourraient être considérés comme des forfaits de sinistres à leur ouverture. Si c'est le cas, nous pourrions en faire une nouvelle estimation ou établir des seuils différents.

Une autre représentation se trouve figure 33 où nous voyons que l'estimation est soit juste, soit en dessous du coût réel. Il est rare de voir l'estimation du sinistre dépasser celle du coût réel. Il y a donc une réelle progression possible.

A noter que pour les sinistres extrêmes, nous pouvons voir que certains ont été anticipés alors que d'autre absolument pas.

De manière chiffrée, sur l'ensemble de ce portefeuille, seul 66% du risque a été provisionné lors de l'ouverture de dossier. Ces données justifient donc l'étude que nous menons.

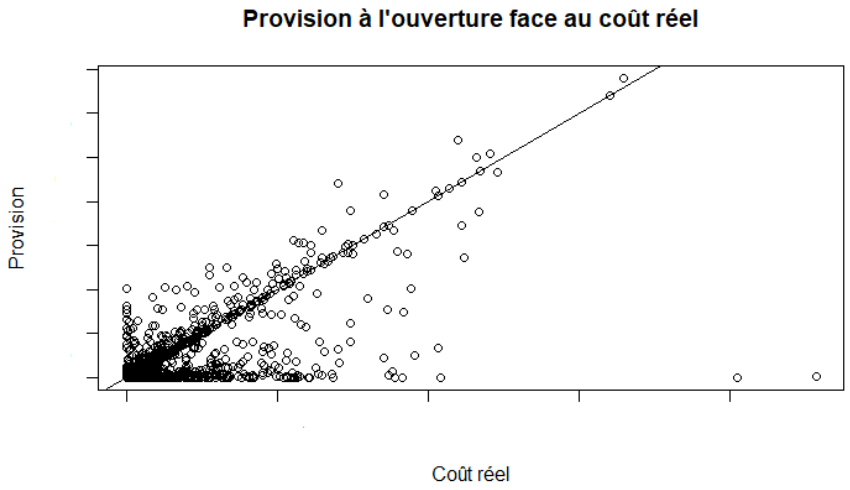

FIGURE 33 – Graphique représentant la provision par rapport au coût réel à l'ouverture

On remarque tout de même que de manière générale, la charge sinistre est surestimée pour la majeure partie des sinistres.

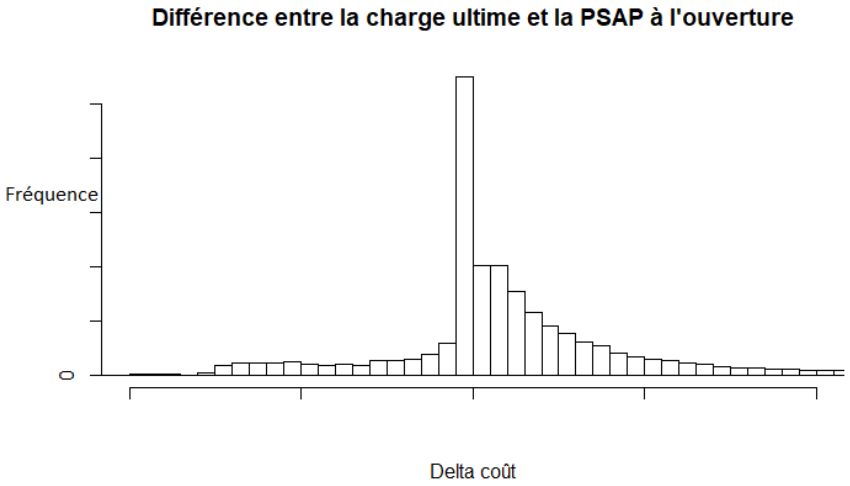

FIGURE 34 – Ecart entre la charge ultime et la provision à l'ouverture

II.3 Modèles

Notre but est de mettre en place des modèles qui vont tenter de sur-performer, que ce soit de manière locale ou globale, l'estimation du gestionnaire. Si certaines estimations sont plus proches du coût réel des sinistres, alors ces modèles pourront permettre à la fonction actuarielle d'échanger avec les gestionnaires afin de leur permettre d'améliorer leurs règles de gestion.

Nous avons tout d'abord scindé notre base de sinistres en deux afin de créer une base d'apprentissage représentant 80% de la base de sinistres clos. Les 20% restants forment notre base de test. C'est en fonction de la précision des modèles sur cette base de test que nous choisirons le modèle pour faire une prédiction sur les sinistres en cours.

Il est important de vérifier que la base d'apprentissage et la base de test possèdent les même propriétés statistiques (conservation de la moyenne, de la médiane ainsi que des minimums et des maximums proches). En effet, les algorithmes apprennent des données. Si les données n'existent pas dans la base d'apprentissage mais qu'elles sont dans la base de test alors il sera impossible de les prédire correctement.

Il est aussi important de mélanger sa base. En effet, il faut à tout prix éviter une structure implicite des données qui soit due à leur enregistrement dans la base de données. Par exemple, si les données sont enregistrées en fonction de la date de survenance, alors une tendance dans les données pourrait s'instaurer et se répercuter sur la précision des modèles si l'on prenait les 80 premiers pourcent de la base (données les plus anciennes) pour tester sur les 20 derniers pourcent (données les plus récentes).

Les algorithmes ne doivent prendre que l'information des variables explicatives et ne doivent pas subir une structure imposée par la personne qui renseigne les données.

Afin de savoir si notre modèle apporte une amélioration à la prédiction faite par le gestionnaire nous allons nous baser sur deux indices : l'erreur quadratique moyenne (MSE, *Mean Squared Error*) et la prédiction de la charge ultime globale.

Ces deux indices pour les estimations faites par les gestionnaires sont les suivantes :

Résultats sur la base de test		
Méthode	MSE	Prédiction de la charge totale
Gestionnaire	402 345	0,670

TABLE 7 – Résultats de la méthode Gestionnaire à l'ouverture

Nous voyons donc qu'une amélioration sur la prédiction de la charge totale ne semble pas impossible. Le MSE nous donne une indication quand à la proximité de l'estimation faite avec le coût réel.

Le premier modèle que nous considérons est l'arbre de décision CART. Nous avons tout d'abord regardé les résultats que nous obtenions via un arbre élagué. Ensuite nous avons eu recours au bagging afin d'obtenir un modèle plus robuste. Enfin, nous avons modifié le paramètre de scission pour qu'il prenne en compte l'entropie et non plus le critère de Gini. Tous les arbres ont été élagué en regardant l'évolution de l'erreur en fonction de la taille des branches.

Les résultats sont les suivants :

Résultats sur la base de test		
Méthodes	MSE	Prédiction de la charge totale
CART - Elagué	526 976	1,038
CART - Bagging	524 909	1,042
CART - Entropie	519 068	1,039

TABLE 8 – Résultats des méthodes CART à l'ouverture

Nous voyons ici que le modèle est moins bon que la prévision des gestionnaires lorsque l'on s'inté-

resse aux sinistres un par un. Mais l'arbre de décision réussi à généraliser le processus et à donner une estimation de la charge globale bien plus proche que celle des gestionnaires.

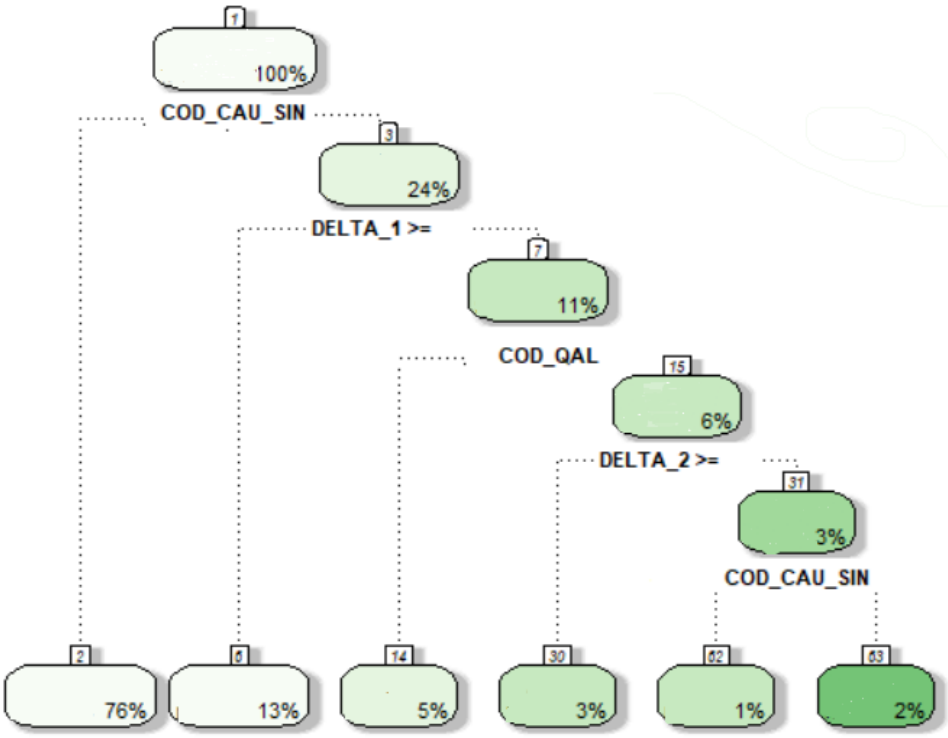

FIGURE 35 – Visualisation des variables importantes avec l'arbre élagué à l'ouverture

Nous pouvons voir sur la figure 35 que seules quatre variables sont nécessaires afin d'obtenir une segmentation des sinistres : la cause du sinistre, l'intervalle de temps entre la survenance et la déclaration du sinistre, la qualité de l'occupant ainsi que l'intervalle de temps entre la déclaration du sinistre et son ouverture. Nous voyons donc qu'en plus des variables liées au sinistre, au bien, au contrat et à l'assuré, le temps de gestion de l'organisme d'assurance reflète d'une certaine manière la gravité (ou non) du sinistre. Celle-ci apparaît en tant que deuxième et quatrième variable la plus segmentante avec cette méthode.

Sur les nœuds de l'arbre où la cause du sinistre ou le code qualité est la variable segmentante, nous devrions avoir le choix que fait l'arbre à ce moment là. De la même manière, lorsque les variables Delta 1 et Delta 2 sont les variables de segmentation nous avons à nous poser la question : Delta 1 est-il supérieur ou égal à 3,5 et si ce n'est pas le cas alors dans le sous arbre suivant alors la question Delta 2 est-il supérieur à 0,5 pourrait intervenir. Ces choix n'ont pas été mentionnés sur l'arbre car ils alourdissent la lecture et la compréhension de l'arbre.

Nous voyons donc que pour un assureur qui voudrait mettre cette méthode en place, nul besoin d'aller chercher pléthore de variables et des informations dans des bases utilisées par d'autres services.

Il est a noter que la cause du sinistre est la variable la plus segmentante, et qu'elle apparaît deux fois ! Il est donc nécessaire pour l'organisme d'assurance de pouvoir déterminer et catégoriser de manière claire les causes de sinistres.

Variable qui dans notre portefeuille n'est pas remplie de manière toujours efficiente. En effet, cette variable à été conservée malgré le fait qu'elle ne soit pas entièrement renseignée (légèrement plus de 75%) et qu'une modalité représentative de cette variable soit : cause inconnue.

Il est donc impératif d'avoir plus de renseignements sur cette variable afin de pouvoir estimer le plus précisément possible la provision à l'ouverture.

En effet, les trois méthodes prédisent une charge globale avec au plus 4,2% d'erreur alors que les gestionnaires faisaient une sous estimation de la charge globale de 33%. Les arbres de décisions peuvent donc aiguiller la fonction actuarielle, et donc les gestionnaires, quant à la charge globale des sinistres qu'ils gèrent.

Le MSE est bien plus élevé pour les modèles d'arbres de décision même s'il diminue avec le bagging ainsi qu"avec l'utilisation de l'entropie comme critère d'information.

Exemple : Prenons un exemple afin de visualiser le processus de sélection fait par l'arbre. Lorsque le sinistre survient, nous avons à notre disposition toutes les variables mises en avant lors de la sélection de variable. L'arbre nous montre que nous n'avons qu'à sélectionner les variables Cause du sinistre, Delta 1, la qualité de l'occupant ainsi que Delta 2 pour avoir une prédiction du coût du sinistre.

Prenons les sinistres de la table 9 [15].

Les sinistres 1, 4 et 5 ont des causes de sinistres qui respectent le premier critère de segmentation de l'arbre. Ils sont donc directement mis dans la feuille de gauche et comme ayant un coût de sinistre de 2 399.

Les sinistres 2 et 3 tombent quant à eux dans la feuille de droite et arrivent à un nouveau nœud. Le sinistre 3 a bien un Delta 1 supérieur à 3,5, il part donc dans la feuille terminale de gauche et lui est attribué un coût de sinistre de 4 882.

Seul reste le sinistre 2 qui passe au nœud suivant. Sa qualité d'occupant ne permet pas de finir la régression. Il tombe dans la feuille de droite et arrive à un nouveau nœud. Si sa qualité d'occupant avait correspondu à la segmentation alors un coût de 14 000 lui aurait été attribué. Comme ce n'est pas le cas nous nous intéressons à son Delta 2 afin de savoir s'il est supérieur à 0,5. Son Delta 2 est nul, le processus continue. Encore une fois nous nous intéressons à la cause du sinistre. Celle-ci fait bien partie de la segmentation et le sinistre 2 part donc dans la branche de gauche où lui est attribué un coût de sinistre de 30 000.

Afin d'augmenter la robustesse de ces modèles et d'obtenir de meilleurs résultats nous utilisons la méthode des forêts aléatoires. Rappelons que pour cette méthode il est nécessaire de déterminer deux paramètres : la taille de la forêt (nombre d'arbres) ainsi que le nombre de variables m conservées à chaque nœud. Nous avons suivi la règle tacite expliquant que les meilleurs résultats pour la régression s'obtiennent en prenant m égal au nombre de variables explicatives divisé par 3, ce qui représente ici m égal à 6.

15. Ces exemples sont fictifs et ne représentent pas les données réelles du portefeuille de la société pour qui cette étude est menée

Exemple de sinistres				
Sinistre	Cause Sinistre	Qualité de l'occupant	Delta 1	Delta 2
Sinistre 1	Feu de cheminée	Propriétaire occupant total	27	0
Sinistre 2	Court Circuit	Nu propriétaire	0	0
Sinistre 3	Criminelle	Copropriétaire non occupant	5	0
Sinistre 4	Explosion	Usufruitier	1	1
Sinistre 5	Foudre	Locataire occupant partiel	4	95

TABLE 9 – Exemple de sinistres à l'ouverture

Nous décidons donc de construire les modèles de forêts aléatoires avec m allant de 2 à 6 afin de regarder l'évolution de nos indicateurs.

En ce qui concerne le nombre d'arbres utilisés, nous avons regardé l'évolution de l'erreur en fonction de la taille de la forêt et pris soit la valeur qui minimisait l'erreur, soit un nombre d'arbres permettant une erreur proche de l'erreur minimale asymptotique (un léger compromis a du être fait entre un gain marginal de l'erreur et le fait de faire exploser le nombre d'arbres, et ce pour des raisons de temps d'exécution).

Les résultats sont visibles dans le tableau suivant :

Résultats sur la base de test		
Méthodes	MSE	Prédiction de la charge totale
RF - m=2	422 642	1,058
RF - m=3	433 923	1,073
RF - m=4	439 331	1,086
RF - m=5	453 383	1,101
RF - m=6	456 466	1,114

TABLE 10 – Résultats des méthodes de forêts aléatoires à l'ouverture

Avec une certaine surprise nous voyons que c'est la méthode avec le nombre de variables sélectionnées à chaque nœud le plus petit qui nous offre le meilleur modèle.

Les modèles des forêts aléatoires offrent un MSE beaucoup plus proche de la méthode des gestionnaires que les arbres de décision mais font une prédiction de la charge totale légèrement plus élevée. Cette prédiction oscille entre 5,8% et 11,4% de plus que la réalité, ce qui est plus que les arbres de décision mais toujours plus précis que les gestionnaires.

Nous voyons qu'en augmentant m dans les modèles, les deux indices montrent que les modèles se détériorent.

Ce premier modèle avec m égal à deux nous donne de l'espoir dans le fait de pouvoir aiguiller la fonction actuarielle dans la compréhension des provisions dossier/dosser et éventuellement aider les gestionnaires car les prédictions qu'il fait sont similaires (en terme d'erreur) à celles faites par les gestionnaires, avec une erreur globale bien plus faible.

Cette amélioration s'explique car les forêts aléatoires sont beaucoup plus robustes que les arbres de

décision et qu'elles ont un meilleur pouvoir de généralisation. Cette méthode sur-apprend moins que les arbres simples.

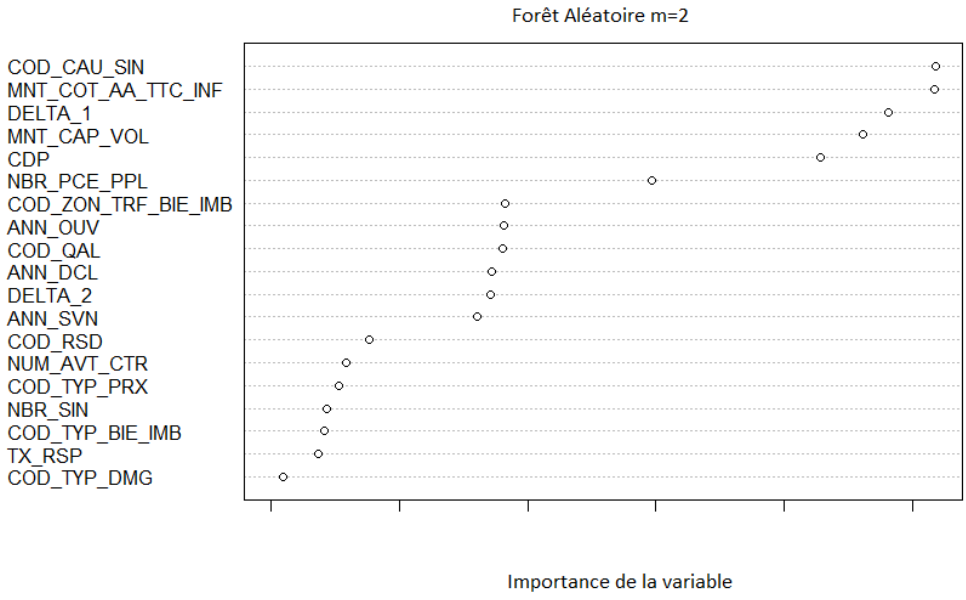

FIGURE 36 – Importance des variables pour la forêt aléatoire à l'ouverture

Avec l'utilisation de la méthode des forêts aléatoires, nous avons une redistribution dans l'importance des variables que nous pouvons voir figure 36.

Encore une fois nous voyons la prédominance de la cause du sinistre comme variable explicative de la charge ultime. Mais les variables suivantes ne sont plus exactement celles qui avaient une importance cruciale avec les arbres de décision.

En effet, à part Delta 1 en troisième place, les autres variables ont totalement été éclipsées par le montant de la cotisation payée et le montant de capacité de vol. Ces deux variables représentent une estimation du coût de l'appartement ainsi que de ce qu'il contient. Ils sont donc deux indicateurs de la valeur remboursable possible lors d'un incident.

La qualité de l'occupant et Delta 2 sont relégués respectivement à la 9^{è} et 11^{è} place dans l'ordre d'importance des variables.

Contrairement à la méthode précédente il n'est pas possible de donner un arbre représentant la forêt. En effet, la régression faite se base sur l'agrégation et la moyenne des différentes estimations faites par les arbres de la forêt. Seules les variables importantes peuvent être extraites.

Il serait légitime de penser que nous perdons en interprétation mais nous pourrions tout à fait représenter tous les arbres simulés par cette forêt et montrer le moyennage des résultats pour chaque sinistre subissant la régression mais ceci serait contre productif. Le processus étant plus long que celui des arbres CART, il est plus difficilement explicable visuellement de manière simple mais il n'en n'est pas pour autant plus complexe.

Afin de terminer avec les méthodes basées sur les arbres de décision nous décidons de mettre en

œuvre la méthode de Gradient Boosting Machine. Cette méthode permet d'avoir une souplesse plus grande qu'avec les forêts aléatoires mais les paramètres sont plus nombreux (nombre d'arbres, distribution supposée de la variable à expliquer, nombre minimum d'observations par nœud, nombre de permutation de la base d'apprentissage).

Ici nous n'avons influé que sur le nombre d'arbres générés et fait l'hypothèse que les coûts étaient distribués suivant une fonction normale (ce qui est une hypothèse discutable mais la mise en place d'une fonction réponse Gamma s'est avérée trop complexe à mettre en œuvre). De plus, nous avons mis le nombre de permutations à zéro afin que l'apprentissage soit fait sur exactement la même base que les autres méthodes pour ne pas fausser la comparaison des résultats.

Les résultats sont table 16.

Résultats sur la base de test		
Méthodes	MSE	Prédiction de la charge totale
GBM - 100 arbres	454 456	1,058
GBM - 1000 arbres	442 175	1,059

TABLE 11 – Résultats des méthodes de gradient boosting machine à l'ouverture

Ici nous voyons que ces modèles sont très similaires. L'apprentissage fait sur 1 000 arbres apporte un MSE légèrement inférieur à celui obtenu avec 100 arbres pour une erreur d'estimation de la charge globale identique de quelques millièmes.

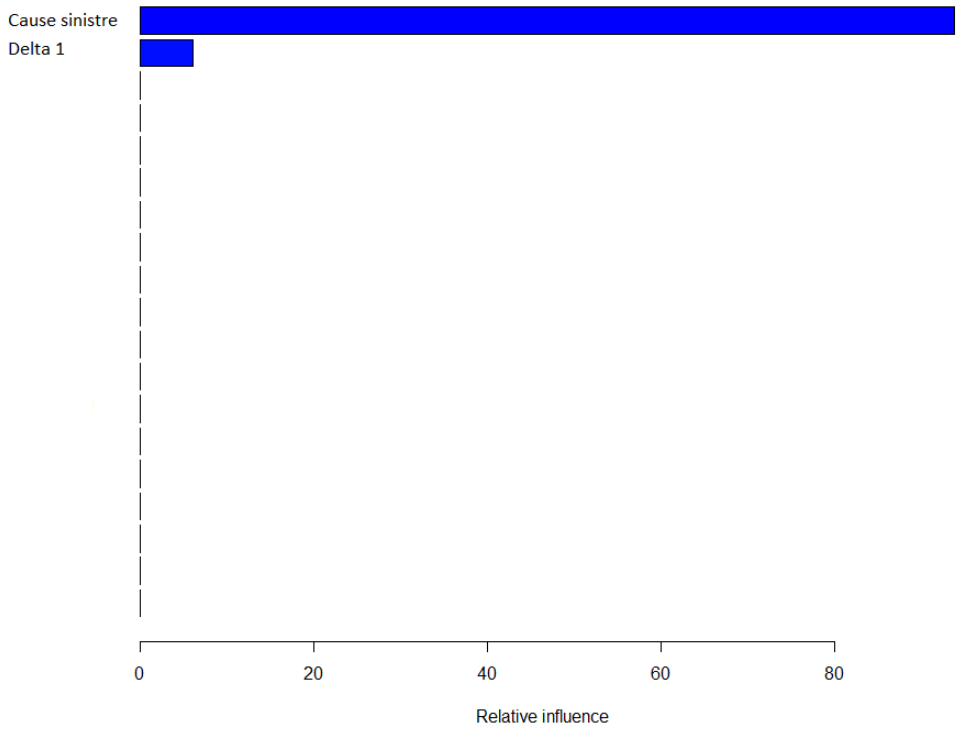

FIGURE 37 – Importance des variables avec la méthode GBM à l'ouverture

Dans ces modèles, de la même manière que pour les arbres de décisions, la cause du sinistre et l'intervalle de temps entre la survenance et la déclaration du sinistre sont les deux seules variables ayant une importance dans la prédiction du modèle (voir figure 37). Ceci nous confirme que ces deux variables sont d'une importance capitale.

Ces modèles n'arrivent pas à surpasser notre meilleur modèle de forêts aléatoires. Ceci peut s'expliquer car ces modèles ont tendance à sur-pondérer les erreurs qui ne sont pas bien prédites. Hors, les sinistres graves ici ont un coût extrêmement différents du coût moyen. Leur mauvaise prédiction et leur sur-pondération ont du avoir une influence sur le MSE final.

La dernière méthode que nous mettons en place est celui des réseaux de neurones. Dans celle-ci, nous devons choisir le nombre de couches ainsi que le nombre de neurones par couche. Il faut aussi déterminer le nombre d'itérations jusqu'à la convergence ainsi que le paramètre de decay. Ces modèles sont donc plus complexes à mettre en place mais peuvent apporter des résultats plus précis que les méthodes précédentes mises en place.

Ici nous décidons de fixer le nombre d'itérations à 100. Le paramètre de decay aura pris quatre valeurs différentes pour les modèles à une couche : 0, 0,1, 0,01 et 0,001.

Le modèle le plus précis a été obtenu par tâtonnements et les résultats sont visibles figure 38. Nous voyons tout de suite que tous les modèles (sauf un) ont un MSE inférieur à celui obtenu par les gestionnaires de sinistres.

Ces modèles améliorent donc en moyenne la prédiction du gestionnaire !

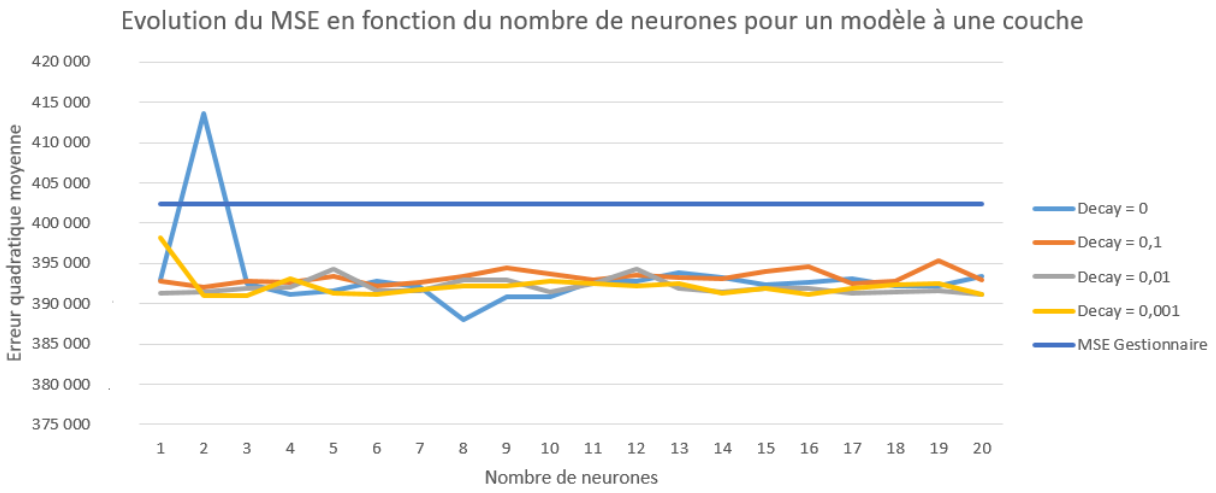

FIGURE 38 – Evolution du MSE pour un modèle à une couche pour différents decay

L'erreur reste quasiment constamment dans une bande large d'un pourcent (entre 390 000 et 395 000) quel que soit le decay choisi, mis à part pour deux modèles : celui à deux neurones et celui à huit neurones avec un decay nul.

Les modèles apportant tour à tour un MSE plus faible que les autres pour un nombre de neurones fixés nous décidons pour la suite de conserver un decay de 0 (car il nous apporte notre meilleur modèle) même si en moyenne le decay de 0,001 apporte les meilleurs résultats avec les MSE les plus faibles.

Nous voyons aussi que la prédiction de la charge globale est meilleure que celle qui est faite par les gestionnaires à l'ouverture. Celle-ci ne rivalise pas toujours avec la précision apportée par les méthodes précédentes mais elle améliore tout de même nettement celle sans modèles.

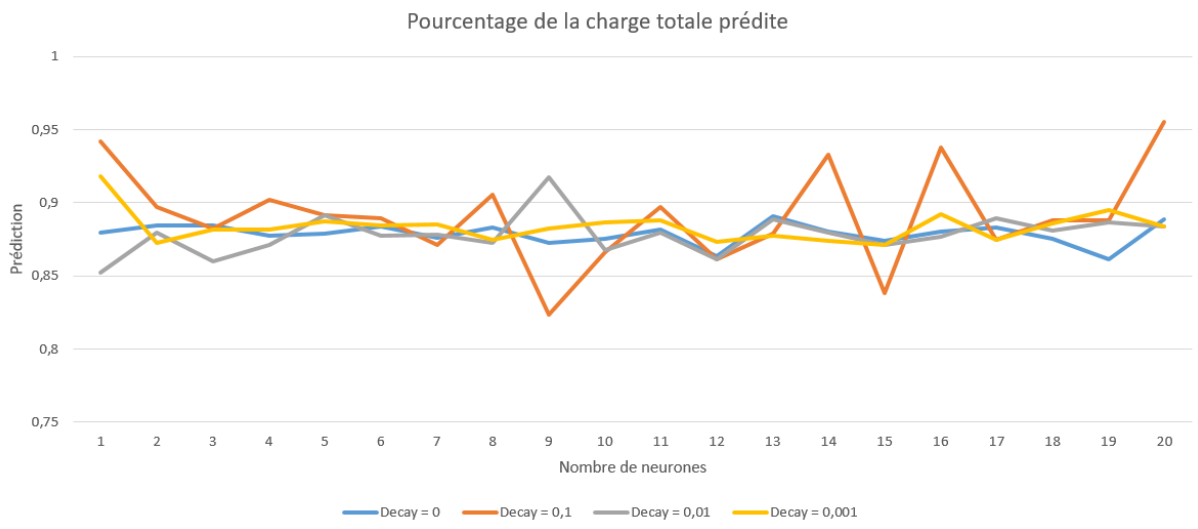

FIGURE 39 – Part de la charge totale prédite par des réseaux de neurones à une couche à l'ouverture

Nous voyons que la part de la prédiction de la charge totale est plus volatile que les MSE pour les modèles à une couche. La précision de la prédiction de la charge globale n'est pas linéairement corrélée au nombre de neurones dans notre couche unique.

Le modèle que nous avons conservé (8 neurones et un decay nul) prédit 88% de la charge globale. Ce qui est dans la moyenne des prédictions des modèles sans être exceptionnel. Nous confortons notre choix dans la conservation de ce modèle car le MSE de ce modèle est le plus faible (388 004).

Mais la configuration à une couche n'est pas la seule possible. En effet, nous avons aussi testé des configurations de réseau de neurones à deux couches, mais les résultats se sont détériorés. Les modèles testés étaient de la configuration suivante : $(1,i)$, $(i,1)$ ou (j,k) avec $i \in [\![1,20]\!], (j,k) \in [\![2,5]\!]^2$.

Rappelons qu'un modèle (a,b) signifie que la première couche cachée possède a neurones et la seconde couche cachée b neurones.

Les résultats de ces modèles ne sont pas présentés ici car ils n'apportent aucune information supplémentaire intéressante à l'étude.

La représentation du modèle avec seulement une couche cachée de 8 neurones (notés $H_i, i \in [\![1,8]\!]$) se trouve figure 40. La couche à gauche s'appelle la couche d'entrée où sont prises en compte les variables explicatives (les neurones notés I). La couche se trouvant à droite est la couche de sortie (le neurone noté O). Ici il n'y a qu'un seul neurone car nous faisons de la régression. Les neurones notés B sont des biais introduits afin d'améliorer la précision.

Ici, les variables les plus importantes recoupent celles qui avaient été mises en avant pas les précédents modèles. Nous pouvons voir celles qui ont le plus d'impact sur la figure 41. Les variables ayant

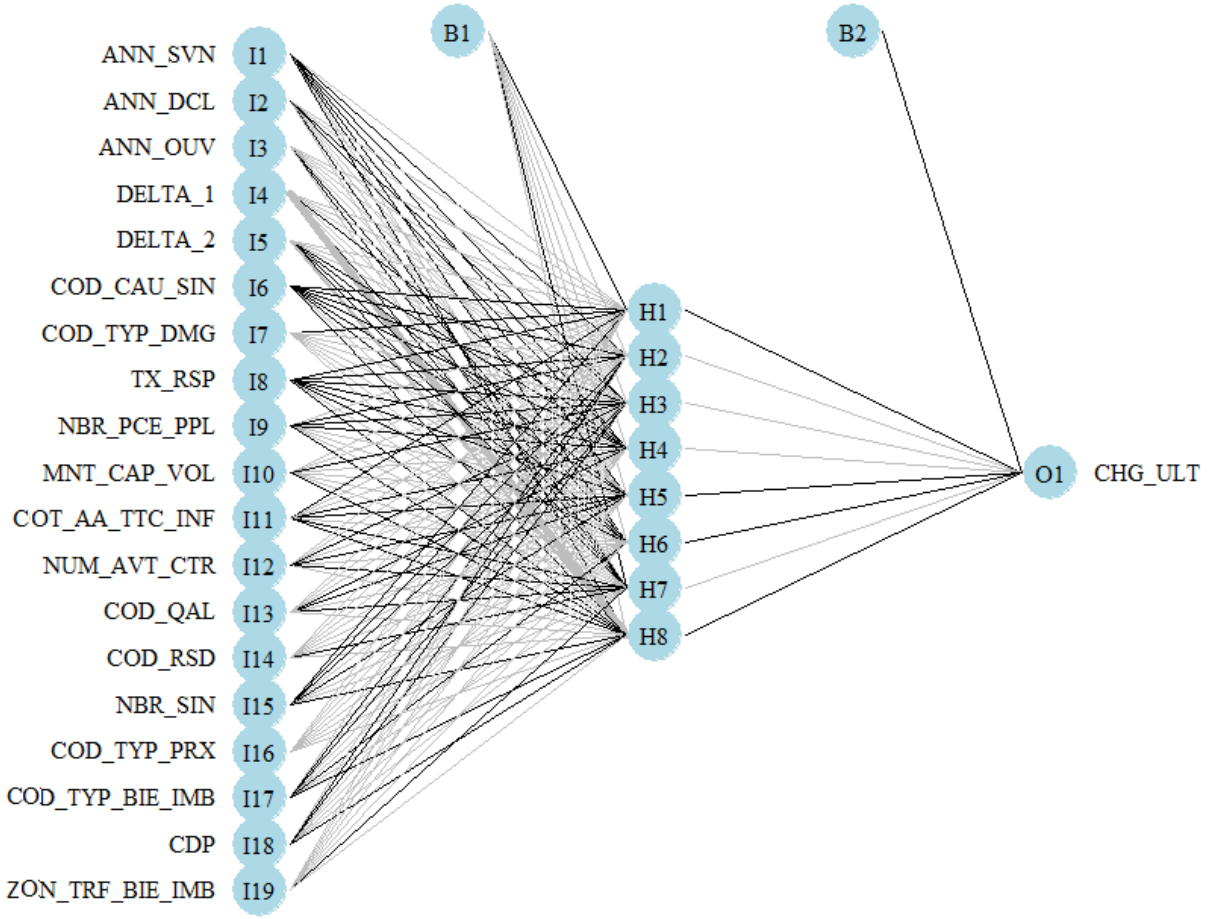

FIGURE 40 – Réseau de neurones apportant le meilleur MSE à l'ouverture

le plus d'importance sont Delta 1 et Delta 2 loin devant toutes les autres. S'en suit ensuite la cause du sinistre avec le nombre de pièces principales.

Le réseau de neurone apporte donc une importance primordiale aux deux temps de traitement du sinistre. C'est une information importante car ces deux temps ne sont pas des variables données directement par les bases sources. Ce sont des variables déduites et qui ne sont pas prises en compte par le gestionnaire de sinistre lorsqu'il intègre les éléments du sinistre.

Les réseaux de neurones sont donc le meilleur outil de prédiction que nous avons a notre disposition dans toutes les différentes méthodes mises en place.

Ils permettent une meilleure appréciation de tous les différents facteurs comme nous pouvons le voir avec l'apparition de la variable type de dommage dans les variables les plus importantes alors qu'elle n'est soit pas prise en compte du tout par les modèles ou alors très peu (20^{è} position pour les forêts aléatoires).

Il est important d'apprécier ce fait car le type de dommage impacte énormément la finalité du sinistre. Les coûts ne sont pas du tout les mêmes (comme le montre les différences d'évolution de l'inflation pour les dommages matériels et corporels).

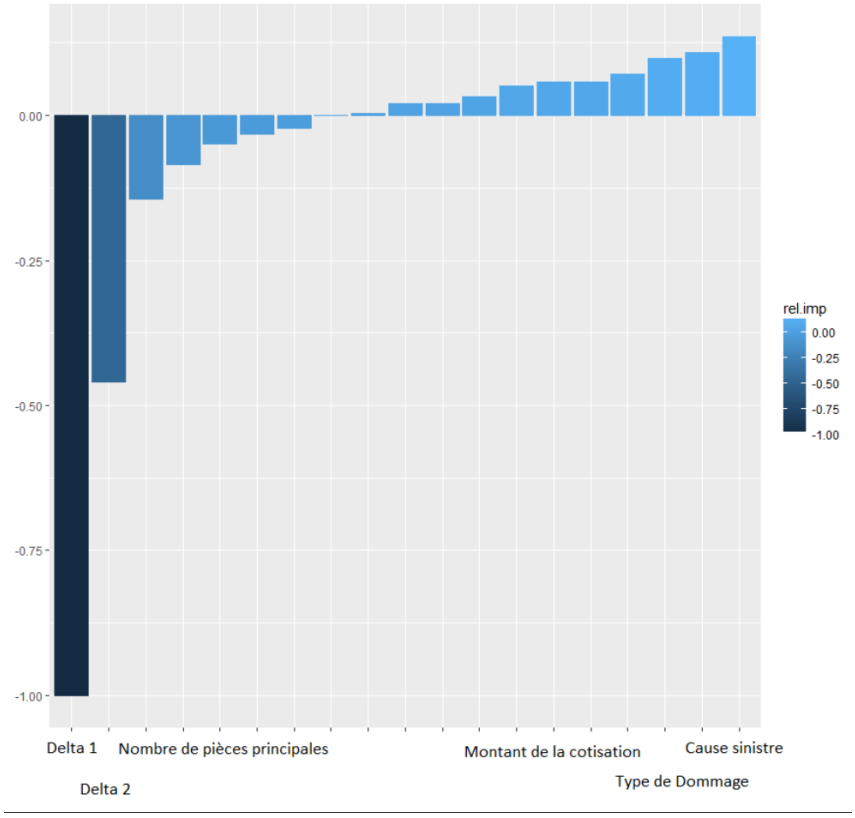

FIGURE 41 – Importance des variables dans le modèle de réseau de neurones retenu à l'ouverture

Nous retenons donc, parmi tous les modèles testés avec toutes les méthodes, le modèle de 8 neurones afin de voir quelle seront les prédictions faites sur les sinistres ouverts afin de les comparer avec les estimations faites par les gestionnaires. Cette étude se retrouve dans la partie suivante.

Une question mérite d'être soulevée concernant le lien entre ces deux paramètres que sont le MSE et l'estimation de la charge ultime globale. Comment est-il possible d'avoir deux modèles avec un MSE_1 inférieur à MSE_2 tout en ayant une estimation de la charge ultime globale moins précise pour le premier modèle que le second modèle ?

Ceci peut s'expliquer en terme de volume d'estimations correctes. En effet, le premier modèle va prédire un groupe de sinistre (appelons le α) de manière plus précise que le second modèle. Inversement pour un autre groupe de sinistres β. Considérons de plus que le premier modèle prédise mal sur β et que le second modèle prédise moins bien que le premier sur α mais mieux sur α que le premier modèle sur β. Il suffit que le volume de sinistres α soit plus important que celui de β pour que le MSE du premier modèle soit inférieur au second modèle. Si le premier modèle sous prédit sur β alors que le second modèle fait une estimation en moyenne correcte sur α, alors le second modèle prédira mieux la charge ultime totale tout en ayant un MSE plus élevé que le premier modèle.

Il est clair que l'utilisation de cette méthode est chronophage et bien plus complexe que les autres méthodes utilisées précédemment. En effet, la sélection de paramètres ne peut se faire qu'en testant un à un les modèles. Nous n'avons même pas exploré tous les modèles recommandés à deux neurones avec 20 entrées. Ajuster 20 neurones par couche dans deux couches représente 400 modèles différents. Nous avions choisi de prendre 4 decay différents et de laisser le nombre d'itérations à 100. Faire une étude

exhaustive de ces paramètres afin de trouver le meilleur modèle représente donc au moins 1 600 modèles avec des temps de traitement qui sont plus élevés. Et ceci ne représente que les réseaux de neurones à deux couches dont nous avons empiriquement fixé le nombre maximal de neurones par couche à 20.

A titre de comparaison, pour les arbres de régression il suffit de considérer le critère de segmentation. Une fois l'arbre maximal construit, l'élagation se fait naturellement en fonction de l'évolution de l'erreur de prédiction. Le nombre de branches est donc obtenu directement.

Pour les forêts aléatoires nous choisissons la taille de la forêt. Mais encore une fois l'évolution de l'erreur est rapidement visible en fonction du nombre d'arbres. Seul le nombre de variables sélectionnées aléatoirement se fait de manière empirique, ce qui peut amener jusqu'à un maximum de $n - 1$ modèles, avec n le nombre de variables explicatives (prendre $m = n$ revient un considérer un arbre maximal normal). Nous pouvons aussi jouer sur le nombre d'individus minimal possible dans chaque feuille terminale, sur le nombre maximal de feuilles terminales ainsi que sur le rythme d'apprentissage de la forêt (ceci revient à dire que l'algorithme fait une segmentation si cela améliore la prédiction de $x\%$). Ceci peut sembler certes représenter un nombre de modèles importants mais il y a une certaine linéarité des résultats dans la modification des paramètres (sauf pour m).

Augmenter le nombre d'arbres augmente la précision, diminuer le taux d'apprentissage nécessaire à chaque étape permet d'avoir des arbres plus volumineux et donc encore une fois d'augmenter la précision. Ceci est très différent des réseaux de neurones où chaque modification des paramètres n'est pas anticipable (ou très difficilement) par l'utilisateur.

Concernant le Gradient Boosting Machine, le nombre d'arbres entre encore en jeu mais désormais la forme de la réponse aussi. Ceci permet d'adapter la fonction de coût. Il faut aussi considérer le nombre minimal d'individus par nœud mais aussi le taux d'apprentissage.

Nous aurions pu décider d'utiliser plus de couches avec un moins grand nombre de neurones mais ceci complexifie le modèle et les résultats étaient assez éloignés des modèles avec une seule couche. Cette dégradation des résultats nous a poussé à arrêter la recherche d'un meilleur modèle.

II.4 Résultats sur les sinistres ouverts

Cette partie analysera les divergences entre les prédictions faites par le modèle sélectionné dans la partie précédente, un réseau de neurones à 8 neurones et celles faites par les gestionnaires sur les sinistres ouverts.

Les sinistres ouverts représentent 5% de la base et cette partie a pour but de voir quelles sont les différences notables entre nos prévisions pour chaque sinistre.

Il faut bien comprendre que les sinistres que nous allons étudier ne sont pas clos. Leurs coûts ne sont pas définitifs. Le dossier est toujours ouvert, le sinistre est en traitement. Le coût que nous avons est un coût arrêté à la date d'extraction de la base. Celui-ci est un coût plancher car il ne pourra diminuer (il représente la somme des versements faits) mais pourra augmenter. Cet argument est très important car même si l'estimation d'un des modèles est proche du coup actuel, peut être ne sera-t-il pas si proche du véritable coût final.

Reprenons nos indicateurs précédents, les résultats sont visibles dans la table 12. Dans ce tableau, le coût total représente la somme cumulée des règlements faits pour les sinistres ouverts.

Méthode	MSE	Charge Globale Prédite
Gestionnaire	1 543 307	62%
Réseau de neurones	3 942 511	74%

TABLE 12 – Etude à l'ouverture des dossiers ouverts au 31 décembre 2016

A première vue, l'estimation du gestionnaire semble bien plus précise individuellement (le MSE du réseau de neurones étant 2,5 fois plus élevé) même si encore une fois l'estimation de la charge globale est plus précise avec les réseaux de neurones. Ce qui doit être pris en compte ici est le fait que l'estimation des méthodes est arrêtée alors que le coût du sinistre peut évoluer. Les deux estimations sont donc bien en deçà de la réalité future, au terme de tous ces dossiers ouverts.

En effet, dans 53% des cas, le coût du sinistre en cours au 31 décembre 2016 est supérieur à celui de la provision d'ouverture du gestionnaire. Ce n'est vrai que dans 45% des cas avec l'estimation du réseau de neurones. La prédiction du réseau de neurones est supérieure à celle du gestionnaire dans 62% des cas.

Il est aussi intéressant de noter que pour 16% des sinistres, le gestionnaire a fixé un coût nul a des sinistres et pourtant certains ont un coût. Le réseau de neurones a estimé que 25% des sinistres auraient un coût nul. Mais de manière combinée, un coût nul a été attribué à seulement 4,6% des sinistres via les deux méthodes. Seul un sinistre grave a filtré entre les deux approches sur la cinquantaine de sinistres catégorisables comme graves à la date d'extraction.

Mais même si le réseau de neurones estime un plus grand nombre de sinistres nuls, la moyenne actuelle réelle des coûts de ces sinistres s'élève à 4 930 alors que la moyenne du coût des sinistres lorsque le gestionnaire estime un coût nul est presque deux fois plus élevée. Il semble donc que l'absence de coût prédite par le réseau de neurones soit plus fiable.

Il faut ajouter que lorsque le gestionnaire fait une estimation à l'ouverture nulle, le réseau de neurones, lorsque son estimation est inférieure à 21 000, fait une erreur moyenne moins élevée relativement au coût du sinistre. Ceci représente une estimation plus correcte pour 73% de ces sinistres. Soit 12% rapporté à l'ensemble de la base.

En ce qui concerne les sinistres graves, il est rare que l'estimation du gestionnaire soit complètement erronée. En effet, celui-ci possède, dans une grande partie des cas, des informations supplémentaires à celles implémentées dans les modèles afin d'anticiper la gravité du sinistre. Mais ce n'est pas forcément toujours le cas. Nous voyons que dès que la PSAP à l'ouverture du gestionnaire dépasse de plus de 20 fois la provision d'ouverture de gestion médiane alors le gestionnaire est automatiquement plus précis que le réseau de neurones.

Une tendance se dégage dans les données. On peut notamment remarquer que lorsque la prédiction du réseau de neurone est quatre fois plus élevée que celle du gestionnaire (ce qui représente 36% des sinistres ouverts), alors en moyenne, une fois les sinistres ayant une provision du gestionnaire nulle mis de côté, le coût des règlements est en effet en moyenne 4,2 fois plus élevé que ce qu'avait prévu le gestionnaire. De plus, il est aisé de démarquer les sinistres ayant reçu un versement forfaitaire (pour des raisons qui ne nous ont pas été communiquées). Ceci permet d'isoler des groupes de sinistres dont statistiquement les coûts seront bien plus élevés que ceux prédits.

De plus, lorsque l'on met de côté les provisions d'ouverture inférieures à 1 000 et que l'estimation du réseau de neurones est au moins quatre fois plus élevée que l'estimation faite par le gestionnaire de sinistre alors dans 88% des cas l'une de deux choses est réalisée :

— L'estimation du réseau de neurone est plus proche que celle du gestionnaire
— Il y a au moins 10 000 d'écart entre l'estimation du gestionnaire et le coût du sinistre (ce qui représente presque 4 fois le coût moyen d'un sinistre)

Dans les 12% des cas restants, le coût réglé est systématiquement deux fois plus élevé que la provision d'ouverture faite par le gestionnaire.

Nous voyons donc que ces méthodes peuvent avoir une certaine complémentarité. Là où le gestionnaire n'a pas d'information le réseau de neurones peut en donner une et vice-versa.

Il est difficile de tirer des généralités plus fortes que les règles qui viennent d'être énoncées précédemment. En effet, la méthode mise en place est très sensible aux données utilisées. C'est de la qu'elle tire ses intérêts mais aussi ses désavantages.

Grâce à cette méthode nous avons pu tirer des enseignements sur les sinistres et mettre en lumière des points qui aideront la fonction actuarielle à aiguiller les services de gestion pour repérer un certain type de sinistres grâce au processus mis en place au cours de cette partie d'application.

II.5 Conclusion

Nous avons donc vu que les algorithmes de machine learning peuvent permettre de détecter d'éventuels dysfonctionnements et ainsi d'adapter et de revoir certaines règles de gestion de sinistre.

Certains algorithmes prédisent mieux la charge ultime globale mais sont moins satisfaisants au niveau ligne à ligne. C'est le cas des méthodes basées sur les arbres.

Ces résultats obtenus permettent de constater qu'il est possible d'ajuster et d'adapter les règles de provisionnement dossier/dossier pour intégrer les résultats de l'algorithme. Nous avons pu voir en effet qu'ils sont partiellement généralisables sur un certain segment de la sinistralité.

De cette première partie de l'étude nous pouvons tirer plusieurs enseignements d'une relative importance.

Pour toutes les méthodes qui ont été mises en place, deux paramètres ont été mis en avant : la prédiction de la charge ultime globale ainsi que le MSE.

Il est donc important de savoir, au premier abord, quel paramètre est le plus intéressant afin de se focaliser sur le meilleur modèle. Pour les besoins de cette étude, le paramètre qui a été privilégié est le MSE afin de savoir si les algorithmes pouvaient apporter une information qui permettrait par exemple d'épauler, d'aiguiller ou potentiellement de guider, le gestionnaire de sinistre dans son évaluation à l'ouverture.

Nous nous sommes donc intéressés au modèle qui nous apportait le meilleur MSE et donc à un réseau de neurones en particulier. Mais pour un besoin d'estimation de la charge ultime globale nous avons pu remarquer que les méthodes basées sur les arbres de régressions apportaient une meilleure précision. Ces

modèles étant moins chronophages et plus aisés à paramétrer. Si avoir une indication de la charge ultime globale est le seul point d'intérêt alors nul besoin d'utiliser les réseaux de neurones.

Si maintenant c'est le MSE qui provoque un intérêt alors les modèles des réseaux de neurones semblent mieux adaptés. Seuls ceux-ci on pu obtenir une erreur quadratique moyenne plus faible que les estimations des gestionnaires. Pour les modèles à une couche cela a même été la norme pour tous les modèles (sauf un).

Les résultats obtenus sur les sinistres ouverts en sont la preuve. Les prédictions obtenues par les réseaux de neurones peuvent permettre dans certains cas d'obtenir une prédiction plus ajustée que celle du gestionnaire à l'ouverture du sinistre.

Le modèle utilisé permet donc dans certains cas de figure (précisés dans la partie précédente) spécifiques d'être plus proches que l'estimation donnée par le gestionnaire. Il ne permet pas toutefois d'être généralisable à tous les cas possibles.

En ce qui concerne l'estimation de la charge ultime, le réseau de neurones n'est pas en reste car son estimation de cette charge globale est plus précise que celle faite par le gestionnaire à l'ouverture. Le réseau de neurones, en étant plus prudent est plus proche de la réalité au global.

Nous voyons donc à travers la première partie de cette étude que l'utilisation des méthodes de machine learning peuvent permettre d'améliorer les règles d'évaluation des dossiers sinistres.

III Etude à date d'extraction

La première partie de notre étude s'intéressait aux sinistres historiques à la date d'ouverture et reposait alors sur les informations connues à cette date.

Cette partie sera en revanche consacrée à l'étude des dossiers ouverts à la date d'extraction de la base. Ces dossiers auront donc reçu des éléments complémentaires qui complètent les informations disponibles à l'ouverture.

Afin d'appliquer cette méthode, nous devons toutefois modifier le champ d'application. En effet, notre but est de challenger la provision du gestionnaire à la date d'extraction de la base. Hors les sinistres clos ne possèdent plus de provision. La charge ultime est égale à la somme des règlements et des honoraires. Alors que pour les sinistres ouverts, la charge ultime est égale à la somme des versements et de la provision du gestionnaire.

Ici nous prédirons la charge ultime sur les dossiers clos et ferons de même sur les dossiers ouverts. Une fois cette charge ultime prédite des dossiers ouverts, nous soustrairons les honoraires et les règlements versés afin de comparer la valeur obtenue à la somme de la provision du gestionnaire à la date d'extraction avec la provision d'honoraires à la date d'extraction.

L'étude menée a donc le même but que dans la première partie.

Données présentes dans la base au moment de l'extraction

Données issues de la base à l'ouverture		
Variable	Type	Modalités
Date de survenance	Entier	
Date de déclaration	Entier	
Date d'ouverture	Entier	
Cause du sinistre	Catégorielle	23
Delta 1	Continue	
Delta 2	Continue	
Nombre de pièces principales	Entier	
Type de dommage	Catégorielle	2
Taux de responsabilité	Continue	
Capacité de vol	Continue	
Cotisation annuelle	Continue	
Nombre d'avenants	Entier	
Qualité de l'occupant	Catégorielle	17
Code Résidence	Catégorielle	4
Nombre de sinistres	Entier	
Type de prix	Catégorielle	2
Type de bien	Catégorielle	5
Code Postal	Entier	
Zone Tarifaire	Entier	

Données supplémentaires issues de la base à la date d'extraction		
Variable	Type	Modalités
Delta	Entier	
Top Réouverture	Booléen	2
Top Annulation	Booléen	2

TABLE 13 – Récapitulatif des variables dans la base de données à la date de l'extraction

III.1 Données à la date d'extraction

Les informations sur le client, le contrat et le bien sont restées inchangées mais la base sinistre a elle été agrémentée d'informations nouvelles.

Nous conservons les variables sélectionnées à l'ouverture du dossier mais à celles-ci s'ajoutent un indicateur signifiant si le dossier a été réouvert et un autre indicateur permettant de savoir si le dossier est annulé (cela contient les dossiers sans suite, les dossiers n'étant finalement pas couverts par la garantie ou des dossiers purement annulés). A ces deux variables, nous rajoutons une variable d'importance : la durée d'ouverture du dossier qui représente l'intervalle de temps entre l'ouverture du dossier et sa clôture (que nous nommerons Delta). Pour les sinistres ouverts, cet intervalle de temps représente le nombre de jours entre l'ouverture du dossier et la date d'extraction de la base. C'est cette donnée qui est censurée à droite. Il a donc fallu, grâce à l'estimateur de Kaplan-Meier, déterminer la durée de vie résiduelle des dossiers de sinistres ouverts (c'est-à-dire la durée du vie restante du dossier sachant que le dossier est déjà ouvert depuis x jours).

Nous ne prenons pas en compte les estimations de recours ou les provisions de recours faites car l'étude est menée en brut de recours. La démarche sur le brut pourra bien entendue être étendue aux recours à l'avenir.

Un récapitulatif des variables utilisées à la date d'extraction se trouve dans la table 13.

III.2 Postulats

Dans cette partie, il nous est impossible de challenger directement la provision faite par le gestionnaire lors de l'évaluation à date d'extraction.

En effet, la relation suivante est vérifiée à tout instant :

$$\text{Provision} + \text{Règlements} = \text{Charge Ultime}$$

Dans notre première approche, à l'ouverture du sinistre, il n'y a aucun règlements de fait. La provision, à l'ouverture, est directement l'estimation de la charge ultime : $\text{Provision}_{\text{ouverture}} = \text{Charge Ultime}_{\text{ouverture}}$. Mais à date d'extraction, la provision a été liquidée au cours du temps.

Malheureusement, pour les sinistres clos, la provision est nulle et nous avons : $\text{Règlements}_{\text{clos}} = \text{Charge Ultime}_{\text{clos}}$.

Nous ne pouvons donc pas simplement apprendre sur la charge ultime contrairement à ce qui a été fait dans la première partie.

La démarche a été la suivante :

- Estimation de la charge ultime sur les sinistres clos
- Détermination du meilleur modèle
- Prédiction de la charge ultime sur les ouverts
- Déduction des règlements et des honoraires à la date d'extraction pour obtenir la provision dossier/dossier estimée à date d'extraction
- Comparaison des provisions estimées avec les provisions en cours

La validité et l'interprétation des résultats seront plus complexe dans cette partie car nous allons comparer deux estimations. Celles obtenues par les gestionnaires et celles obtenues algorithmiquement.

Nous ne sommes plus en mesure d'avoir un référentiel fixe établi (comme l'était la charge ultime à l'ouverture) car nous ne possédons que la charge ultime à la date d'extraction, ce n'est qu'un minimum.

III.3 Modèles

Concernant les modèles mis en place, ils sont identiques à ceux qui ont été utilisés dans la première partie.

Nous avons commencé l'étude de la même manière que précédemment, c'est-à-dire avec les modèles CART. Les résultats se trouvent dans la table 14.

Résultats sur la base de test		
Méthodes	MSE	Prédiction de la charge totale
CART - Elagué	650 922	0,986
CART - Bagging	647 337	0,988
CART - Entropie	650 922	0,986

TABLE 14 – Résultats des méthodes CART à la date d'extraction

Le modèle qui apporte le plus satisfaction est celui faisant une agrégation de modèle (Bootstrap Aggragating), tant d'un point de vue MSE que prédiction de la charge globale. Les modèles élagués, que ce soit avec le critère de Gini ou le critère d'information, sont identiques. La prédiction globale sous estime très légèrement la charge totale de 1,5%.

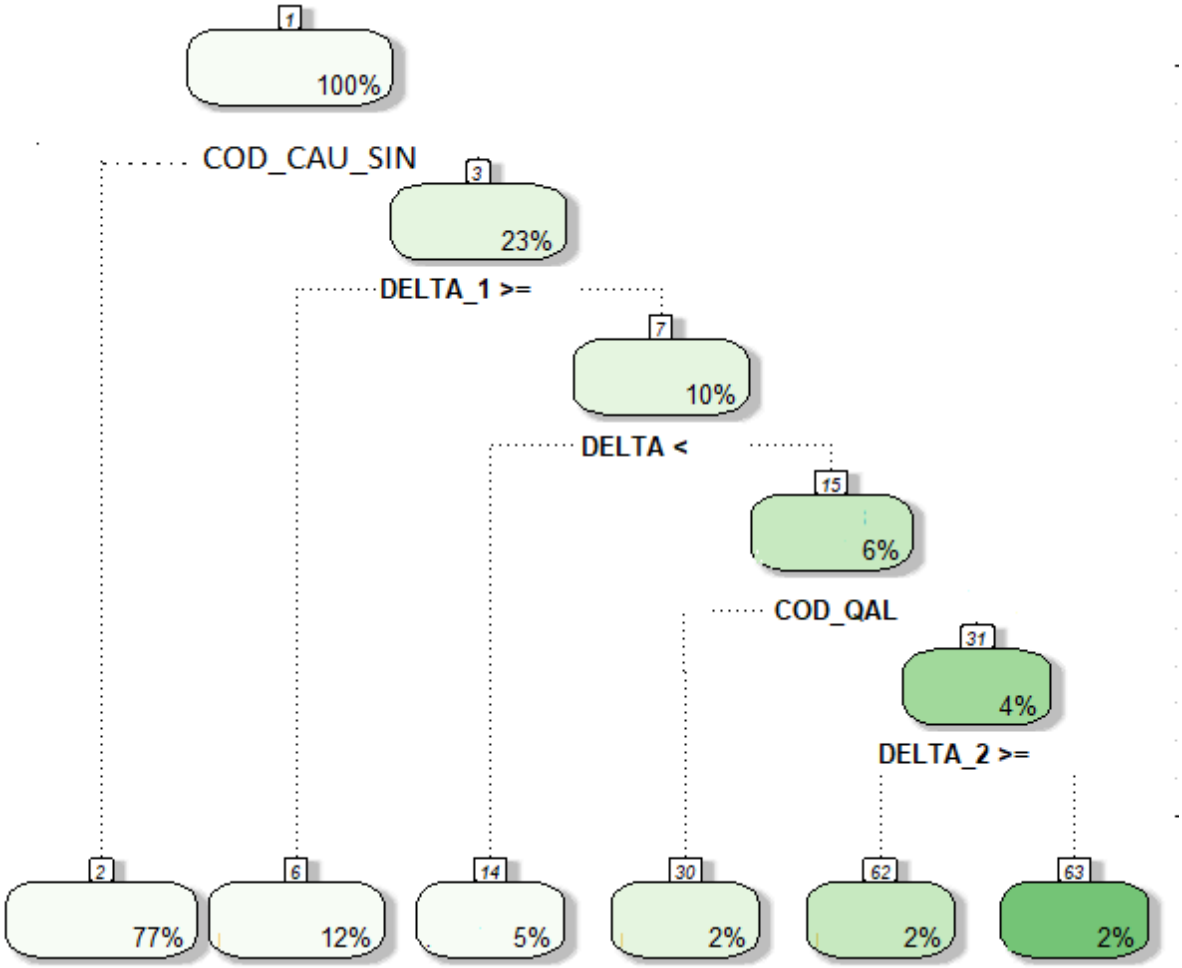

FIGURE 42 – Visualisation des variables importantes avec le modèle CART Elagué à la date d'extraction

La figure 42 nous montre l'arbre élagué ainsi que l'importance des variables associées au modèle CART. Il n'est pas possible de représenter l'arbre pour le modèle Bootstrap car il s'agit d'une moyenne d'estimateurs, mais les modèles étant extrêmement proches cela nous oriente tout de même.

De la même manière que dans la partie à l'ouverture, les variables que sont la cause du sinistre (COD_CAU_SIN), la qualité de l'occupant (COD_QAL) ainsi que les durées Delta_1 et Delta_2 font leur apparition.

Nous voyons de plus que la variable Delta (que nous avons du modéliser pour cause de censure à droite) est elle aussi présente. D'où l'intérêt de la prise en compte de la censure.

Nous voyons donc que tous les temps de traitement de gestion et d'enregistrement du dossier sinistre par l'organisme ont un impact direct sur le coût de celui-ci. Ceci pourrait paraître surprenant car ces durées pourraient sembler indépendantes du sinistres mais ce n'est pas forcément le cas, plus un sinistre possède d'éléments ou est complexe, plus il est difficile à constater et donc plus il est long de le déclarer.

La durée d'ouverture Delta représente aussi les complications qui peuvent intervenir lors de la gestion du dossier. Plus le dossier est complexe (sinistre corporel grave par exemple) plus celui-ci aura un coût élevé.

La prochaine méthode mise en place est celle des forêts aléatoires. Les résultats sont visibles dans la table 15.

Résultats sur la base de test		
Méthodes	MSE	Prédiction de la charge totale
RF - m=2	621 614	1,011
RF - m=3	606 285	1,029
RF - m=4	602 810	1,040
RF - m=5	603 117	1,038
RF - m=6	602 436	1,047
RF - m=7	608 973	1,052
RF - m=8	608 576	1,051
RF - m=9	613 014	1,057

TABLE 15 – Résultats des méthodes de forêts aléatoires à la date d'extraction

Dans la méthode précédente, notre meilleur modèle prédisait en même temps le meilleur MSE et la meilleure prédiction globale. Ici ce n'est pas le cas, notre meilleur MSE est modélisé lorsque $m = 6$ alors que la meilleure prédiction au global est pour $m = 2$ avec une amélioration de 3,5%.

Malgré cela, nous voyons que le modèle $m = 2$ est meilleur pour les deux prédictions que le modèle CART Bootstrappé à la seule différence que la forêt aléatoire surestime la charge globale alors que l'autre méthode la sous-estime.

L'importance des variables se trouve dans la figure 43. Ici nous voyons la prépondérance de l'importance de la variable Delta qui est loin devant la deuxième variable significative. Encore une fois, la cause du sinistre ainsi que Delta_1 font partie des variables les plus importantes. A noter que le montant de la cotisation annuelle et le montant de la capacité de vol ferment ce groupe de variables les plus importantes avec le code postal.

La variable Delta_2 est quant à elle reléguée plus bas après le nombre de pièces et la qualité de l'occupant.

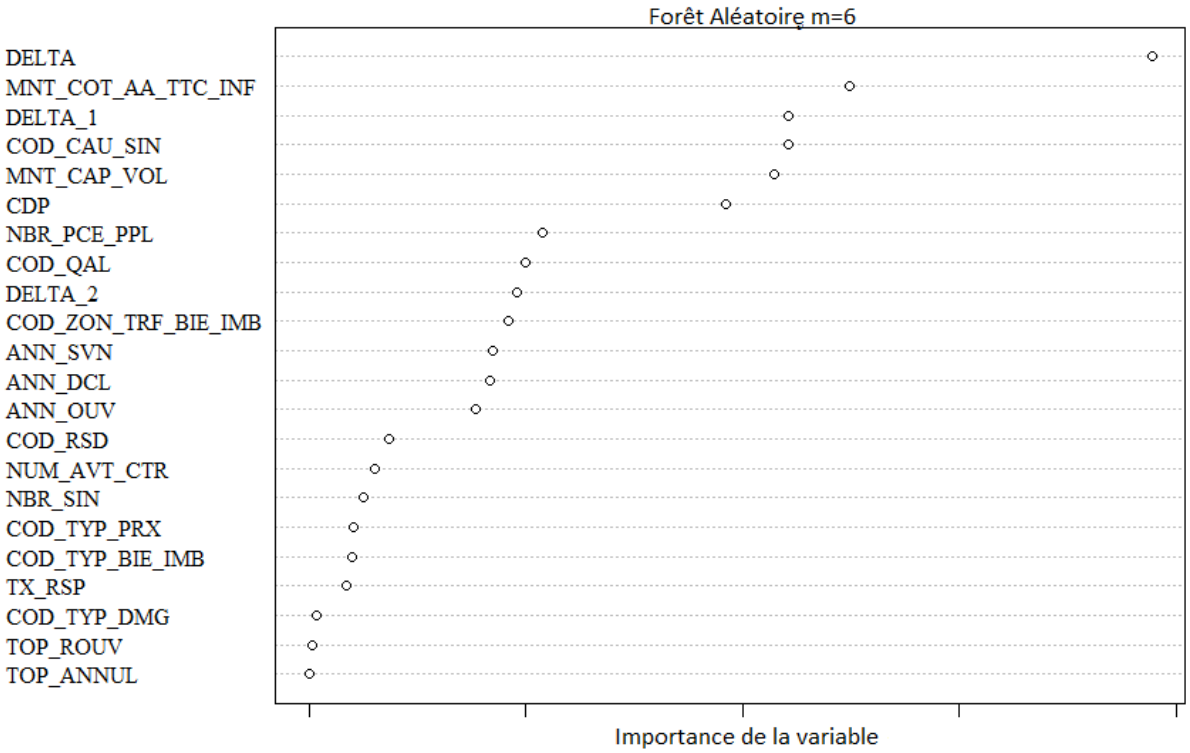

FIGURE 43 – Importance des variables avec la méthode de Random Forest à la date d'extraction

Nous voyons que les deux autres variables ajoutées lors de l'extraction de la base que sont l'annulation du sinistre (plus précisément l'annulation de sa prise en charge par l'organisme assureur) ainsi que la réouverture ou non du sinistre ne sont pas du tout impactantes sur le modèle et sont reléguées au rang des deux variables les moins importantes.

Notre prochaine méthode est celle du GBM, les résultats se trouvent 16.

Résultats sur la base de test		
Méthodes	MSE	Prédiction de la charge totale
GBM - 100 arbres	748 929	1,017
GBM - 1000 arbres	721 630	1,025

TABLE 16 – Résultats des méthodes de gradient boosting machine à la date d'extraction

Les résultats ne sont pas ici meilleurs que pour les modèles précédents. Le MSE est bien supérieur à celui des modèles vus précédemment. L'estimation de la charge globale est proche mais l'erreur est tout de même plus élevée.

Encore une fois, cette modélisation mériterait plus d'attention notamment vis-à-vis de la forme de la réponse attention. L'hypothèse ici que la réponse est gaussienne est très forte et biaise les résultats du modèle fortement.

En ce qui concerne l'importance des variables nous pouvons voir dans la figure 44 que les variables

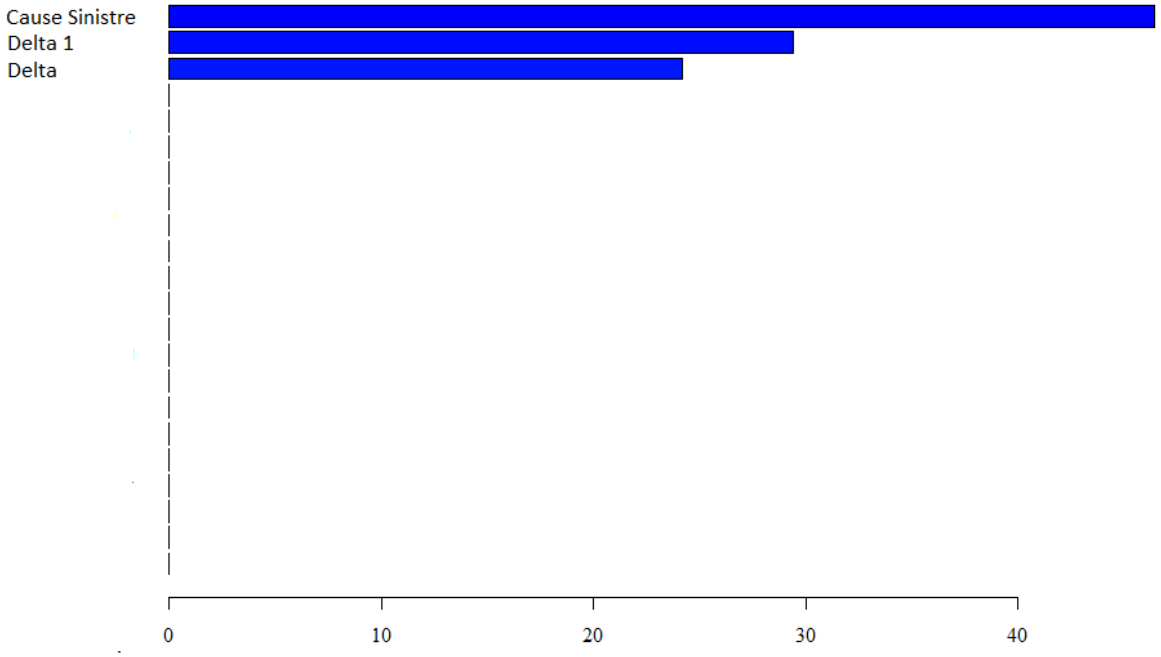

FIGURE 44 – Importance des variables avec la méthode GBM à la date d'extraction

les plus importantes restent relativement inchangées. Ici la cause du sinistre, Delta_1 et Delta sont les seules variables nécessaires afin d'obtenir une estimation du modèle. Les autres variables ne sont même pas considérées.

La prochaine méthode mise en place est celle des réseaux de neurones. Exactement les mêmes modèles que dans la partie à l'ouverture ont été utilisés dans le processus d'estimation.

Les résultats de cette méthode n'améliorant pas ceux des méthodes précédentes, seules les informations sur le meilleur réseau de neurones sont visibles dans la table 17.

Résultats sur la base de test			
Méthode	Construction	MSE	Prédiction de la charge totale
Réseau de neurones - 100 arbres	(1,2)	636 564	0,993

TABLE 17 – Résultats des méthodes de gradient boosting machine à la date d'extraction

Ce modèle à deux couches, avec un neurone dans la première et deux dans la seconde (visibles dans la figure 45) n'améliore pas le MSE des modèles précédents mais apporte une estimation de la charge globale avec une erreur en dessous du pourcent.

Ce qui nous intéresse est de challenger les provisions ligne à ligne, nous nous concentrons donc sur le modèle qui apporte le meilleur MSE. Mais pour une estimation de la charge globale seulement, le modèle de réseau de neurones est à privilégier.

En effet, malgré des MSE plus élevés que les forêts aléatoires, voir plus élevés que pour les modèles GBM, l'estimation de la charge globale est souvent bien plus précise dans l'ensemble que celle donnée

par les autres modèles.

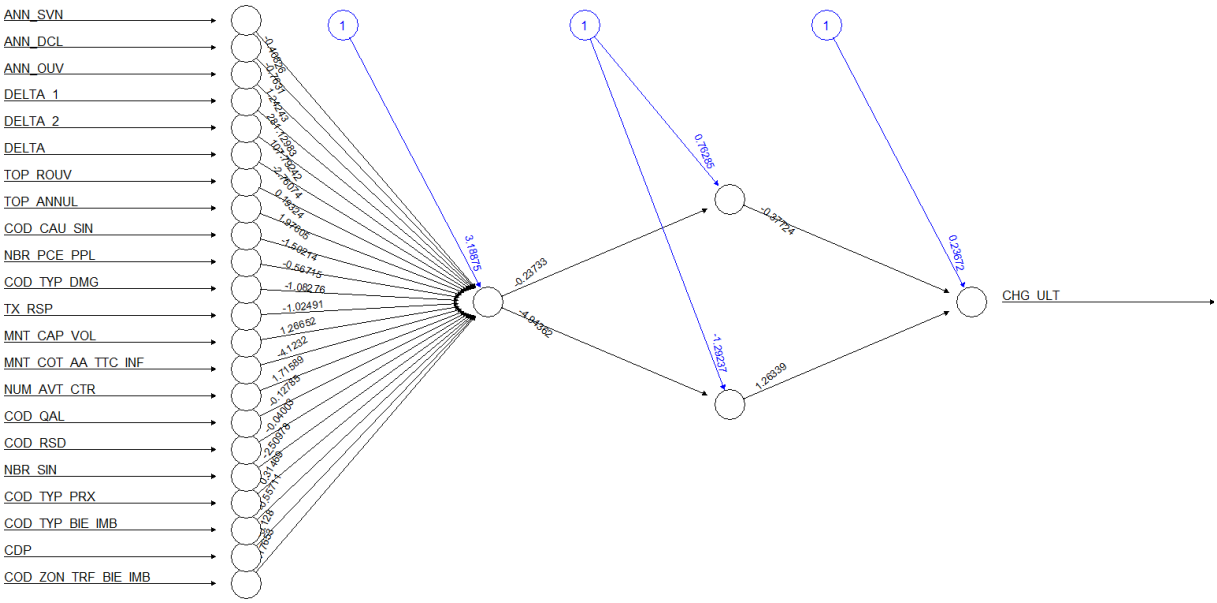

FIGURE 45 – Meilleur réseau de neurones prédicteur à la date d'extraction

En ce qui concerne l'importance des variables, nous pouvons la voir dans la figure 46.

Les variables les plus importantes sont presques identiques à celles obtenues par les modèles précédents. La variable Delta_1 est incontournable, les variables Delta, la cause du sinistre et le type de bien sont encore une fois présentes.

Comme pour la partie précédente, l'étude à la date d'ouverture des dossiers de sinistres, le réseau de neurones est le seul à capter l'importance du type de dommage. En effet, le type de dommage est la seconde variable la plus importante dans ce modèle.

Ceci confirme l'intérêt de modélisations inflationnistes différentes pour les sinistres corporels et matériels.

Nous voyons donc, à la suite de la mise en place de ces modèles, que celui qui nous apporte le MSE le plus faible est le modèle de forêt aléatoire avec un $m = 6$, nous le considérerons dans la partie suivante qui porte sur l'étude des sinistres ouverts.

Encore une fois, nous aurions pu choisir un modèle qui possède une meilleure estimation de la charge globale mais la prédiction du modèle avec $m = 6$ n'est pas si éloignée du meilleur modèle de réseau de neurones et possède un MSE bien meilleur.

A l'inverse de la partie précédente, les résultats obtenus via les arbres CART sont très proches de ceux obtenus par forêt aléatoires avec deux bonnes prédictions pour les deux indices considérés, malgré la robustesse accrue du modèle de forêt aléatoire. Les modèles GBM n'ont pas apporté satisfaction. L'utilisation des réseaux de neurones est ici importante car la modélisation est bien plus chronophage pour des résultats qui localement ne sont pas précis (moins que d'autres modèles) mais globalement font une meilleure estimation.

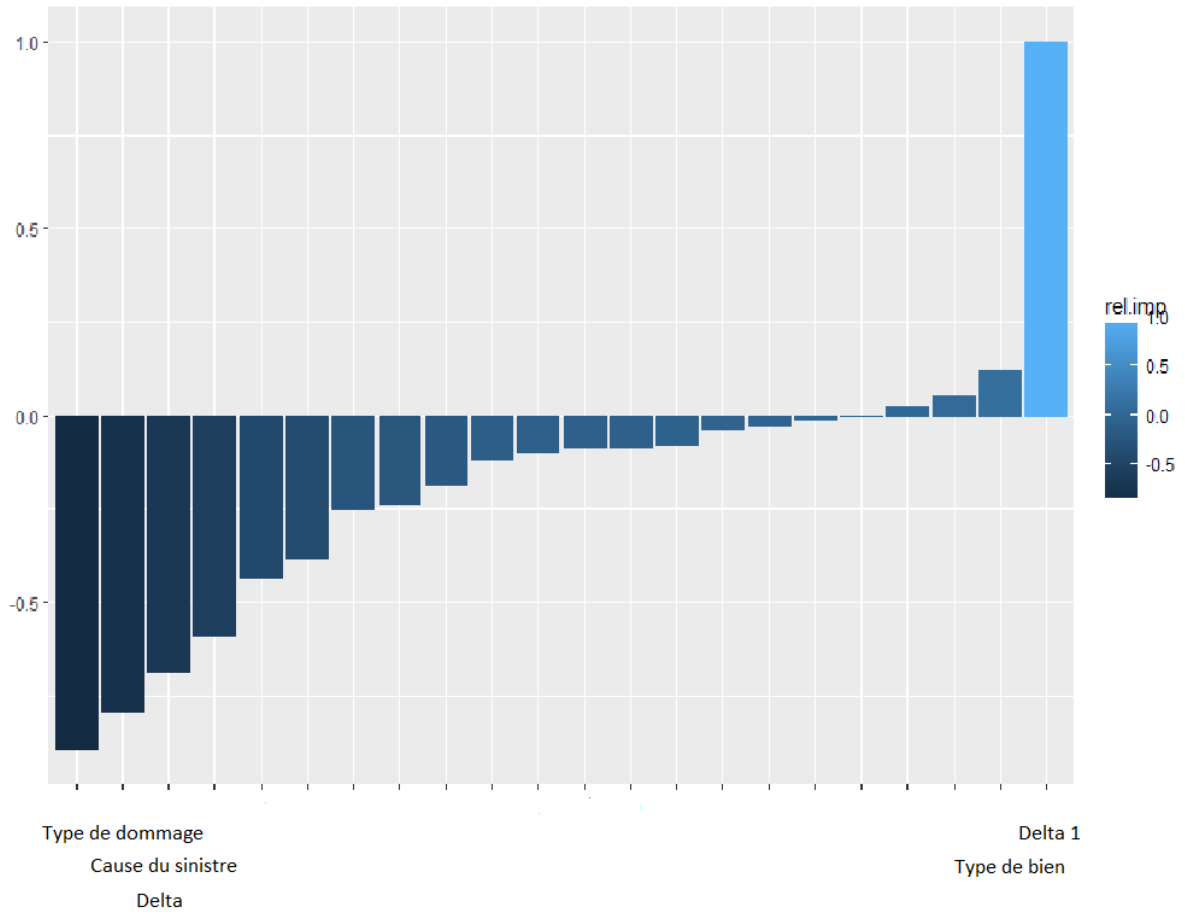

FIGURE 46 – Importance des variables avec la méthodes des réseaux de neurones à la date d'extraction

III.4 Résultats sur les sinistres ouverts

Lors de la phase d'apprentissage et de test sur les clos, les variables ont été utilisées telles qu'elles sont dans les bases. La seule différence avec la projection pour les ouverts est contenue dans la variable Delta qui a été modifiée. En effet, elle est connue pour les sinistres clos mais censurée pour les sinistres ouverts. Nous avons donc rajouté à la durée extraite sa durée de vie résiduelle pour avoir une estimation correcte de la variable Delta finale (la durée effective d'ouverture des dossiers).

Pour les sinistres clos, Delta a été inchangée mais pour les sinistres ouverts, la variable a été modifiée de la manière suivante afin de prendre en compte la censure :

$$\text{Delta}^*_i = \text{Delta}_i + \mathbb{E}(\text{Delta}_i | \text{Delta}_i \geq t_i)$$

Avec Delta* le Delta ajusté pour les sinistres ouverts, Delta la durée d'ouverture du dossier jusqu'à date d'extraction et t l'instant de censure et i l'indice indentifiant les sinistres ouverts.

A la vue de l'importance de la variable Delta dans les modèles précédents, il est était nécessaire de prendre en compte cette modification.

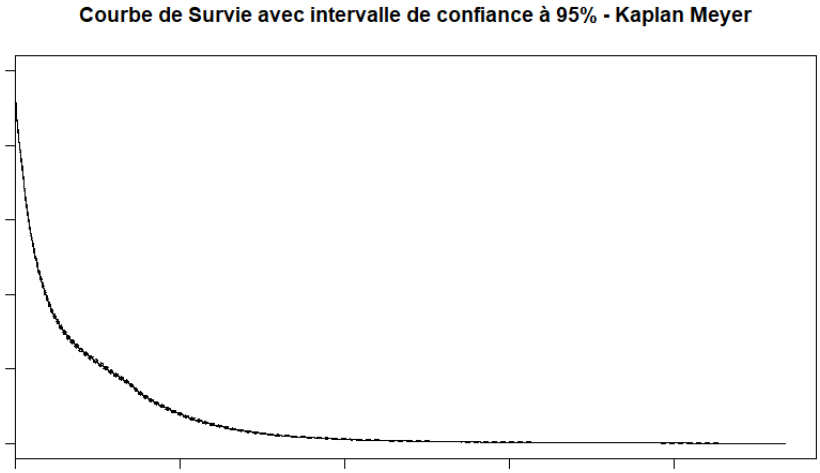

FIGURE 47 – Fonction de survie estimée par Kaplan-Meier avec intervalle de confiance à 95%

Nous voyons dans la figure 47 qu'à peine 80% des dossiers sont traités en moins de 500 jours.

Les statistiques sur cette variable sont table 18.

Statistiques de Delta	
Minimum	0
1er Quart	60
Médiane	160
Moyenne	360
3^{è} Quart	460
Maximum	4600

TABLE 18 – Statistiques sur la variable Delta

Nous voyons que le premier quart des dossiers est réglé très rapidement, entre sept et huit semaines. Alors que la moitié des dossiers sont traités en quatre mois.

Cette première partie de traitement concerne des sinistres attritionnels. En ce qui concerne le troisième quart des sinistres ceux-ci sont réglés dans l'année. Nous pouvons voir que sur le dernier quart, la durée d'ouverture des sinistres peut s'éterniser. Ce peut être le cas lorsqu'un litige n'est pas réglé et que les torts ne sont pas reconnus.

Nous pouvons voir sur la figure que les sinistres corporels restent ouverts plus longtemps de manière générale que les sinistres matériels. Mais nous voyons aussi qu'ils ne s'éternisent pas à la manière de certain sinistres matériels.

Sur cette figure, au-delà du dernier sinistre corporel, il reste 0,7% de sinistres (tous matériels). Ces sinistres ne représentent pas un coût particulièrement élevé. En effet, prêt d'un tiers ont un coût nuls, ceux du deuxième tiers ont un coût moyen et le dernier tiers ont un coût que l'on peut considérer comme élevé.

Sur la figure 49, le paramètre mis en avant est la responsabilité du client. Nous pouvons voir que

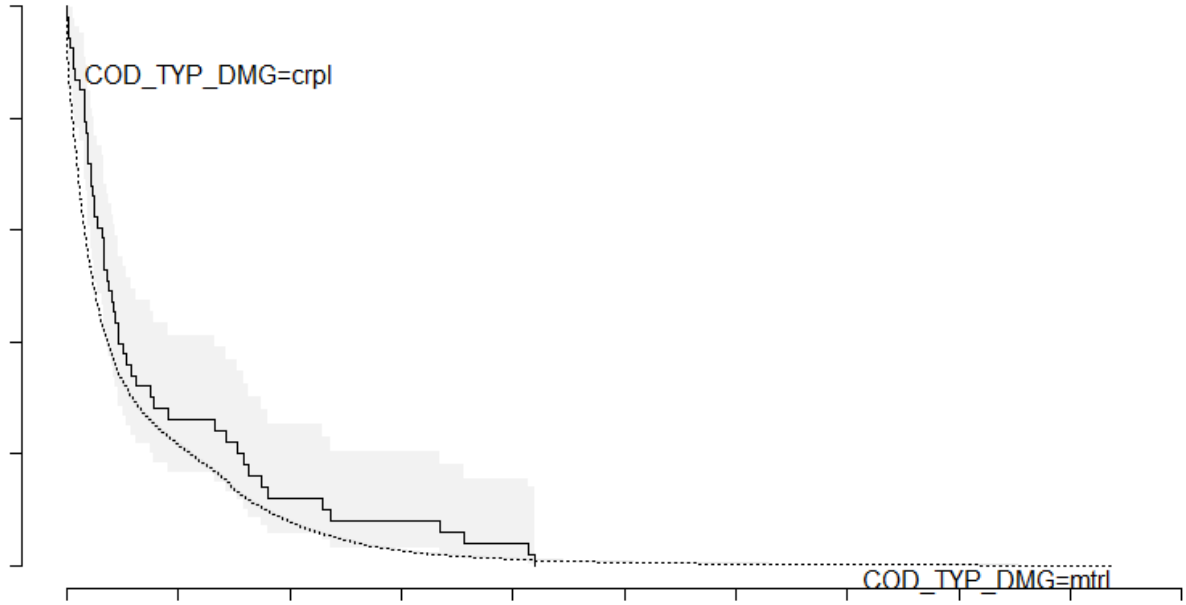

FIGURE 48 – Fonction de survie en fonction du type de dommage

lorsque le client n'est pas responsable, les dossiers restent ouverts en moyenne plus longtemps. Ceci doit être du à la recherche de la responsabilité et au temps de traitement et d'échange entre les organismes d'assurance.

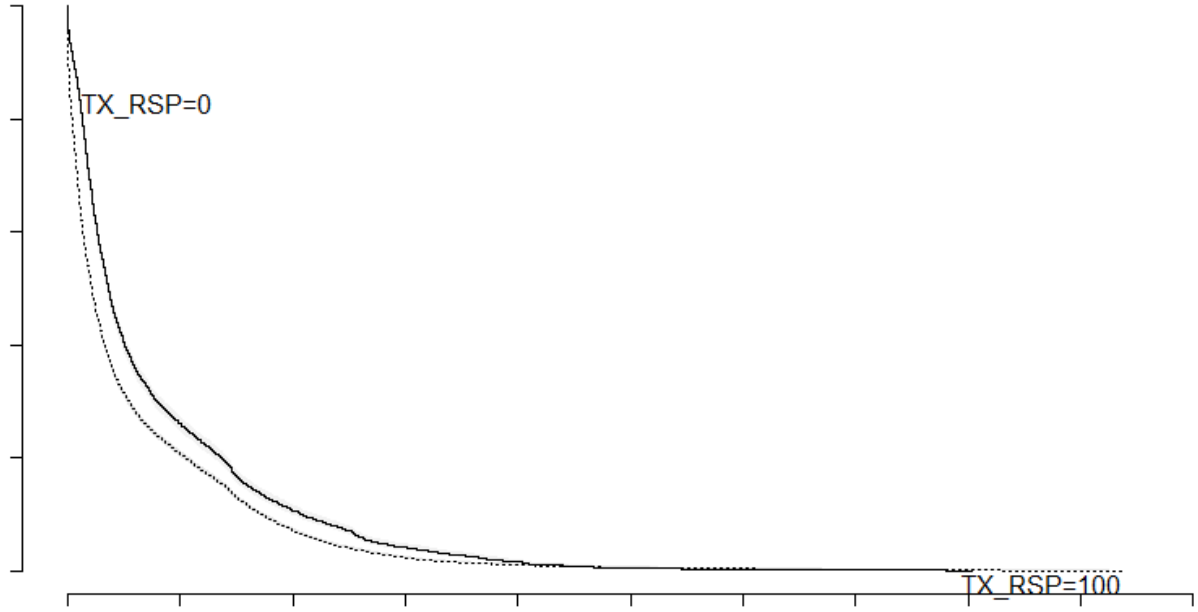

FIGURE 49 – Fonction de survie en fonction du taux de responsabilité

Nous pourrions faire une analyse de la durée d'ouverture des dossiers sur les autres variables importantes comme la qualité de l'occupant ou la cause du sinistre mais elle ne sera pas mise en avant ici car

une figure lisible avec une quinzaine de paramètres et de courbes sera peu lisible.

Nous pourrions par conséquent déterminer quels seraient les cas les plus complexes à traiter et les identifier à l'ouverture afin qu'ils soient surveillés plus attentivement.

Concernant la prédiction, la forêt aléatoire permettait une approximation de la charge globale plutôt fiable. Mais sinistre par sinistre c'est plus compliqué.

En effet, avec les estimations des charges sinistres obtenues nous voyons un certain nombre de provisions estimées négatives. Ceci ne peut être possible car nous ne prenons que les règlements en compte et ce brut de recours.

Pour la prédiction de notre charge, les résultats obtenus sont les suivants : notre modèle estime une charge globale d'un peu plus d'un million supérieure au règlements qui ont été effectués jusqu'à présent.

La prédiction est d'une charge ultime pour les ouverts est 9% plus élevée par rapport à la charge à date d'extraction.

Malheureusement, l'estimation pour 22% des sinistres est inférieure à ce qui a déjà été réglé.

Pour tous ces sinistres, la provision associée est positive, ce qui veut dire que le gestionnaire estime qu'il reste encore des règlements associés à ce sinistre.

Si nous nous attardons sur les coûts positifs l'estimation du reste à payer par l'algorithme est 45% plus élevée que celle faite par les gestionnaires. Ce qui est 16 fois plus que la variation de prédiction sur toute la charge.

Comparer les provisions obtenues pour ces sinistres n'a aucun sens car nous déduisons une provision négative (du au fait que l'estimation du coût ultime est inférieure à ce qui a déjà été réglé).

En ce qui concerne les sinistres ayant une charge à la date d'extraction inférieure à 1 000, l'algorithme propose pour ces 19% de sinistres une provision 32 fois supérieure à celle faite par les gestionnaires. Les provisions estimées sont systématiquement au moins 3 à 4 fois supérieures à celles des gestionnaires sauf dans de très rares cas (moins d'une dizaine) où elle est très proche.

Il serait intéressant de suivre l'évolution de ces sinistres pour voir si l'algorithme surestime à juste titre la provision associée à ces sinistres ou si l'estimation est réellement incorrecte.

Il est nécessaire de mettre de côté les sinistres ayant une estimation de la provision négative (23%) et ceux ayant une charge à la date d'extraction inférieure à 1 000 (19%). En effet, ceux qui ont une provision négative ont une vision automatiquement erronée. Pour ceux qui ont une charge inférieure à 1 000, les estimations sont beaucoup trop disparates pour tirer une conclusion.

Les statistiques concernant les écarts entre la provision estimée et la provision des gestionnaires sont les suivantes (en ne prenant en compte que les provisions positives de l'algorithme, soit 79% du portefeuille) se retrouvent figure 50.

Au global, sur ces sinistres, la provision globale prédite algorithmiquement est 7 fois plus élevée que celle estimée par le gestionnaire.

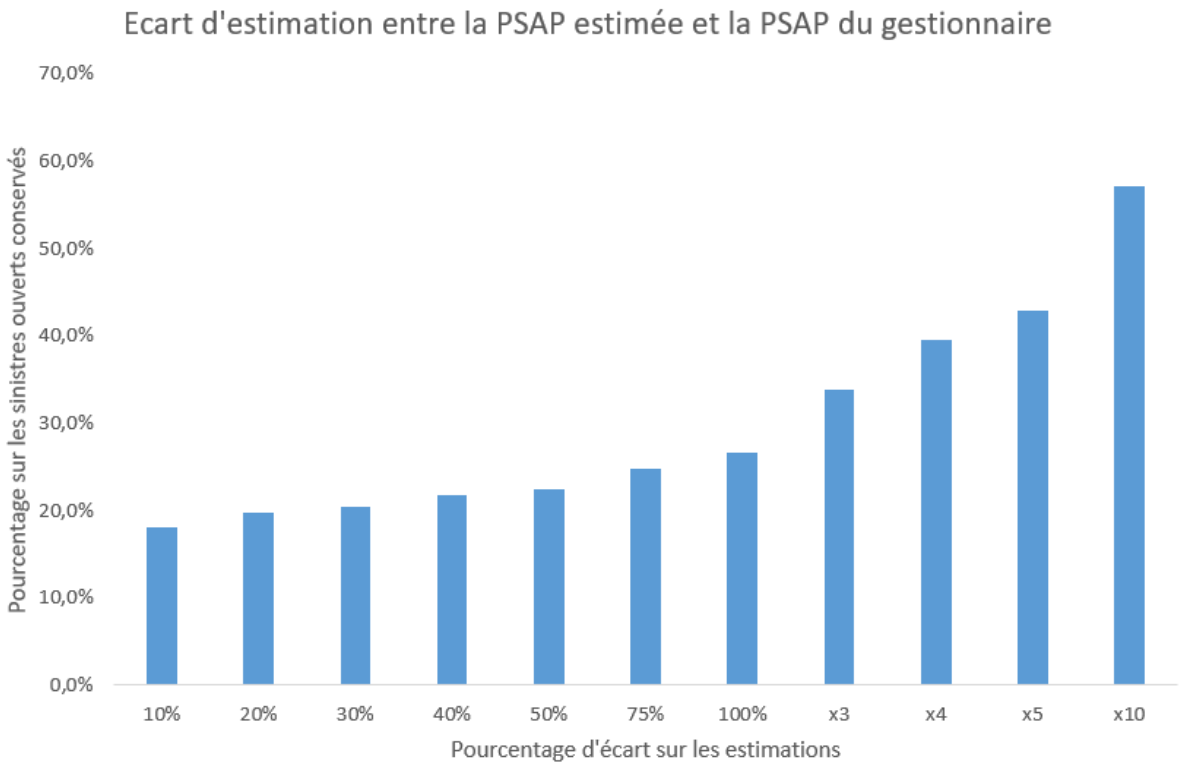

FIGURE 50 – Visualisation des pourcentage de sinistres et leur répartition en fonction de l'écart de prédiction avec la provision du gestionnaire

Nous allons donc nous intéresser aux caractéristiques des sinistres ayant une estimation proche (inférieure à deux fois celle du gestionnaire).

Dans ce cas, l'estimation de l'algorithme est supérieure à celle du gestionnaire 35% du temps. Alors que pour les sinistres où l'écart est supérieur à 100%, l'algorithme prédit une valeur supérieure à celle du gestionnaire dans tous les cas.

Il est malheureusement impossible d'identifier quels sinistres ont spécifiquement un écart faible et quels sinistres ont spécifiquement un écart important. En effet, l'écart peut être faible sur un sinistre à coût faible mais il peut aussi bien être très élevé. Le même postulat est fait sur des sinistres dont le coût est moyen ou élevé.

Afin de tirer plus de conclusions, il faudrait étudier les sinistres et leurs caractéristiques respectives afin de dégager une trame.

III.5 Conclusion

Nous avons pu voir dans cette partie de l'étude que sur une grande partie des sinistres, la prédiction algorithmique n'apportait pas satisfaction.

En effet, les résultats obtenus par cette estimation sont difficilement interprétables, mais encore une fois nous voyons une faiblesse dans les estimations faites par les algorithmes pour des sinistres à coût

faible (inférieur à 1 000). Ici, il a aussi été le cas d'une sous estimation des sinistres graves.

Quant aux sinistres conservés pour une étude plus attentive, nous avons pu constater que les prédictions, dans plus de 70% des cas, étaient plus de 2 fois supérieures aux estimations du gestionnaire.

De plus, nous avons pu voir que sans nous plonger plus précisément dans les sinistres et leurs caractéristiques intrinsèques il n'est à ce stade pas possible de dégager une règle concernant les proximités de prédiction entre l'algorithme et le gestionnaire.

En ce qui concerne l'estimation de la charge globale, les deux points de vue divergent. En considérant tous les sinistres, l'algorithme estime un écart global 10 fois inférieur à celui estimé par le gestionnaire.

Une autre amélioration possible serait dans la modélisation de la censure de la variable Delta en prenant une autre fonction de survie que celle de Kaplan-Meier (celui-ci a été mis en place car n'est pas paramétrique mais un modèle calibré, de Cox par exemple, pourrait être plus approprié) et de comparer les résultats obtenus. Une modélisation de la via les séries temporelles pourrait aussi être envisageable.

Afin d'avoir un avis plus tranché sur la prédiction de la provision sinistre par sinistre obtenue par l'algorithme il faudrait- avoir recours d'une part au backtesting, et d'autre part consulter en gestion les données sinistres dont les écarts entre les deux approches (celle des gestionnaires et des algorithmes) sont les plus matérielles. Concernant le backtesting, cela signifierait tronquer l'étude quelques années avant la date d'extraction et mener la même démarche. Mais pour cela il faudrait avoir des éléments supplémentaires sur les sinistres clos, notamment la provision associée et le montant des règlements à la nouvelle date d'extraction. Une fois cette étude menée et les modèles calibrés il faudrait se ramener à l'étude qui vient d'être faite et nous aurions ainsi un véritable point de comparaison. S'agissant de la consultation des dossiers sinistres du service indemnisation, cela impliquerait de prendre connaissance de l'ensemble des pièces du dossier pour pouvoir se forger une opinion sur l'évaluation. Notamment pour savoir laquelle des deux approches est la plus pertinente et s'il est possible de dégager une nouvelle règle de gestion si le provisionnement dossier/dossier par le service d'indemnisation n'est pas adapté.

Il faudrait de plus pouvoir inclure la somme des règlements dans l'algorithme. Cette optimisation sous contrainte permettrait de ne plus avoir des charges ultimes prédites inférieures à la somme des règlements déjà faits et serait donc plus cohérente avec la réalité.

Conclusion

Durant ce mémoire nous avons eu pour objectif ambitieux de challenger les provisions dossier/dossier établies par les gestionnaires à l'ouverture des dossiers sinistres ainsi qu'à la date d'extraction de la base.

Notre démarche a été faite afin de pouvoir créer un outil permettant à la fonction actuarielle de pouvoir valider les provisions faites par les gestionnaires sinistres. Bien évidemment, l'idée de cette approche n'est pas de se substituer aux autres bonnes pratiques existantes au sein des services d'actariat qui permettent également d'apprécier le provisionnement dossier/dossier telles que le suivi de la liquidation des sinistres avec la surveillance des boni et des mali dossier/dossier ou le sondage direct d'un échantillon de dossier sinistres dans les services d'indemnisation afin de valider ou non la cohérence des provisions dossier/dossier estimées. Notre approche a pour but de mettre en place cette troisième option qu'est la validation par modèle.

Voici un résumé graphique de la démarche qui a été menée tout au long de ce projet ainsi que des étapes réalisées :

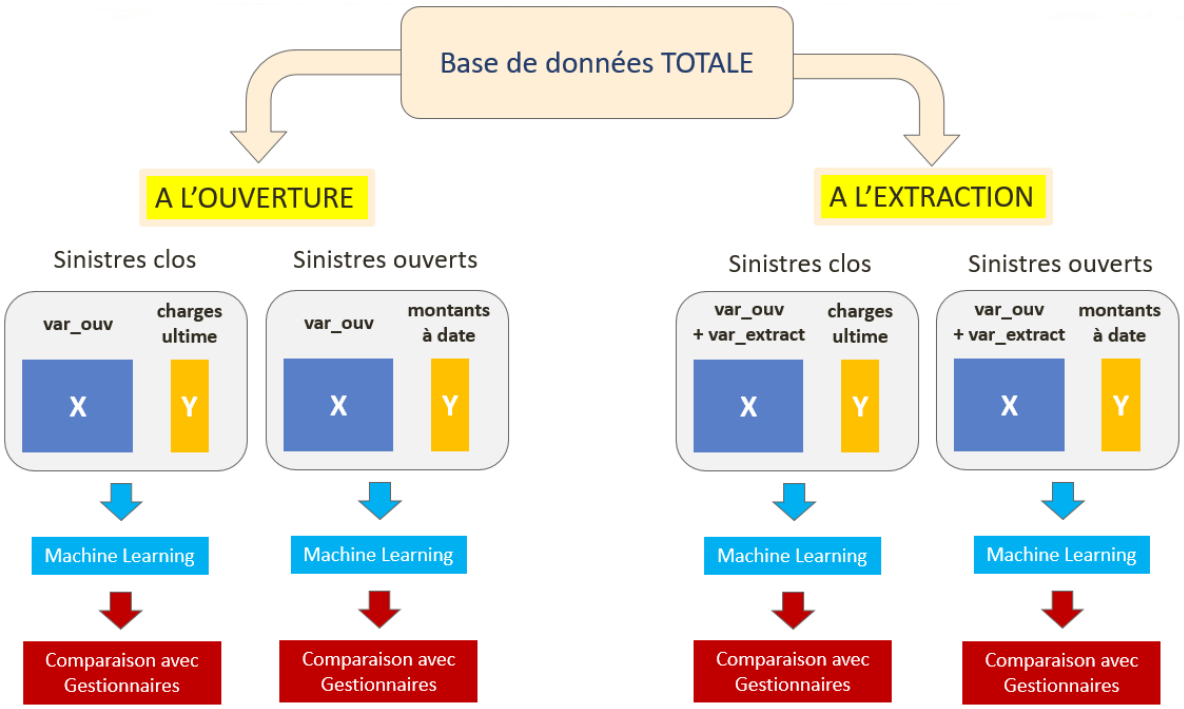

FIGURE 51 – Résumé de la démarche

Les modèles qui s'offraient à nous afin de prendre en compte le maximum de données possibles sont les modèles de machine learning. Nous avons mis en place des modèles d'arbres de régression, des forêts aléatoires, du gradient boosting machine ainsi que des réseaux de neurones.

Ces modèles ont pu prendre une information sinistre ligne à ligne et assimiler des informations provenant de la base sinistre, de la base contrat, de la base client ainsi que la de base bien.

Le seul retraitement fait sur les données de la base finale aura été la prise en compte de l'inflation, celle sur les dommages matériels d'un côté et celle sur les dommages corporels de l'autre. Il aura fallu

ensuite sélectionner les variables pertinentes à notre étude.

A l'ouverture :

Nous avons, tout d'abord, fait une étude sur les sinistres à l'instant d'ouverture où nous avons pu challenger la provision d'ouverture en faisant apprendre nos algorithmes sur les charges ultimes de sinistres clos.

Nous avons vu que les algorithmes arrivaient à prédire une charge ultime globale bien plus précise que les gestionnaires à l'ouverture. Mais seuls les modèles de réseaux de neurones ont eu une prévision sinistre par sinistre légèrement plus précise que les gestionnaires.

Les sinistres les moins coûteux n'ont pas été bien assimilés par les réseaux de neurones. Les gestionnaires avaient donc une meilleure visibilité que les algorithmes sur ces sinistres. Il en est de même pour les sinistres graves voire extrêmes. Le réseau de neurones arrive à prédire un sinistre à coût élevé mais pas de manière aussi précise que les gestionnaires qui ont vraisemblablement plus d'information à leur connaissance que l'algorithme (des rapports d'expertise notamment).

Les réseaux de neurones se sont montrés utiles lorsque les gestionnaires attribuaient au sinistre un coût moyen. Dans toutes les évaluations faites par le réseau de neurones, le résultat était soit plus proche que l'évaluation du gestionnaire ou permettait de déterminer que l'évaluation du gestionnaire sous estimait d'au moins de moitié le coût du sinistre.

Nous avons pu remarquer que les variables les plus explicatives étaient la cause du sinistre, ce qui semble logique, mais aussi le temps de traitement (survenance et déclaration) de l'organisme de l'assurance qui s'occupe du sinistre. Il faut en effet absolument prendre en compte le temps de traitement de l'assureur. Ces temps semblent traduire la complexité du dossier.

La modélisation de l'inflation est cruciale car la distinction entre sinistre matériel ou sinistre corporel est assimilée par le réseau de neurones. Il est donc essentiel de la prendre en compte.

Une piste d'amélioration serait de pouvoir automatiquement relever les sinistres avec un coût très faible pour les exclure. En effet, ceux-ci ne sont pas bien approximés par les réseaux de neurones et déforment donc leur processus d'apprentissage. Il faudrait aussi pouvoir tester plus de réseaux de neurones avec un nombre de neurones et un nombre de couches plus élevés que ceux qui ont été testés ici. En effet, ils permettent de créer des modèles plus précis que les modèles basés sur les arbres. Cette approche est toutefois chronophage et nécessite une appréciation des paramètres des plus délicate.

A la date d'extraction :

En ce qui concerne l'étude à la base d'extraction, une variable a dû être ajustée. En effet, la durée d'ouverture des dossiers est censurée à droite [16] et il a fallu prendre en compte cette censure pour ne pas sous estimer le coût final des sinistres. Nous avons utilisé l'estimateur de Kaplan-Meier pour obtenir une fonction de survie de la durée d'ouverture des dossiers sinistres que nous avons ajustée en conséquence.

Ceci n'a pas été inutile car cette variable s'est révélée être l'une des plus explicatives. C'est même une des variables prédominante dans tous les modèles avec la cause du sinistre et l'intervalle de temps

16. Nous n'observons pas toute la durée du dossier, simplement une durée minimum

entre la survenance du sinistre et sa déclaration.

Nous avons pu voir que la durée d'ouverture du dossier est fortement impactée par le type de dommage ou la responsabilité du client.

Toutefois, les résultats de l'étude à la date d'extraction ne permettent pas de tirer, à ce stade des investigations, des généralités bénéfiques de l'utilisation des modèles.

Ici, le modèle le plus intéressant a été la forêt aléatoire, permettant une prédiction de la charge ultime globale très précise.

Problèmes rencontrés :

Une des difficultés réside dans l'interprétation des résultats. En effet, nous avons estimé la charge globale sur les sinistres clos afin de challenger les provisions dossier/dossier à la date d'ouverture. Nous avons fait cela en prédisant la charge ultime des dossiers ouverts et en leur retranchant les règlements faits ainsi que les honoraires versés. Ce montant a été comparé à la provision dossier/dossier à date d'extraction des gestionnaires (provision sinistre à laquelle nous avons rajouté la provision d'honoraires).

Ces études ayant été faites brutes de recours nous avons pu voir les limites des modèles, notamment dans la seconde partie de l'étude. Les provisions estimées négatives montrent que les modèles testés ne reflètent pas suffisamment la réalité. Compte tenu de ces incohérences nous avons dû nous passer de presque la moitié de notre base de sinistres ouverts (plus de 40%) ce qui ne facilite pas l'interprétation de nos résultats. Il faudrait surtout pouvoir intégrer la somme des règlements déjà fait dans l'algorithme. Une piste d'amélioration est celle explorée dans la parution Non parametric individual claim reserving in insurance de *Maximilien Baudry* et *Christian Y. Robert* où le coût restant à régler est déterminé directement en fonction du coût du sinistre. Leur approche palie à ce type d'incohérence car ils intègrent les règlements déjà faits.

Le problème majeur dans les résultats obtenus dans la seconde partie de l'étude émane du fait que nous n'avons pas d'étalon comme dans la première partie de l'étude. En effet dans la première partie nous savons exactement quel est le coût final du sinistre. Nous pouvons donc voir à quel moment il est plus intéressant de faire confiance au gestionnaire plutôt qu'à l'algorithme et vice-versa.

Dans la seconde partie de l'application numérique, c'est plus complexe car nous nous retrouvons devant deux estimations. D'un côté, la forêt aléatoire, clairement faillible, qui sous estime des provisions de sinistres d'un côté mais qui fait une meilleure estimation de la charge globale et de l'autre un gestionnaire qui sous estime la charge globale mais qui semble mieux maitriser ses dossiers avec les informations supplémentaires dont il dispose.

Axes d'amélioration :

Il faudrait avoir recours au back-testing, qui consiste en la troncature de la base quelques années avant la date d'extraction, pour tester la véracité et la pertinence des modèles utilisés afin de pouvoir voir de manière pertinente si les provisions prédites à l'extraction sont plus pertinentes que les provisions établies pas les gestionnaires. Pouvoir évaluer différentes estimations de l'évolution de la durée d'ouverture des dossiers serait aussi pertinente.

Pour améliorer la modélisation, il faudrait surtout pouvoir prendre en compte les règlements déjà

faits, sous la contrainte que les variables doivent être d'une temporalité commune. En effet, si l'on extrait l'avant dernière somme des règlements sur un sinistre alors celle-ci sera très proche de la charge finale. Ce qui nuit à l'utilisation de la variable dans l'estimation de la charge finale (à cause de la trop forte corrélation entre les deux variables).

L'intégration de variables supplémentaires à l'extraction à disposition, mis à part la durée d'ouverture du dossier n'a pas apporté d'amélioration nette dans la modélisation. En effet, la réouverture ou l'annulation des dossiers n'ont pas pesé en tant que variables pertinentes lors de l'étude.

Mener l'étude avec l'utilisation des recours comme variables supplémentaires pourrait être un axe d'amélioration.

Sur l'ensemble de l'étude il est tout de même positif de voir qu'il est possible de challenger les provisions dossier/dossier à l'ouverture des gestionnaires avec l'utilisation des réseaux de neurones. Une discussion et un back-testing sont primordiaux afin d'interpréter les résultats à la date d'extraction.

Cet outil pourrait donc permettre à la fonction actuarielle d'avoir une meilleure visibilité et une meilleure interprétation, dans certains cas précis, de la cohérence des estimations faites par les gestionnaires de sinistres.

Les pistes d'améliorations sont nombreuses. Nous pouvons mentionner l'utilisation d'un modèle de machine learning : les SVM (*Support Vector Machine* ou Support à Vaste Marge), qui pourraient permettre d'avoir des résultats plus précis. Il est aussi possible de modifier les paramètres des modèles utilisés, notamment pour les GBM en implémentant une fonction réponse plus appropriée aux données. Ici ceci n'a pas été possible car d'autres pistes ont été explorées.

Les réseaux de neurones ont montré ici leur efficacité. Ils possèdent eux aussi de nombreux paramètres qui leurs permettent une extrême appréciation des données. Il faudrait donc pouvoir explorer plus de modèles que ceux qui ont été mises en place ici.

Finalement, il est impératif de mentionner que cette étude est fortement corrélée au volume et à la qualité des données qui ont été mises à disposition lors de cette étude.

En effet, les modèles mis en place ici sont sensibles aux variations de la qualité des données. Ils n'ont pas pour vocation aujourd'hui à remplacer les gestionnaires de sinistres dans leur estimation mais à permettre à la fonction actuarielle d'émettre un regard critique sur ces estimations pour pouvoir formuler des recommandations.

Perspectives :

L'étude menée à travers ce mémoire nous a donc permis de montrer que ceci était possible dans certains cas même si de nombreux axes d'améliorations sont très évidemment possibles.

Il faut en effet mettre en perspective un certain nombre de choses qui ont été faites à travers ce mémoire.

Il est nécessaire que ces algorithmes fassent leurs preuves au niveau des service d'actuariat avant d'envisager une déclinaison opérationnelle dans les services d'indemnisation (le risque opérationnel étant accru étant donné la complexité de la mise en place).

La faculté de faire des backtesting est primordiale afin de pouvoir valider les modèles mis en place et de vérifier que ces derniers reflètent globalement bien la réalité du métier. La qualité des données fait aussi partie des piliers d'intérêt.

Il semble par ailleurs nécessaire de débattre de ce sujet avec les services d'indemnisations et la direction générale.

Ces algorithmes semblent toutefois pouvoir permettre de détecter certains dysfonctionnements qui échapperaient aux autres analyses et ainsi d'adapter certaines règles de gestion de sinistres.

Certains algorithmes prédisent mieux la charge ultime globale mais sont moins pertinents lorsqu'ils font une prédiction ligne à ligne et vice-versa. Seuls ils ne sont donc pas entièrement satisfaisants mais peuvent se compléter les uns les autres.

En revanche, identifier les caractéristiques de certains sinistres mieux prédits par les algorithmes permettrait aussi de revoir et d'adapter certaines règles de gestion. L'algorithme n'est pas systématiquement meilleur que le gestionnaire mais il permet de temps en temps un gain d'information.

Ces algorithmes ne sont pas une science exacte et même s'ils venaient à être déployés un jour dans les services de gestion des sinistres il semble qu'une validation de chaque provision dossier par l'homme soit recommandée.

Pour que ces algorithmes soient déclinés opérationnellement ils faudraient qu'ils soient d'abord validés suite à une phase d'expérimentation en service d'actuariat. Ensuite il est crucial que les personnes mettant en place ces modèles soient en mesure de les comprendre, notamment quels sont leurs points forts et leurs points faibles. De plus, il faut être capable de confronter ces résultats à la vision métier en échangeant avec le service d'indemnisation afin de prendre du recul sur la provision proposée par l'algorithme.

A court terme, l'idée n'est donc pas de directement doter les services de gestion de sinistres de ces algorithmes (dû à leur manque de recul lié à la complexité des méthodes mise en œuvre) mais d'en doter les services d'actuariat qui pourraient les expérimenter afin de compléter leurs études.

Liste des tableaux

1 Exemple d'un extrait de base potentielle 51
2 Information contenues dans la base à l'ouverture du dossier 64
3 Variables déduites de la base à l'ouverture 64
4 Variables retenues par les différentes méthodes de sélection à l'ouverture 65
5 Apparition des variables dans les modèles de sélection à l'ouverture 66
6 Résumé des variables conservées à l'ouverture 66
7 Résultats de la méthode Gestionnaire à l'ouverture 69
8 Résultats des méthodes CART à l'ouverture 69
9 Exemple de sinistres à l'ouverture 72
10 Résultats des méthodes de forêts aléatoires à l'ouverture 72
11 Résultats des méthodes de gradient boosting machine à l'ouverture 74
12 Etude à l'ouverture des dossiers ouverts au 31 décembre 2016 80
13 Récapitulatif des variables dans la base de données à la date de l'extraction 83
14 Résultats des méthodes CART à la date d'extraction 85
15 Résultats des méthodes de forêts aléatoires à la date d'extraction 86
16 Résultats des méthodes de gradient boosting machine à la date d'extraction 87
17 Résultats des méthodes de gradient boosting machine à la date d'extraction 88
18 Statistiques sur la variable Delta 91

Table des figures

1 Les 3 piliers de Solvabilité 2 . . . 3
2 Bilan sous Solvabilité 2 . . . 9
3 Evolution de la vie d'un sinistre . . . 13
4 Challenge de la PSAP à l'ouverture . . . 15
5 Exemple 1 : segmentation d'un univers bivarié . . . 23
6 Exemple 1 : Décision 1 . . . 24
7 Exemple 1 : Décision 2 . . . 24
8 Exemple 1 : Décision 3 . . . 25
9 Exemple 1 : Décision 4 . . . 25
10 Exemple 2 : segmentation d'un univers bivarié . . . 27
11 Exemple 2 : segmentation maximale . . . 27
12 Exemple 2 : segmentation après élagage . . . 28
13 Création d'échantillons Bootstrap . . . 29
14 Processus du Bagging . . . 31
15 Principes de construction d'une forêt aléatoire . . . 34
16 Illustration d'un perceptron simple . . . 40
17 Réseau de neurones multicouches . . . 41
18 Visualisation du biais et de la variance des modèles . . . 47
19 Le compromis biais-variance . . . 48
20 Visualisation à date d'extraction . . . 52
21 Censure non prise en compte . . . 53
22 Censure prise en compte . . . 53
23 Démarche de validation et de comparaison . . . 56
24 Base d'étude . . . 57
25 Indice d'inflation FFB pour les sinistres matériels . . . 59
26 Indice d'inflation SCOR pour les sinistres corporels . . . 60
27 Nombre de sinistres déclarés par année de survenance . . . 60
28 Représentation de la charge sinistre par année de survenance . . . 61
29 Evolution des dossiers clos . . . 62
30 Représentation de l'envergure des coûts des sinistres du portefeuille . . . 62
31 Visualisation du saut de coût . . . 63
32 Visualisation de l'écart entre l'estimation des coûts à l'ouverture et des coûts réels à l'ouverture . . . 67
33 Graphique représentant la provision par rapport au coût réel à l'ouverture . . . 68
34 Ecart entre la charge ultime et la provision à l'ouverture . . . 68
35 Visualisation des variables importantes avec l'arbre élagué à l'ouverture . . . 70
36 Importance des variables pour la forêt aléatoire à l'ouverture . . . 73
37 Importance des variables avec la méthode GBM à l'ouverture . . . 74
38 Evolution du MSE pour un modèle à une couche pour différents decay . . . 75
39 Part de la charge totale prédite par des réseaux de neurones à une couche à l'ouverture . 76
40 Réseau de neurones apportant le meilleur MSE à l'ouverture . . . 77
41 Importance des variables dans le modèle de réseau de neurones retenu à l'ouverture . . . 78
42 Visualisation des variables importantes avec le modèle CART Elagué à la date d'extraction 85
43 Importance des variables avec la méthode de Random Forest à la date d'extraction . . . 87
44 Importance des variables avec la méthode GBM à la date d'extraction . . . 88
45 Meilleur réseau de neurones prédicteur à la date d'extraction . . . 89
46 Importance des variables avec la méthodes des réseaux de neurones à la date d'extraction 90

47 Fonction de survie estimée par Kaplan-Meier avec intervalle de confiance à 95% 91
48 Fonction de survie en fonction du type de dommage 92
49 Fonction de survie en fonction du taux de responsabilité 92
50 Visualisation des pourcentage de sinistres et leur répartition en fonction de l'écart de prédiction avec la provision du gestionnaire 94
51 Résumé de la démarche 97

Références

[1] Maximilien Baudry and Christian Y. Robert. Non parametric individual claim reserving in insurance. Chair Dami, ISFA, 2017.

[2] Guillaume Beneteau. Modèle de Provisionnement sur Données détaillées en Assurance Non-Vie. Towers-Perrin Tillinghast, 2004.

[3] Pierre Connault. Présentation du domaine de recherche ; Algorithme CART, 2008.

[4] Hastie Daniela, Tibshirani Trevor, James Robert, and Jerome Friedman. The Elements of Statistical Learning, 2017.

[5] Gaël Gibaud and Mathieu Wolf. Research in Property & Casualty Reserving Determination of the reserves for a fire risk portfolio. ISFA, 2016.

[6] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning with Applications in R, 2017.

[7] Raphaël Lagier. Provisionnement non-vie sur la branche Responsabilité Civile Professionnelle. Predica, 2013.

[8] A. Le Tesson, A. Lenain, S. Samba, and J. Ung. Estimation de l'erreur de prédiction dans le cas de l'utilisation d'une combinaison de méthodes pour le calcul de provisions en assurance IARD. EURIA, 2014.

[9] Olivier Lopez, Xavier Milhaud, and Pierre-Emmanuel Thérond. Tree based censored resgression with applications to insurance, 2015.

[10] Xavier Milhaud. Techniques déterministes et stochastiques de provisionnement en assurance. ISFA, 2016.

[11] Xavier Milhaud. Cours de Data Science. ISFA, 2017.

[12] Frédéric Planchet. Modèles de durée : statistiques de modèles non paramétriques. ISFA, 2016.

[13] Eric Rakotomalala. Arbres De Décision. Université Lumière Lyon 2.

[14] Eric Rakotomalala. Arbres de Régression. Université Lumière Lyon 2.

[15] Eric Rakotomalala. Bagging - RandomForest - Boosting, Techniques ensemblistes pour l'analyse prédictive. Université Lumière Lyon 2.

[16] Eric Rakotomalala. Gradient Boosting. Université Lumière Lyon 2.

[17] Eric Rakotomalala. Les méthodes d'Inductions d'Arbres. Université Lumière Lyon 2.

[18] Eric Rakotomalala. Réseaux de neurones artificiels, perceptrons simple et multi-couches. Université Lumière Lyon 2.

[19] Eric Rakotomalala. Support Vector Machine. Université Lumière Lyon 2.

[20] Noémie Rose. Provisionnement en assurance non-vie : Utilisation de modèles paramétriques censurés. Winter & Associés, 2009.

[21] Alex Smola and Bernhard Schölkopf. A tutorial on support vector regression, 2003.

[22] Thérèse Winterholer. Responsabilité Civile Médicale : Enjeux actuels sur le provisionnement des sinistres. Modélisation des trajectoires individuelles. AXA, 2012.