Calcul et projection d'un score

Auteur : Théo Mardoc

Maître d'apprentissage : Aurélien Bellet
Tutrice pédagogique : Myriam Maumy-Bertrand

2020

Résumé

La mise en place d'un score pour évaluer les sociétaires commence à se répandre dans les organismes d'assurances. Chaque entreprise a sa propre formule qui se base généralement sur les cotisations reçues et le coût des sinistres survenus.

Sous la forme d'un solde cumulable année après année, le score conçu dans ce mémoire apporte de nombreuses déductions sur sa dynamique que ce soit pour un sociétaire, un groupe de sociétaire ou pour le portefeuille entier. Une analyse du score sur chaque produit est réalisable mais aussi l'analyse d'une combinaison de plusieurs produits et ce pour différentes périodes de temps.

Ce nouvel outil est idéal pour identifier et sélectionner au mieux les mauvais risques dont il est préférable de se séparer, et au contraire, sélectionner les bons risques à conserver dans le portefeuille. Si certains sociétaires peuvent être considérés comme de mauvais risques pour une branche, ce n'est pas toujours le cas une fois l'ensemble des contrats qu'ils détiennent pris en compte. L'intégration des frais et l'agrégation des scores par foyer sont des idées novatrices pour acquérir une vision exhaustive de l'état d'un sociétaire.

Renforcée de quelques indicateurs, la valeur actuelle permet une projection convenable de ce score à plusieurs horizons à l'aide de régressions logistiques, de forêts aléatoires ou de matrices de transition empiriques. Les valeurs futures peuvent alors influencer certaines décisions présentes. Bien qu'il soit naissant, le score est utilisable sous sa forme actuelle et montre son potentiel pour les années à venir.

Mots clés :

Score sociétaire, Frais, Foyer, Valeur actuelle, analyse descriptive, Valeur Future, Régression Logistique, Forêt aléatoire, Matrice de transition.

Abstract

Setting up a score to rate insurance members starts spreading across insurance organisations. Each company got its own formula which is usually based on received contributions and the cost of insurance claims.

Shaped as a cumulative balance made year after year, the master's thesis' score brings numerous conclusions about its dynamic which is for an insurance member, a group of insurance members or for the whole insurance portfolio. A score's analysis on each product is doable but also the analysis of a combination of products and that for different time periods.

This new tool is ideal to flag and select at best the bad risks that it is better to dissociate with, and at the contrary, select the good risks to keep in the portfolio. If some insurance members could be considered as bad risks to an insurance branch, it's not always the case once all contracts put together. The integration of the fees and the aggregation of scores by household are new ideas to get an exhaustive vision of the member's value.

Reinforced with some indicators, the present value makes a suitable projection of this score possible to multiple time horizons using logistic regression, random forest or empirical transition matrices. Then, future values can influence some today's decisions. Even though it's incipient, the score can be use in its actual form and shows potential for future years to come.

Key-words :

Member's score, Fees, Household, Present value, Descriptive analysis, Future value, Logistic Regression, Random forest, transition matrices.

Note de synthèse

Introduction

Aujourd'hui, la tarification en assurance est à la fois maîtrisée et réglementée. La cotisation d'un contrat d'assurance est déterminée selon le niveau de sinistralité anticipée par l'assureur. Cependant, la sinistralité réelle peut se révéler différente de celle attendue.

Afin d'identifier au mieux les bons et mauvais risques, plusieurs assureurs et bancassureurs ont récemment mis au point un système de score qui leur est propre. L'idée est d'attribuer un score à chaque sociétaire basé sur ses cotisations et ses sinistres. Ce nouvel indicateur peut alors être utilisé à différentes fins commerciales.

L'objectif de la Matmut est de concevoir son propre score. À l'inverse de ses concurrents, ce score n'évaluera pas un unique sociétaire mais l'ensemble de son foyer sur tous les contrats souscrits. Une fois mis en place, il sera plus facile d'identifier les contrats à résilier et les foyers éligibles aux gestes commerciaux.

Un score exhaustif

Le score a pour but d'intégrer les cotisations, les sinistres et les frais de tous les produits détenus par le foyer étudié. Il est composé de la Valeur Actuelle calculée à partir des exercices passés, et de la Valeur Future obtenue par projection sur les exercices futurs.

Pour connaître le montant des frais à affecter, une ventilation des frais généraux se fait par destination comptable que sont les frais d'acquisitions, de gestion des sinistres et d'administration. Les frais sont ensuite redistribués proportionnellement à l'activité de gestion nécessaire à chaque contrat.

Pour former les foyers de sociétaires, certaines données déclarées par l'assuré sont utilisées afin d'en déduire les relations filiales ou conjugales existantes. Le mauvais score d'un sociétaire peut alors être compenser par les autres scores de son foyer lui évitant dans des cas extrêmes d'être résilié, poussant les autres membres de sa famille vers la concurrence.

Calcul et étude du score

Une des étapes de conception du score est de calculer la différence entre les cotisations acquises et le coût des sinistres pour chaque sociétaire par année d'exercice passée. Une séparation par produit est également applicable. Pour obtenir la vision cumulée qui sera étudiée, les résultat d'un même sociétaire sont sommés année après année. Les scores peuvent ensuite être regroupés soit par année de survenance soit par maturité, c'est à dire en fonction du nombre d'années de construction du score.

Plus que le score lui-même, c'est l'évolution de ce dernier qui est intéressant. Les distributions par maturité permettent de visualiser la hausse linéaire de la moyenne et logarithmique de l'écart-type.

Plusieurs indicateurs de cette évolution ont été créés à partir du score cumulé afin de résumé l'état de chaque sociétaire. Savoir si le score a toujours été positif ou combien d'année le score a été négatif est alors possible. Ces indicateurs servent à la fois d'un point de vue macroscopique lorsqu'une tendance est étudiée, mais aussi d'un point de vue individuel lorsque l'état d'un sociétaire particulier est observé.

Les bons risques auront un score croissant avec le temps et sont majoritairement détenteur de plusieurs contrats tandis que les mauvais auront un score plus oscillant à cause d'une sinistralité importante.

Identifier les meilleurs risques présents ou futurs est essentiel, ce sont ceux à conserver en priorité. Avec le score, il est plus simple de les sélectionner et de les rendre éligibles à un geste commercial en cas de menace de résiliation de leur part. Il s'agit de la défense de portefeuille.

À l'inverse, le score peut aussi être utilisé pour identifier les contrats qui auraient dû être résiliés par l'assureur grâce à un outil et des indicateurs plus pertinent que ceux déjà en place.

Projection du score

Les différentes valeurs et indicateurs mis en place servent également à projeter le score de l'exercice présent vers celui des exercices futurs. Pour cela deux méthodes alimentées par ces variables et une troisième plus empirique qui n'a besoin que du score cumulé présent sont étudiées.

La première approche est la projection par régression logistique. Ce modèle particulier de modèle linéaire généralisé estime si le score d'un sera positif ou non dans les années à venir. Après calibration et d'après les AUC calculés, plus le modèle projette loin, moins il est précis. En revanche, plus la maturité de départ est élevée, plus le modèle est de qualité.

La deuxième approche est celle des forêts aléatoires. Elle utilise des algorithmes de *machine learning* pour générer une multitude d'arbres de décisions qui segmentent de manière homogène les données d'apprentissage pour créer un modèle de projection robuste une fois bien calibré avec suffisamment de données. Une forêt de classification aura le même rôle que la régression logistique avec de meilleurs résultats alors qu'une forêt de régression donnera en sortie un score futur de même nature que celui présent.

La troisième approche utilise des matrices de transition empiriques. Tout d'abord, les scores cumulés d'une même maturité sont classés puis divisés en rang selon leur classement. Les probabilités de passer d'un rang à l'autre avec le temps est calculé et permet de construire des matrices stochastiques. Il est alors possible de calculer soit le score moyen de l'espérance du rang futur sachant le rang présent soit l'espérance du score moyen à une maturité future sachant le rang présent. Avec le nombre de données important disposé, les résultats obtenus sont bons avec vingt rangs. Cependant, cela limite le nombre de sorties à vingt.

Conclusion

Dans son état actuel le score permet déjà une évaluation de chaque sociétaire qui sera renforcé avec l'ajout effectif des frais et de la dimension foyer dans son calcul. Il respecte l'idée d'un score simple à comprendre, utiliser et interpréter.

L'outil mis en place pour sa conception est également flexible dans son utilisation puisqu'il peut s'adapter aux différents produits et périodes d'exercices étudiés.

Plusieurs pistes de projections ont été abordées, non sans l'aide de la quantité importante de données disponible, ouvrant la voie à de futurs modèles. L'utilisation des forêts aléatoires semble être la meilleure approche parmi les trois étudiées.

Summary

Introduction

Today, insurance pricing is both mastered and regulated. The contribution of an insurance contract is determined according to the level of loss experience anticipated by the insurer. However, the actual loss experience may turn out to be different from the one expected.

In order to improve good and bad risks, several insurers and bank insurers have recently developed a scoring system of their own. The idea is to assign a score to each insurance member based on their contributions and claims. This new indicator can then be used for different business purposes.

Matmut's goal is to design its own score. Unlike its competitors, this score will not evaluate a single insurance member but the entire household on all the contracts taken out. Once set up, it will be easier to verify the contracts that need to be terminated and the households eligible for commercial gestures.

An exhaustive score

The score's target is to integrate the contributions, the claims and the expenses of all the products held by the examined household. It is made up of the Present Value calculated on the basis of past financial years, and of the Future Value realized by projection on future years.

To find out the amount of costs to be allocated, general costs are broken down by accounting destination, namely acquisition costs, claims management and administration costs. The fees are then redistributed in proportion to the management activity required for each contract.

To define households of insurance members, some data declared by the insured are used in order to deduce family links. The bad score of a insurance member can then be compensated by the other scores of his household, preventing him in extreme cases from being terminated, pushing the other members of his family to the competitor.

Calculation and study of the score

One of the steps in the design of the score is to calculate the difference between the contributions acquired and the cost of claims for each member by past financial year. A separation by product is also applicable. To obtain the cumulative vision that will be studied, the results of the same member are summed up year after year. The scores can then be grouped either by year of occurrence or by maturity, ie according to the number of years of construction of the score.

More than the score itself, it is the evolution of the latter that is interesting. The distributions by maturity allow us to visualize the linear increase in the mean and the logarithmic increase in the standard deviation.

Several indicators of this evolution were created from the cumulative score in order to summarize the status of each insurance member. Knowing if the score has always been positive or for how many years the score has been negative is then possible. These indicators are used both from a macroscopic point of view when a trend is studied, but also from an individual point of view when the state of a particular insurance member is observed.

The good risks have an increasing score over time and are mostly holders of several contracts while the bad ones have a more oscillating score because of a significant loss experience.

Identifying the best present or future risks is essential, they are the ones to be kept as a priority. With the score, it is easier to select and make eligible for a commercial gesture in the event of a threat of termination on their part. This is portfolio defense.

On the contrary, the score can also be used to identify contracts that should have been terminated by the insurer using a tool and indicators that are more relevant than those already in place.

Score projection

The various values and indicators set up are also used to project the current financial year score towards that of future financial years. For this, two methods fed by these variables and a third more empirical which only needs the present cumulative score are studied.

The first approach is the projection by logistic regression. This particular model of generalized linear model estimates whether a score is positive or not in the coming years. After calibration and based on the calculated AUCs, the further the model projects, the less accurate it is. On the other hand, the higher the initial maturity, the better the quality of the model.

The second approach is the random forests one. It uses machine learning algorithms to generate a multitude of decision trees that homogeneously segment the training data to create a robust projection model when properly calibrated with enough data. A classification forest will have the same role as logistic regression with better results while a regression forest produces a future score of the same type as the present output.

The third approach uses empirical transition matrices. First of all, the cumulative scores of the same maturity are classified and then divided into rank according to their classification. The probability of moving from one rank to another over time is calculated and allows the construction of stochastic matrices. It is then possible to calculate either the average score of the expected future rank knowing the present rank or the expected mean score at a future maturity knowing the present rank. With the large amount of data arranged, the results obtained are good with twenty ranks. However, this limits the number of outputs to twenty.

Conclusion

In its current state, the score already allows an evaluation of each insurance member, which will be reinforced with the effective addition of costs and the household dimension in its calculation. It respects the idea of a score that is easy to understand, use and interpret.

The tool set up for its design is also flexible in its use since it can be adapted to the different products and periods of financial years studied.

Several thought of projections were discussed, not without the help of the large amount of data available, paving the way for future models. The use of random forests seems to be the best approach among the three studied.

Remerciements

Je tiens en premier lieu à remercier Aurélien Bellet, responsable du pôle tarification de la Matmut et maître d'apprentissage pour sa patience et ses encouragements.

J'aimerai également remercier toutes les personnes du pôle tarification pour leur accueil et leur soutien.

Je remercie l'université de Strasbourg et le DUAS, sa formation en actuariat, de m'avoir accepté et formé ces trois dernières années.

Enfin je remercie ma famille et mes amis pour leur soutien permanent.

Table des matières

Chapitre 1

Introduction

La tarification en assurance est une pratique aujourd'hui bien encadrée et se base sur un ensemble de segmentations pour déterminer la cotisation de chaque contrat. Chaque sociétaire détient alors un ou plusieurs contrats dont la cotisation diffère en fonction des risques assurés et des garanties qui les accompagnent. Les contrats aux cotisations élevées indiquent donc en théorie les plus mauvais risques selon les modèles de tarification. Néanmoins, ce n'est pas toujours le cas à posteriori et se référer uniquement à la cotisation est une erreur car la dimension sinistre est également à prendre en compte.

C'est pourquoi depuis une dizaine d'année une nouvelle pratique spécialisée dans l'identification des bons et mauvais risques, émerge chez les assureurs et bancassureurs. Cette pratique attribue un score propre à chaque sociétaire déterminé par le $\frac{S}{C}$ (CARIA 2012) ou encore par les bénéfices futurs espérés (VANNEAUX 2010) qui dépendent eux même des cotisations et des sinistres futurs de chaque individu. Les objectifs sont divers et dépendent du besoin des entreprises. Cette pratique peut servir à mieux cibler les sociétaires afin de leur proposer un contrat spécifique, mais encore à corriger la tarification pour les années suivantes et à diriger les rabais commerciaux (DURAND 2016).

Tous les scores précédents sont des indicateurs visibles uniquement par l'assureur mais il faut savoir que de nouveaux scores, différents de ces derniers, apparaissent sur le marché de l'assurance dont la valeur est connue des deux parties. Le meilleur exemple est le score de conduite calculé à partir de données recensées directement depuis un boitier à l'intérieur du véhicule. Ce score de conduite impacte généralement la cotisation du conducteur de façon plus directe et sert d'autres objectifs que le score étudié dans ce mémoire.

De par sa nouveauté, la formule du score n'est pas encore normée ce qui laisse une grande liberté dans sa création. Cependant, la plupart des scores rencontrés ne se limitent qu'à un produit ou un groupe de produits tels que des contrats liés soit à l'automobile, soit à l'habitation. De même, la valeur est généralement attribuée à un unique sociétaire. L'ambition dans ce mémoire est de mettre en place un score le plus complet possible qui prend en compte un maximum de contrats et qui étend sa valeur d'une dimension individuelle à une dimension foyer, réunissant plusieurs sociétaires liés entre eux par une relation familiale. La volonté est également d'établir un outil polyvalent et flexible dans son utilisation. Il peut tout aussi bien servir à désigner les sociétaires à résilier dans le cadre de la surveillance de portefeuille, qu'à désigner les sociétaires à retenir dans le cadre de la défense de portefeuille.

Dans un premier temps, ce mémoire expose de manière plus précise la vision du score de la Matmut. Cette partie mettra en lumière la formule du score proposée et plusieurs éléments qui la compose comme l'intégration des frais et l'agrégation des scores par foyer de sociétaires.

Ensuite, la mise en place pratique du score sera détaillée, de la création de la base d'étude au calcul de la valeur actuelle du score. Puis, le score réalisé sera étudié dans le but de montrer concrètement les informations qu'il renferme et les utilisations possibles à travers les deux cas évoqués plus tôt.

Enfin, plusieurs méthodes de projection qui répondent à ces usages seront étudiées. Un modèle de régression logistique et un modèle de classification binaire qui se base sur les forêts aléatoires seront mis en œuvre pour déterminer les sociétaires à résilier. D'autres modèles, plus expérimentaux, basés sur le classement des sociétaires selon le score seront appliqués pour projeter le score dans le temps.

Chapitre 2

Un score exhaustif

La plus-value du score conçu par et pour la Matmut est que celui-ci ne se limite pas qu'à un produit, ni même à un sociétaire. L'entreprise a pour volonté d'observer la dynamique de ce score de la manière la plus exhaustive possible, en recoupant toute l'information disponible sur différentes branches d'assurance comme l'IARD, la Santé, l'Entreprise et la Vie. Le but est de tirer un maximum d'informations passés sur la situation de chaque sociétaire pour mieux anticiper les situations futures. Ainsi il sera plus facile pour la Matmut d'agir en conséquence.

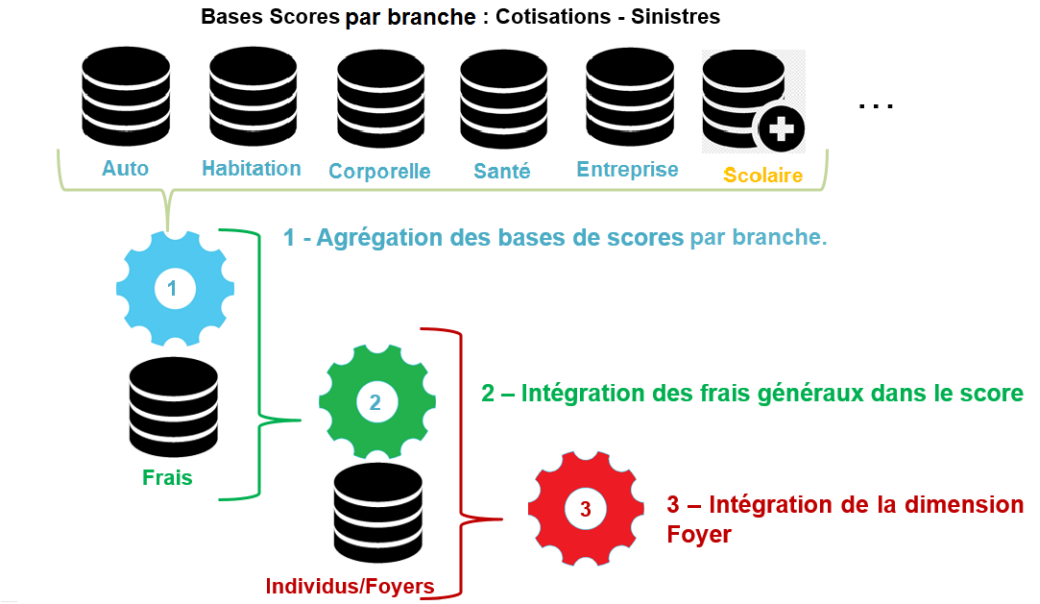

FIGURE 2.1 – Les paliers majeurs de la conception du score exhaustif

Trois grandes étapes sont nécessaires pour la création du score à son plein potentiel. Cependant, la préparation des frais ou l'identification des foyers sont des travaux parallèles au calcul des scores par branche qui seront ensuite agrégés dans une base commune.

2.1 Vision détaillée du score final

L'idée derrière ce score est d'évaluer un sociétaire ou le foyer auquel il appartient, à la fois sur ses états passés mais aussi sur sa valeur future potentielle déterminer par des modèles de projections. Le tout forme alors un score complet utilisable par l'assureur dans divers cas pratiques.

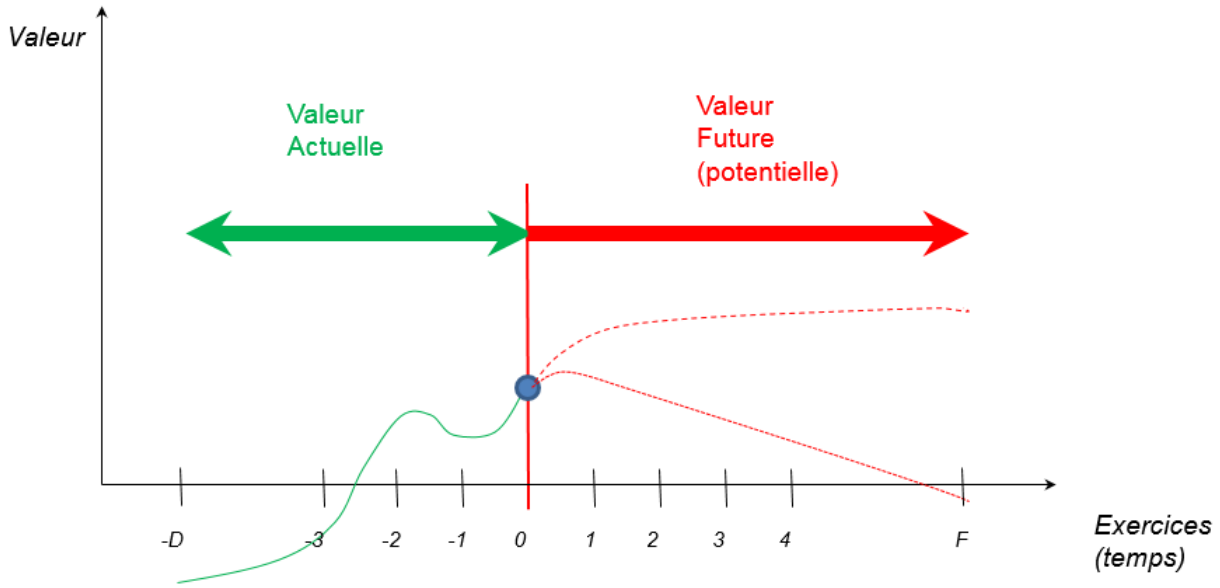

FIGURE 2.2 – Illustration de la valeur actuelle et future en fonction du temps

Formules du score :

— $ScoreComplet = ValeurActuelle + ValeurFuture$

— $ValeurActuelle = \sum_{i=1}^{n} \left(\sum_{t=-D}^{-1} (C_{t,i} - S_{t,i} - F_{t,i}) - A_i \right)$

— $ValeurFuture = \sum_{i=1}^{n} \left(\sum_{t=1}^{F} (E(C_{t,i} - S_{t,i} - F_{t,i}) * \mu_{t,i} * (1 + \sigma_{t,i})) \right)$

Avec :

— D la durée depuis la prise d'effet du contrat, ou la date de début de période d'étude.

— F le nombre d'exercices projetés.

— n le nombre de produits évalués.

— $C_{t,i}$ la cotisation de l'exercice t pour le produit i.

— $S_{t,i}$ les charges sinistres de l'exercice t pour le produit i.

— $F_{t,i}$ les frais d'administrations et de gestion des sinistres de l'exercice t pour le produit i.

— A_i les frais d'acquisitions pour le produit i.

— $\mu_{t,i}$ la probabilité que le produit i soit encore en cours à l'exercice t.

— $\sigma_{t,i}$ la probabilité de souscription d'un contrat i avant ou pendant l'exercice t.

Les probabilités de rétentions et de souscriptions sont indéterminées et nécessiteraient leurs propres modèles. En attendant, les projections mises en œuvre suivent l'hypothèse d'une impossibilité pour un sociétaire de sortir totalement du portefeuille.

2.2 Les frais

L'intégration et la répartition des frais est une étape majeure dans la conception du score. Sans compter l'indemnisation des sinistres, l'entreprise est obligée d'effectuer certaines dépenses inhérentes à son activité. Ces dépenses sont alors ventilées vers trois destinations comptables à savoir :

— Les frais d'acquisitions, qui correspondent principalement à la rémunération des réseaux de distribution pour attirer de nouveaux sociétaires ou pour multi-équiper des sociétaires déjà présents dans le portefeuille.

— Les frais de gestion des sinistres, qui correspondent schématiquement aux salaires des gestionnaires de sinistres mais aussi aux outils mis à leur disposition.

— Les frais d'administration, qui correspondent aux sommes engagées par l'entreprise pour maintenir le bon état de l'activité et des services rendus aux sociétaires. Le maintien de l'espace personnel en ligne et les coûts relatifs à la mise à disposition de l'attestation d'assurance ou des procédures de changements de coordonnées sont ventilés en partie dans ces frais d'administration.

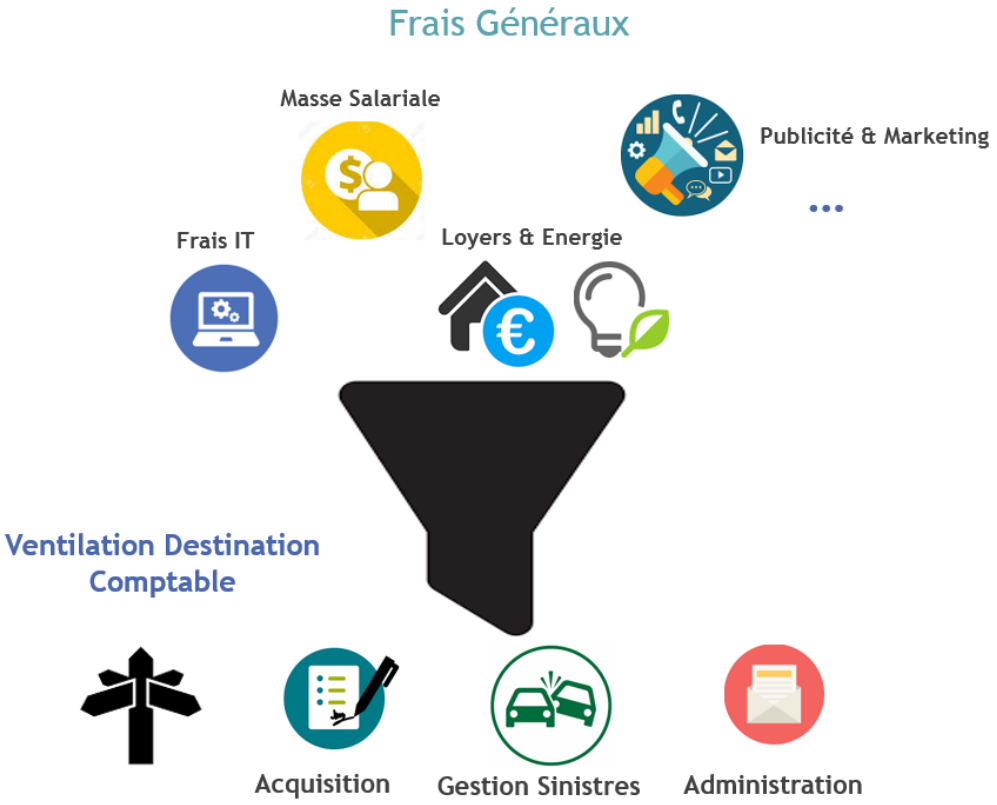

FIGURE 2.3 – Schéma de la ventilation des frais

Une fois cette première ventilation par destination réalisée, chaque frais est redistribué dans les branches qui les concernent. Ainsi les frais de gestion de sinistres sont répartis au prorata des actes de gestion. L'étape suivante consiste à reconstruire une vision sociétaire des frais selon l'attention que requièrent leur indemnisation.

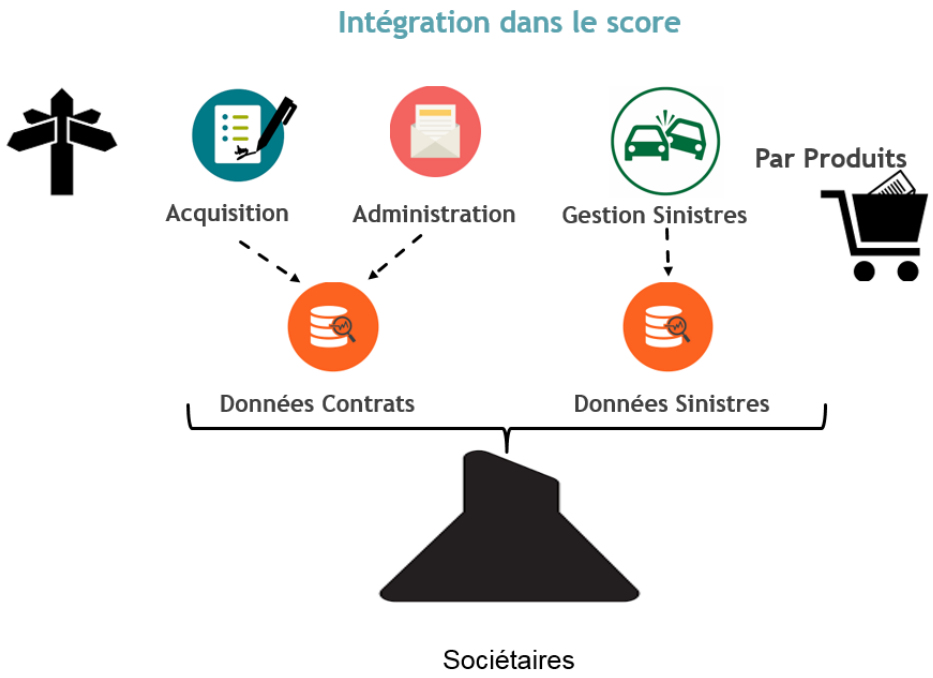

FIGURE 2.4 – Schéma de l'intégration des frais dans le score

Plus simplement, à chaque nouveau contrat un sociétaire se verra attribuer des frais d'acquisition. Puis chaque année, des frais d'administration en fonction de grandeurs fixes telles que les frais d'émission, ou variables comme le nombre d'opérations que le sociétaire réalise sur ses contrats. Enfin, si le sociétaire est sinistré, une part des frais de gestion des sinistres de la branche concernée lui sera associée.

2.3 La dimension foyer

Le score d'une branche ne définit pas forcément l'état du score d'un sociétaire, ce qui incite à la création d'un score prenant en compte plusieurs branches. De la même façon, le regroupement de plusieurs sociétaires est possible pour étudier le score de groupes d'individus.

L'idée est de rassembler des sociétaires qui interagissent entre eux régulièrement pour calculer un score par foyer. Avec une vision par foyer, il est plus facile d'identifier les besoins en assurance. En effet, pour deux personnes qui vivent ensemble une seule souscrira une assurance habitation pour leur résidence principale, ce qui rend la souscription pour un contrat habitation moins probable pour la deuxième personne.

Il existe des cas où un sociétaire considéré comme un mauvais risque soit relier à plusieurs très bons risques. Dans ce type de situation, il peut être dangereux de résilier le mauvais risque si cela pousse les très bons risques vers la concurrence. Ce lien de présence est d'ailleurs représenté sous la forme d'un coefficient dans le mémoire de Choquet (CHOQUET 2011) en addition de la corrélation entre la sinistralité des parents et celle de leurs enfants pour les produits Auto.

Certains numéros de sociétaires font finalement référence à une même personne physique, ça peut être le cas lorsque la personne quitte totalement le portefeuille avant de revenir ou plus simplement si l'individu est directement associé à une personne morale tel quelqu'un qui détiendrait une assurance auto à son nom et une assurance professionnelle au nom de son entreprise.

Il est alors possible de relier les sociétaires entre eux en utilisant les coordonnées postales, ou bien des noms des bénéficiaires de contrats comme l'assurance scolaire pour la filiation ou les noms des différents conducteurs autorisés pour l'assurance auto. Deux types de relations sont ensuite établies :

— Les relations intra-foyer pour tous ceux qui vivent sous le même toit. (Exemple : conjoints, parents et enfants...)

— Les relation inter-foyers qui relient des personnes de la même famille ou très proches qui ne vivent pas sous le même toit. (Exemple : cousins, grands-parents et petits enfants...)

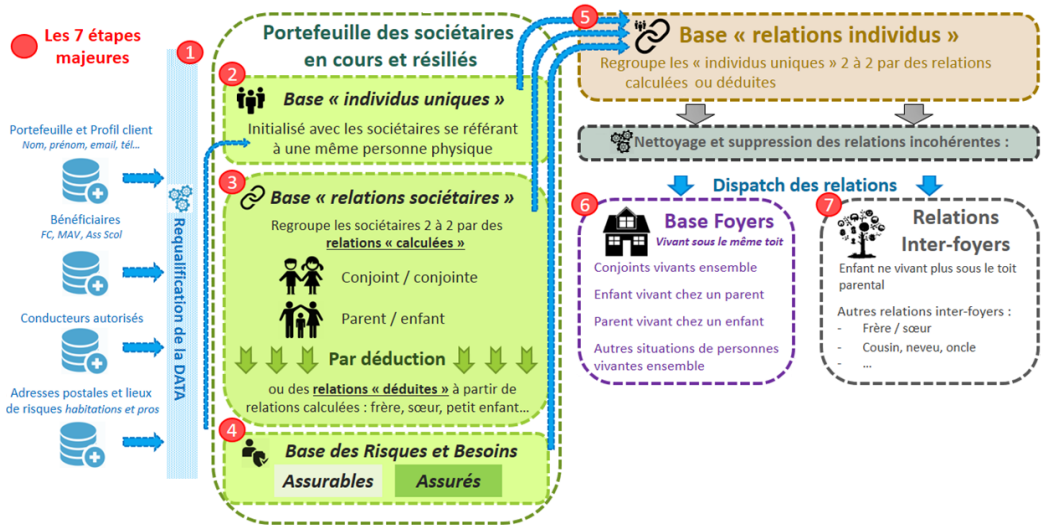

FIGURE 2.5 – Schéma d'identification des foyers de sociétaires

Chapitre 3

Création de la base d'étude

La base de référence pour les scores à calculer est composée de la façon suivante. Chaque observation correspond à l'état d'un sociétaire pour une année d'exercice. Chaque colonne correspond à une branche dans laquelle est inscrit un autre score que celui recherché. Cet autre score est appelé score décumulé et résulte de la différence entre les cotisations et les coûts de chaque sinistre survenu, s'ils existent, le tout pour chaque sociétaire sur l'année d'exercice observée. À partir de cette base, toutes les sous bases sont possibles en fonction des sous-ensembles de branches et d'années voulus.

Il reste à définir quelles branches sont incluses dans cette base dans cette étude, et comment sont recenser les sinistres et les cotisations.

Toutes les manipulations des bases sont effectuées avec SAS Entreprise Guide 7.1.

3.1 Liste des branches étudiées

Une branche est définie ici comme étant une catégorie de contrat. Dans une même branche plusieurs types de contrats de même nature coexistent. Dans cette étude, cinq branches distinctes seront mentionnées :

— La branche MRSQ regroupe les contrats qui concernent l'ensemble des véhicules terrestres à moteur, ainsi que les remorques et les caravanes. Les scores associés à cette branche sont calculables à partir de 2003, et ce jusqu'à 2019.

— La branche MGAR regroupe l'ensemble des contrats habitation ou les contrats qui concernent les terrains. Les scores associés à cette branche sont calculables à partir de 2003, et ce jusqu'à 2019.

— La branche FC regroupe une première gamme de contrats assurant contre les accidents corporels. Les scores associés à cette branche sont calculables à partir de 2011, et ce jusqu'à 2019.

— La branche MAV regroupe une gamme supérieure de contrats assurant contre les accidents corporels. Les scores associés à cette branche sont calculables à partir de 2008, et ce jusqu'à 2019.

— La branche MLCO regroupe les contrats qui concernent l'ensemble des professionnels et des entreprises. Les scores associés à cette branche sont calculables à partir de 2008, et ce jusqu'à 2019.

3.2 Détermination du périmètre de l'étude

La première étape dans la mise en place du score est de définir le périmètre de l'étude. Un des objectifs est de donner un score pour chaque sociétaire étudié et cela pour chaque année civile observée. Le périmètre est déterminé par les sociétaires étudiés et les années observées.

3.2.1 Choix de la fenêtre d'observation

Le choix des années est restreint par la disponibilité des données nécessaires à la mise en place du score. En effet, l'informatisation et la conservation des données sous la forme actuelle ne permettent pas de remonter avant les années 2000. Dans le meilleur des cas, il est possible de calculer un score à partir de 2003. Par conséquent, les sociétaires ayant quitté le portefeuille avant 2003 sont exclus du périmètre d'étude.

Afin de conserver au maximum les sociétaires pouvant être étudiés et afin d'obtenir une profondeur d'historique suffisamment importante pour la suite de l'étude, il est décidé d'étendre la fenêtre d'observation le plus possible en utilisant toutes les années où les données sont disponibles. Concernant la fin de la fenêtre d'observation, elle est aujourd'hui fixée au 31 décembre 2019 inclus, la dernière année civile révolue peut alors être prise en compte.

3.2.2 Contraintes sur les sociétaires évalués

Maintenant que les dates auxquelles seront calculés les scores sont connues, il faut s'intéresser aux sociétaires. Deux grandes interrogations existent sur les individus à étudier :

— Les sociétaires qui ne sont plus présents dans le portefeuille à la dernière date d'observation doivent-ils faire partie de l'étude ?

Actuellement, près de trois millions de sociétaires composent le portefeuille, c'est moins de la moitié de l'ensemble des sociétaires observables dans la fenêtre d'observation. Sachant qu'à terme le score ne sera pas uniquement une vision actuelle de la valeur du sociétaire mais sera également utilisé pour projeter la valeur des sociétaires encore présents. Prendre en compte les sociétaires résiliés pour enrichir la base d'étude est utile dans l'analyse du comportement du score. Se priver des adhérents résiliés entre 2003 et 2019, c'est se priver d'une quantité non-négligeable d'information. De plus, une vision du score par groupe de sociétaires d'un même foyer est prévue dans le futur. Il est fort probable que certains sociétaires aujourd'hui résiliés soient liés à d'autres encore présents.

— Les sociétaires qui sont entrés dans le portefeuille avant la date pivot doivent-ils faire partie de l'étude ?

Il y a environ 2,7 millions de sociétaires qui sont entrés dans le portefeuille avant 2003 dont près de 400 000 qui ont résiliés avant 2003. Ce qui donne 2,3 millions d'individus supplémentaires dont le score peut être étudié. Cela représente plus d'un tiers de l'ensemble des sociétaires observables dans la fenêtre d'observation. Conserver ces sociétaires dans l'étude implique que leur score soit remis à zéro à la première année d'étude, et que tous ces individus doivent être considérés comme s'ils entraient dans le portefeuille à la date pivot, ignorant en partie les évènements de leurs contrats survenus dans le passé. Cependant, certains de ces sociétaires sont encore présents, il semblerait anormal qu'une partie du portefeuille soit sans score quand

une autre est évaluée chaque année. D'ailleurs, il n'est pas impossible que l'ancienneté d'adhésion, qui est une variable connue, soit déterminante dans la prise de décision en addition du score, différenciant ainsi les nouveaux arrivants des anciens.

Pour résumer, tout sociétaire dont le score annuel peut être connu au moins une fois entre 2003 et 2019 entre dans le périmètre et fait partie de la base d'étude.

3.3 Récupération du coût des sinistres

Il existe plusieurs manières de recenser les sinistres. La date d'ouverture en gestion peut être utilisée ou bien la date de survenance du sinistre. Il arrive parfois que l'année de ces deux dates diffère. En effet, la gestion d'un sinistre ne démarre pas instantanément une fois survenu. Le plus souvent l'ouverture du dossier demande quelques jours ce qui justifie cette différence si le sinistre a eu lieu en fin d'année. De plus, certains sinistres ne sont déclarés que bien plus tard, ce sont des sinistres tardifs. Il existe des cas rares où quelques années séparent la survenance et la déclaration, et donc l'ouverture d'un sinistre.

Tous les coûts sont pris en compte à l'année de survenance du sinistre quelle que soit la ou les dates de paiements ou d'évaluations. Cela rend la survenance d'un sinistre et son impact sur le score plus facilement identifiable en regardant uniquement le score d'un sociétaire. De plus, l'attribution d'un score à un sociétaire les années suivant sa résiliation est évitée.

Lors de l'ouverture d'un sinistre en gestion, une valeur lui est attribuée. Cette valeur correspond au coût estimé du sinistre par l'assureur. Elle peut évoluer avec le temps à la hausse comme à la baisse selon les nouvelles informations communiquées à l'assureur. À la clôture du sinistre, la valeur est égale à la somme versée par l'assureur. Cependant, certains coûts peuvent être à la charge d'une autre société d'assurance par voie de recours.

La présence d'un recours indique également la non-responsabilité du sociétaire dans le sinistre. Il serait donc inapproprié de lui imputer la charge d'un sinistre dans son score pour lequel il n'est pas responsable. Dans ce cas, le coût réel pour l'assureur est égal à la valeur du sinistre nette de recours.

Le score se base sur des montant réels de sinistres quelle que soit l'année de score étudiée. Seulement, attendre la clôture du dossier sinistre pour récupérer le coût réel n'est pas concevable puisque certains dossiers prennent plusieurs mois et même plusieurs années à être traités. Pour cette raison, la dernière valeur évaluée du sinistre nette de recours est systématiquement utilisée pour le score comme une appréciation acceptable et temporaire du coût réel final.

Le coût d'un sinistre varie d'une situation à l'autre et la cotisation existe pour amortir ce coût. Malheureusement pour l'assureur, il arrive souvent que la valeur du sinistre nette de recours soit supérieure à la cotisation annuelle correspondante. Par la suite, ces sinistres seront appelés des sinistres coûteux.

3.4 Exploitation des données de cotisations

3.4.1 Recensement des cotisations

Tout comme il existe deux manières de recenser les sinistres, il en existe deux pour recenser les cotisations. Elles peuvent être calculées en valeur acquise ou en valeur émise pour une période donnée. La cotisation émise correspond à l'appel de cotisation annuelle à l'échéance de contrat tandis que la cotisation acquise est la part de la cotisation annuelle dont la période de risque correspond à l'année civile en cours.

Exemple : Si un sociétaire paie une cotisation de 100€ en 2014 pour un contrat d'assurance valable entre le 1er avril 2014 et le 31 mars 2015 alors les 100€ sont émis en 2014 mais il n'y a que 75€ qui sont acquis durant l'exercice 2014. Les 25€ restant sont définis comme une cotisation émise non-acquise pour cette année-là.

Le score recherché est un solde des montants réels de cotisations et de sinistres. Dans un souci de cohérence, chaque sinistre recensé pour une année doit être mis en face de la cotisation correspondante. Les cotisations prises en compte pour le score sont donc les cotisations acquises, par année d'exercice.

Le score est une valeur constatée, l'année associée n'est qu'un repère temporel. Le poids de chaque cotisation est identique quelle que soit l'année observée, il en est de même pour les sinistres. Aucune capitalisation basée sur l'inflation ou tout autre indice n'est effectuée dans cette étude.

3.4.2 Chronologie et évolution des cotisations

Durant toute la vie d'un contrat, il est possible pour le souscripteur de demander un avenant, c'est-à-dire une modification des conditions particulières d'un contrat. Ces modifications peuvent être le signalement d'un changement d'adresse ou de véhicule ou alors un changement de formule au sein du contrat. En fonction de la nature de l'avenant, la cotisation associée au contrat modifié peut augmenter ou diminuer. Si le sociétaire change sa formule pour une autre plus complète en termes de garanties alors la cotisation augmente, et inversement. Un avenant peut être demandé à n'importe quel moment de l'année, ainsi plus vite il prend effet, plus vite son impact sur la cotisation sera visible.

La souscription d'un nouveau contrat par un sociétaire se nomme une «Affaire nouvelle». Il est fréquent qu'un sociétaire détienne plusieurs contrats dans une même branche mais tous ne sont pas souscrits à la même date. Une affaire nouvelle provoque donc très souvent une hausse des cotisations totales perçues dès lors que la somme des cotisations de l'ensemble des autres contrats ne diminue pas ou peu.

En assurance auto, il existe un coefficient de réduction / majoration (CRM) plus communément connu sous le nom de bonus-malus. Il affecte directement la cotisation payée par le souscripteur. Chaque année sans sinistre, et jusqu'une certaine limite, ce coefficient baisse provoquant une baisse de la cotisation demandée. Au contraire, à chaque sinistre où l'assuré est au moins en partie responsable, le coefficient monte. Sous ce principe, la survenance d'un sinistre justifie une hausse des futures cotisations.

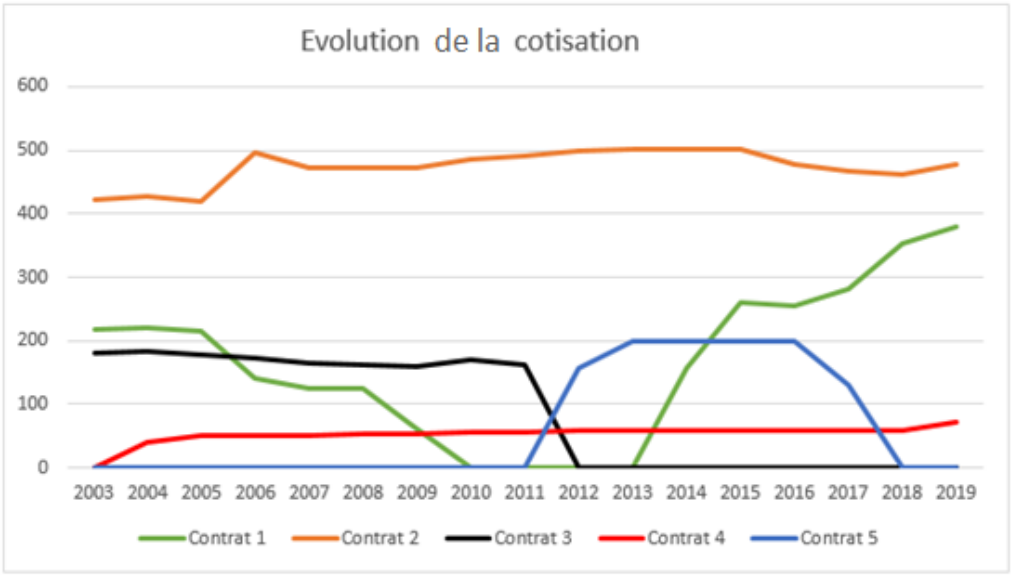

FIGURE 3.1 – Évolution de la cotisation par contrat pour un exemple de sociétaire

Dans cet exemple, il est évident que la cotisation de chaque contrat n'est pas constante même en excluant les années de souscription et de résiliation. Ces années sont particulières puisque ces deux évènements n'interviennent pas nécessairement en début ou en fin d'année. La cotisation associée est donc plus faible car la durée d'exposition au risque l'est aussi. Dans l'ensemble la baisse des cotisations, voire même l'absence de celles-ci, pour certains contrats est ici amortie par la hausse des cotisations des autres mais aussi par la souscription de nouveaux contrats.

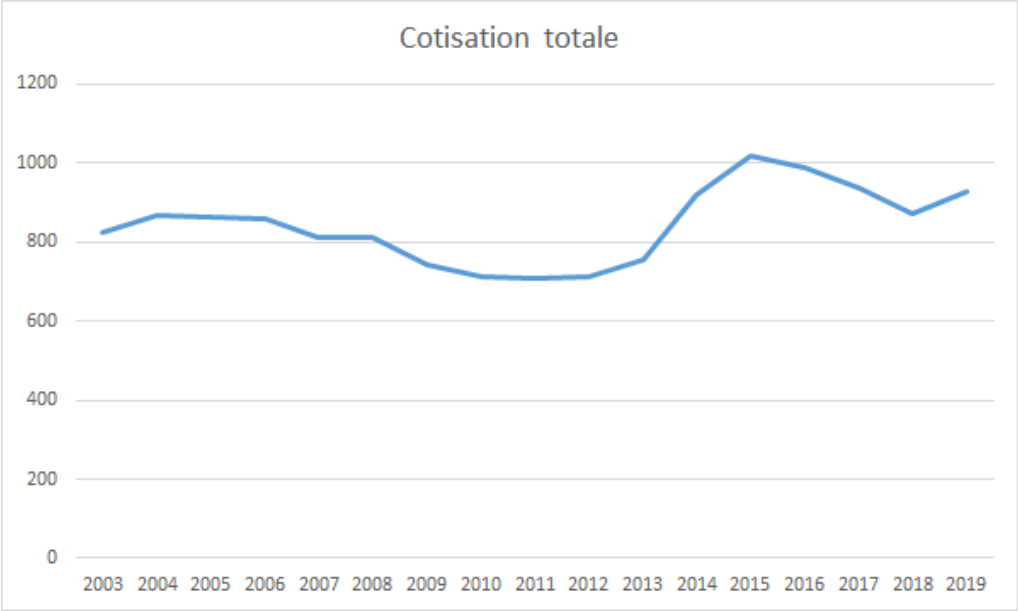

FIGURE 3.2 – Cotisation totale du sociétaire en exemple

L'évolution de la cotisation d'un sociétaire dans le temps est donc loin d'être linéaire.

FIGURE 3.3 – Frise chronologique des évènements des contrats du sociétaire en exemple

Chapitre 4

Calcul et étude du score

4.1 Calcul du score

À partir de la base d'étude de référence maintenant disponible, de multiples sous-bases peuvent être définies en fonction des sous-ensembles de branches et de la période qui peuvent être étudiés. Une fois cette base créée, le véritable score, appelé score cumulé, de chaque sociétaire peut être calculé. Pour cela et pour chaque sociétaire à chaque année de présence dans le portefeuille, la somme des scores décumulés des branches sélectionnées renvoie le score décumulé total de l'année observée. Le score cumulé d'une année, qu'il soit total ou lié à une unique branche, est calculé en sommant le score décumulé correspondant de l'année observée et les scores décumulés de même nature des années précédentes.

Pour faciliter l'étude, les scores cumulés totaux de tous les sociétaires peuvent être regroupés dans différentes bases moins imposantes. En effet, la base d'étude la plus complète à ce jour contient près de 30 millions d'observations. Regrouper les scores par date ou par temps de présence dans l'étude facilite aussi la comparaison et l'analyse.

4.1.1 Base des scores par année

Une première façon de regrouper les scores cumulés des sociétaires est de créer une base pour chaque année civile observée. L'un des avantages est que cela permet d'étudier des scores résultant des mêmes paramètres de tarification. Le second avantage majeur est que regrouper les scores de cette manière donne une vision plus réelle du portefeuille à chaque moment étudié. Cependant, des sociétaires avec des anciennetés différentes seraient alors comparés. Il ne serait pas rare de rencontrer deux scores totalement différents si l'un des sociétaires à un score cumulé sur cinq ans et si l'autre sociétaire vient d'entrer dans le portefeuille. Pour remédier à ce problème, les scores cumulés des sociétaires avec le même nombre d'années de scores étudiés peuvent être regroupés.

4.1.2 Base des scores par maturité

La seconde façon de regrouper les scores cumulés des sociétaires est de créer une base pour chaque maturité, c'est-à-dire que chaque base réunirait les scores ayant été construits sur un nombre d'années identique ce qui permet des comparaisons plus appropriées qu'avec l'autre méthode. Le premier score de chaque sociétaire étudié serait dans une même base, la somme des deux premiers scores décumulés de chaque sociétaire n'ayant pas résilié leurs contrats l'année précédente serait dans une deuxième base, et ainsi de suite. Un nombre de bases inférieur ou égal au nombre d'années étudiées est alors obtenu. La condition pour que le nombre de maturité soit parfaitement égal au nombre d'année civile observée est qu'il existe au moins un sociétaire étudié présent sur toute la fenêtre d'observation. Le grand avantage de cette méthode est qu'il est possible de suivre le comportement des sociétaires et l'évolution de leur score à travers les années de présence dans le portefeuille. À l'inverse de l'autre méthode, l'aspect réel et égalitaire de la vision par année civile est perdu car les contrats disponibles au sociétaire ne sont pas obligatoirement les mêmes selon l'année d'entrée dans la base d'étude.

4.2 Étude du score

Avant d'utiliser le score qui vient d'être calculé, il faut l'étudier. Analyser le score est essentiel pour déterminer les informations qu'il apporte, l'interprétation possible de ces informations et leurs utilités pour l'assureur. Il est important de bien comprendre ce nouvel outil et d'explorer son potentiel.

L'étude commence par la répartition du score sous différentes formes, puis par l'extraction d'indicateurs requis à la compréhension de l'évolution de celui-ci. Des exemples concrets serviront à illustrer et à soutenir la bonne capacité explicative du score et de ses indicateurs. Pour terminer ce chapitre, deux cas d'usages seront étudiés et attesteront du besoin de cet outil.

4.2.1 Distribution des scores

Pour certaines branches, les scores dépendent majoritairement de la sinistralité là où dans d'autres branches c'est la cotisation qui va faire varier les scores. C'est visible avec la branche FC où les contrats sont forfaitaires, les cotisations d'un individu sont extrêmement proches de celle de son voisin. Bien que moins fréquents, les sinistres ont plus de chance d'être couteux sur ce produit car les cotisations sont faibles. Tous ces éléments expliquent la répartition des scores.

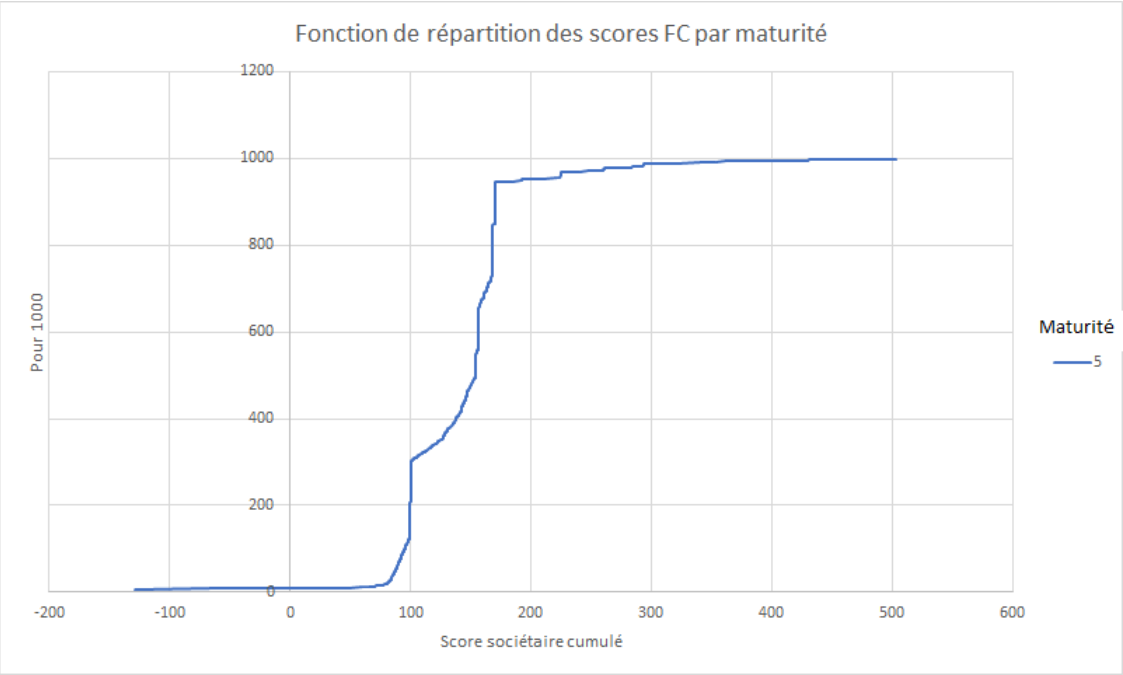

FIGURE 4.1 – Fonction de répartition des score cumulé sur 5 ans pour la branche FC

Les sauts mettent en évidence les différentes formules proposées par l'assureur. Le plus évident est celui quand le score vaut 100 puisqu'une grande partie des sociétaires FC détient un contrat pour lequel la cotisation annuelle est proche de 20€. Avec cinq ans de maturité sans sinistre dans cette branche, la valeur de 100 est retrouvée pour le score cumulé. Les scores situés entre les sauts sont ceux de sociétaires sinistrés ou de sociétaires qui ont changé de formule à un moment.

Tout comme il existe deux manières de réunir les scores, il existe deux façons de les étudier. Une qui utilise les bases par année et une qui utilise les bases par maturité.

Pour la première méthode, il faut fixer une date d'observation et étudier le score cumulé pour tous les sociétaires à cette date. Lorsque la distribution de ces scores est observée, une rupture entre les scores négatifs et positifs est notable. Ceci témoigne du fait que tous les sociétaires n'ont pas cumulé le même nombre d'année de score. Dans la partie positive, les gros scores appartiennent en grande majorité aux sociétaires présents depuis le début de l'étude. Les scores négatifs sont moins présents, les sinistres sont souvent compensés en quelques années par les cotisations de l'ensemble des contrats détenus par le sociétaire associé, à l'exception des sinistres très coûteux mais bien plus rares qui demandent un plus grand nombre d'années pour être amortis ou, dans les cas les plus extrêmes, pour les sinistres dont le coût est bien trop élevé pour être équilibré par les seules cotisations du sociétaire.

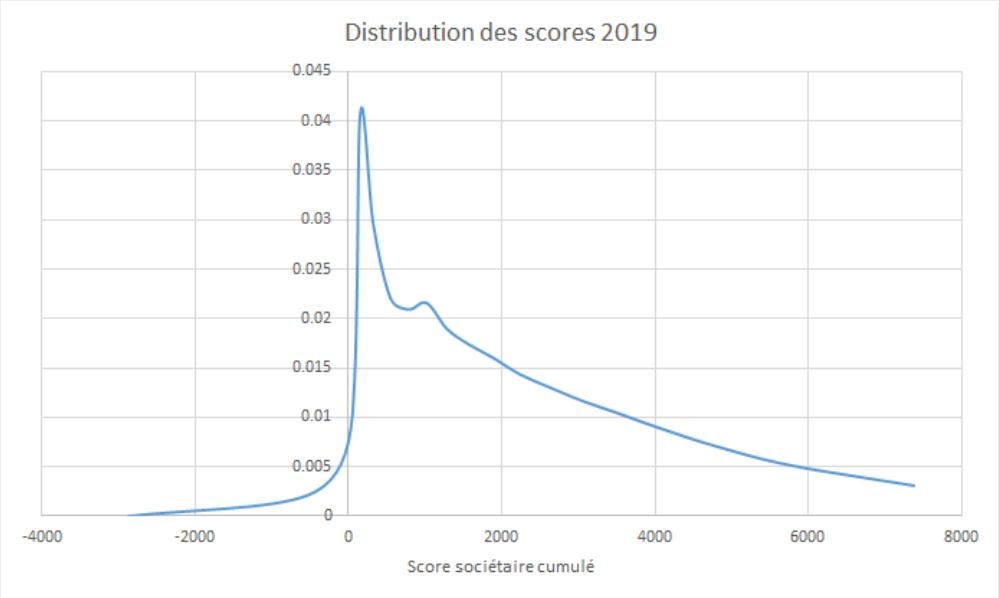

FIGURE 4.2 – Courbe de distribution des scores cumulés vus fin 2019 sur l'ensemble des branches [1]

1. Pour une meilleur visibilité, la distribution n'est pas représentée entièrement. La valeur maximale du score atteinte ici est supérieure à 2,3 millions.

Pour la seconde méthode, il faut fixer une maturité et étudier tous les scores cumulés disponibles à la maturité choisie. La courbe de distribution correspondante est ensuite construite. Si l'opération est répétée pour plusieurs maturités alors le graphique suivant est obtenu :

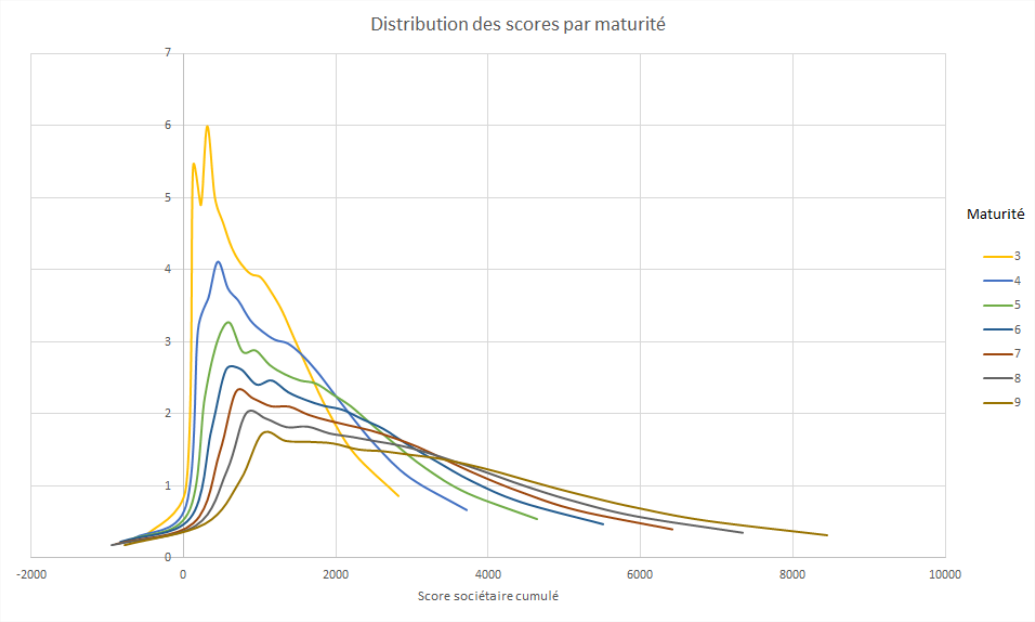

FIGURE 4.3 – Évolution par maturité de la distribution des scores cumulé sur l'ensemble des branches

Chaque courbe correspond à la distribution du score d'une maturité donnée. Dans un souci de lisibilité, les maturités 1 et 2 ne sont pas représentées ici car leurs courbes écrasent les autres du fait d'une grande concentration de scores légèrement supérieurs à zéro. Dans le même esprit, chaque distribution n'est pas entièrement tracée puisque l'étendue des scores peut être très grande, notamment lorsque les dernières maturités sont observées.

Deux comportements sont présents sur la distribution. Tout d'abord, la courbe à tendance à se décaler vers la droite lorsque la maturité augmente. L'explication se trouve dans la construction du score lui-même, des valeurs positives pour la plupart sont sommées. Seul un sinistre coûteux peut faire décroitre le score. Or, si tous les sociétaires paient, en principe, une cotisation chaque année, tous les sociétaires ne sont pas régulièrement victimes d'un sinistre coûteux de manière suffisante pour égaliser la somme de ses cotisations. Il est rassurant de voir que les scores ont tendance à être croissants avec la maturité.

Ensuite, la courbe à tendance à s'aplatir et se lisser lorsque la maturité augmente. La hausse de la variance s'explique par l'étendue des scores qui devient de plus en plus grande. D'un côté, les très bons risques ont un score de plus en plus élevé, le score maximal croît donc. De l'autre, plus la période d'observation est grande, plus la probabilité de survenance d'un sinistre l'est aussi, les scores deviennent alors plus dispersés.

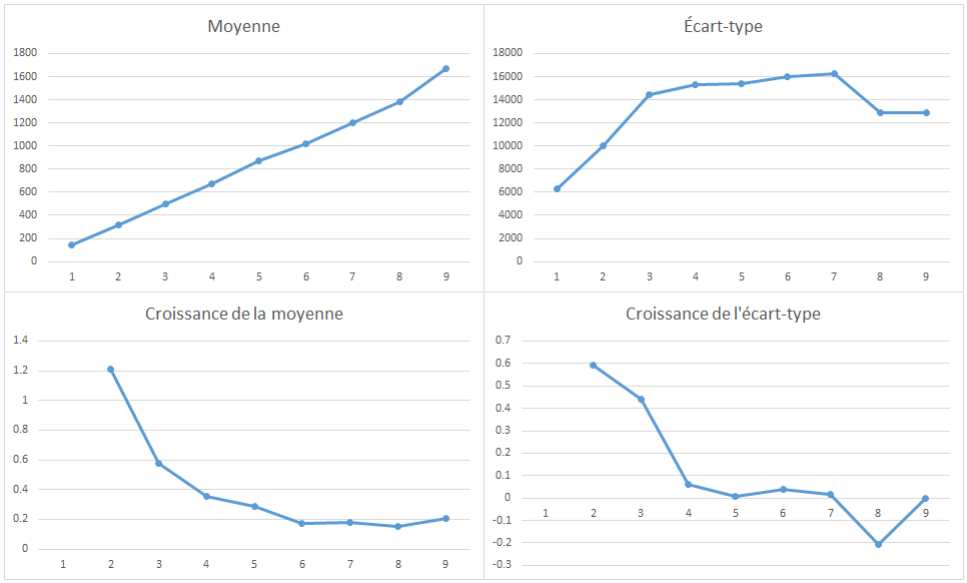

FIGURE 4.4 – Évolution de la moyenne et de l'écart-type du score cumulé en fonction de la maturité

La moyenne a l'air de se comporter de façon linéaire avec les années, la croissance se stabilise. L'écart-type a l'air de se comporter de façon plus logarithmique au départ avant de tendre vers une valeur constante comme en témoigne sa croissance qui semble tendre vers 0.

Pour rappel, ces résultats ne prennent pas en compte les frais qui auront pour conséquences de diminuer la moyenne des scores et de décaler l'ensemble des courbes de distribution vers la gauche. Une fois intégrés, cette analyse permettra d'observer à quelle maturité le pic se situera dans le positif. Cela signifiera qu'une grande partie des frais généraux aura été amortie.

4.2.2 Mise en place d'indicateurs pour interpréter l'évolution du score

Un des obstacles de l'étude de ces scores est le volume massif de données disponibles. L'objectif est de traiter des dizaines de millions de lignes dans une durée raisonnable. C'est la raison pour laquelle l'historique des cotisations ou des scores cumulés et décumulés d'un sociétaire n'est pas rappelé à chaque vue de la base. Avoir une base plus concise au profit d'un temps de traitement plus court est un choix. En revanche, l'évolution du score ne peut être interprétée qu'avec le dernier score connu.

À partir de l'ensemble des scores cumulés d'un sociétaire, différents indicateurs sont créés pour suivre le comportement du score dans le temps. Ces indicateurs sont :

— TjrPos : Un booléen qui renvoie 1 si le score a toujours été positif, et 0 si le score a été négatif au moins une fois par le passé. Par défaut à 1, si la première année le score est négatif, l'indicateur passe directement à 0.

— TjrCroiss : Un booléen qui renvoie 1 si le score a toujours été croissant, et 0 si le score a été strictement décroissant au moins une fois d'une année sur l'autre. Par défaut à 1, si la première année le score est négatif, l'indicateur passe directement à 0.

— Nb_An_negatif : Un compteur qui indique le nombre d'années où le score est strictement inférieur à 0.

— Chgt_Signe : Un compteur qui indique le nombre de fois où le score a changé de signe. Par défaut à 0, si la première année le score est négatif, l'indicateur passe directement à 1.

Ils servent à résumer en quelques variables l'ensemble des informations passées.

Année	Maturité	Score	TjrPos	TjrCroiss	NB_An_negatif	Chgt_Signe
2003	1	173,36	1	1	0	0
2004	2	346,72	1	1	0	0
2005	3	528,60	1	1	0	0
2006	4	718,60	1	1	0	0
2007	5	985,06	1	1	0	0
2008	6	1309,44	1	1	0	0
2009	7	-539,76	0	0	1	1
2010	8	-193,09	0	0	2	1
2011	9	183,67	0	0	2	2
2012	10	595,85	0	0	2	2
2013	11	1030,27	0	0	2	2
2014	12	1476,97	0	0	2	2
2015	13	1923,67	0	0	2	2
2016	14	2393,55	0	0	2	2
2017	15	2884,61	0	0	2	2
2018	16	3375,67	0	0	2	2
2019	17	3925,43	0	0	2	2

TABLE 4.1 – Situation du score d'un sociétaire par année

Ci-contre se trouve l'évolution du score d'un sociétaire sur la branche MGAR. Arrivé dans l'étude en 2003, son score est toujours croissant jusqu'en 2008, ce qui indique une absence de sinistre coûteux sur cette période. Après cela, son score passe dans le négatif en 2009 avant de revenir dans le positif en 2011. Entre 2011 et 2019 le score de ce sociétaire est resté positif sur la branche MGAR. À partir de cette analyse il est déduit que le sociétaire n'est pas un mauvais risque en ce qui concerne cette branche. Les indicateurs sont en partie utiles pour faciliter cette analyse.

Si seul le dernier état du score est regardé, une première idée sur le sociétaire peut tout de même être faite.

Année	Maturité	Score	TjrPos	TjrCroiss	NB_An_negatif	Chgt_Signe
2019	17	3925,43	0	0	2	2

TABLE 4.2 – Dernière situation connue du score du sociétaire

Avec uniquement cette dernière ligne, il est possible de déduire que le sociétaire est présent dans l'étude depuis 17 ans, c'est-à-dire depuis 2003, pendant lesquels son score est passé dans le négatif une seule fois pour une période de deux ans avant de revenir dans le positif. En conclusion et avec uniquement ces informations à disposition, il est fort probable que le sociétaire ne soit pas un mauvais risque.

Contrairement à la table exhaustive, les évènements impactant le score ne sont pas datés. De plus, le score minimal atteint par ce sociétaire n'est pas connu. Dans ce cas, la perte d'information est relativement faible puisque les deux conclusions sont identiques. Utiliser une table restreinte pour étudier un sociétaire en particulier et se référer à la table détaillée en cas de besoin est donc possible.

Ces indicateurs sont tout aussi utiles d'un point de vue plus macroscopique, ils permettent alors de regarder une tendance sur l'ensemble des sociétaires pour une ou plusieurs branches.

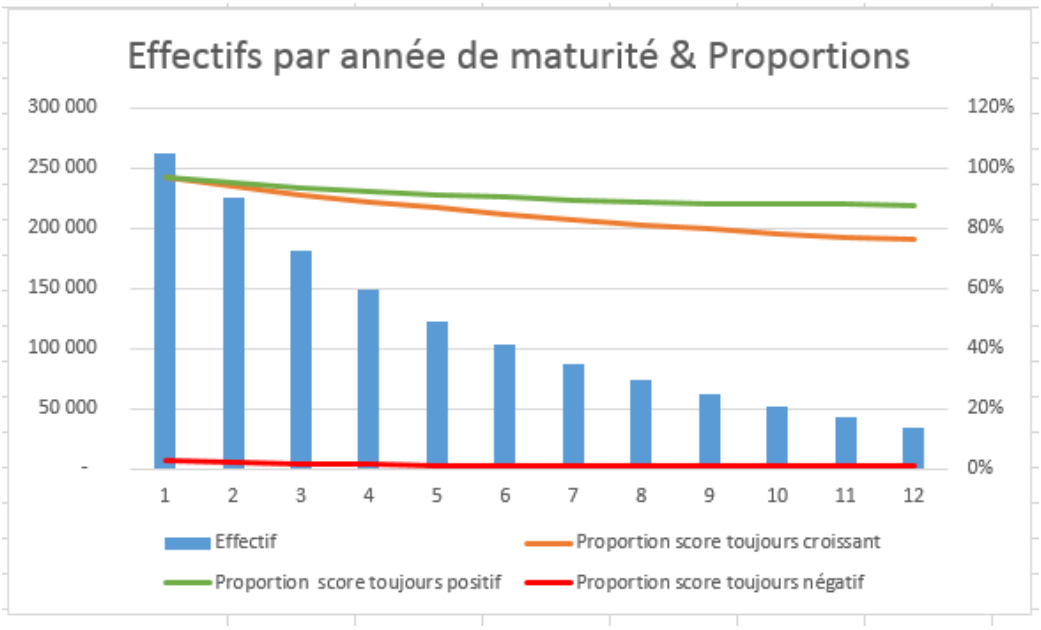

FIGURE 4.5 – Vision macroscopique des indicateurs pour la branche MLCO

Dans ce premier graphique la branche MLCO est uniquement étudiée. Pour chaque sociétaire chacun des scores existants entre 2008 et 2019 dans cette branche est utilisé. En barre, il y a le nombre de sociétaires étudiés à chaque maturité et en courbe, il y a trois informations sur leurs scores, construits à partir des indicateurs définis précédemment.

La courbe verte représente les scores toujours positifs en proportion du nombre de sociétaires étudiés. Ce sont les bons risques, ils leur arrivent d'être impactés par un sinistre couteux mais leurs cotisations passées sont suffisantes pour avoir un score au-dessus de zéro tout leur temps de présence. Malgré douze ans dans le portefeuille, 87% des sociétaires appartiennent à cette catégorie.

La courbe orange représente les scores toujours croissants d'une année sur l'autre en proportion du nombre de sociétaires étudiés. Ce sont les très bons risques, chaque année leur cotisation est supérieure à leur coût de sinistre. Ceci est dû à une fréquence sinistre faible et à un coût par sinistre faible également. Dans l'exemple ci-dessus, 75% des sociétaires restent de très bons risques après douze ans. Cette catégorie est incluse dans la catégorie des bons risques.

La courbe rouge représente les scores toujours négatifs depuis leur arrivée dans la base d'étude en proportion du nombre de sociétaires étudiés. Ce sont les très mauvais risques, leur sinistralité élevée en coût ou en fréquence les empêchent d'être rapidement dans le positif. S'ils représentent 3% de l'effectif la première année de cotisation en MLCO et 1% après douze ans c'est que ces sociétaires sont retombés dans le positif entre temps, ou alors qu'ils n'appartiennent plus à l'étude pour cause de résiliation. Le plus préoccupant ne sont pas les 3% qui peuvent s'expliquer par une première année malchanceuse, mais par le pourcentage de sociétaires toujours négatifs au bout d'une certaine durée. Les garder dans le portefeuille en attendant qu'ils aient un score positif est risqué, rien n'empêche d'autres sinistres de survenir entre temps. Cependant si une branche est étudiée à la fois, prendre la décision de résilier un sociétaire identifié comme un mauvais risque n'est pas sans faille. En effet, sur l'ensemble de ses contrats, un mauvais risque MLCO peut finalement être qualifié comme bon.

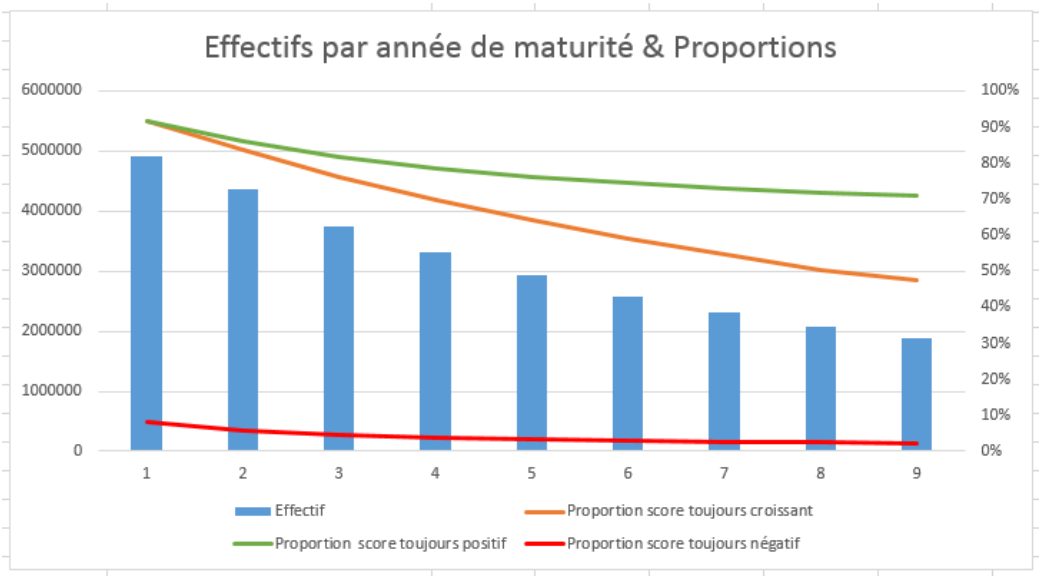

FIGURE 4.6 – Vision macroscopique des indicateurs pour l'ensemble des branches

Cette fois, l'ensemble des contrats des différentes branches étudiées est pris en compte dans le calcul des indicateurs. Pour réaliser ce graphique, tous les scores de chaque sociétaire présent entre 2011 et 2019 ont été utilisés. Au bout de neuf années de scores, les très bons représentent 47% de l'effectif concerné et les bons risques 71%. Les mauvais risques, eux, correspondent à 2% des sociétaires présents depuis neuf ans. Plus le nombre de contrats engagés est élevé, plus le risque de survenance d'un sinistre l'est aussi, ce qui renforce la différence entre les simplement bons et les très bons.

4.2.3 Exemples de scores

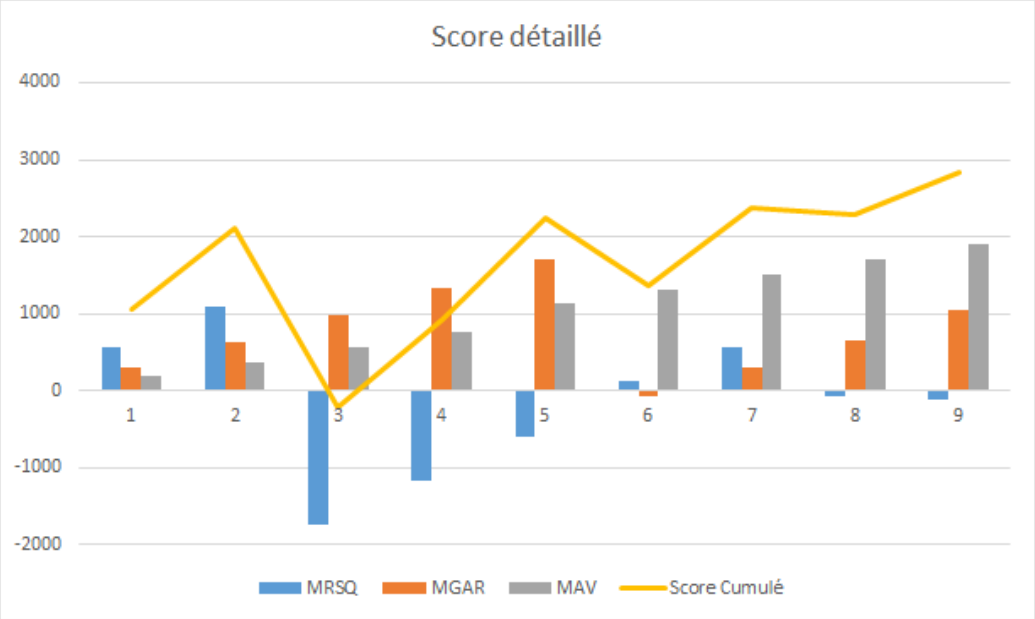

FIGURE 4.7 – Évolution du score cumulé par branche d'un mauvais risque MRSQ

Dans cet exemple, le score MRSQ du sociétaire n'est pas bon quand il est pris seul. En revanche, lorsque la somme du score MRSQ avec les scores MGAR et MAV est calculée, le sociétaire n'est plus à considérer comme un mauvais risque.

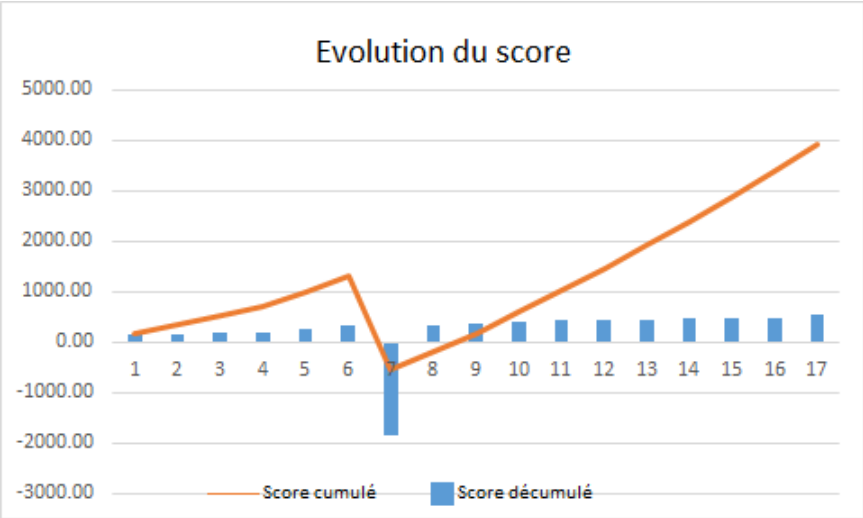

FIGURE 4.8 – Évolution du score cumulé et décumulé d'un bon risque

Avec un score décumulé croissant à l'exception de la septième année d'étude, ce sociétaire est l'exemple typique d'un bon risque. Bien que très néfaste, la sinistralité a été compensée à la fois par un bon score préalablement acquis et des cotisations élevées par la suite. Le score cumulé n'est resté dans le négatif que deux ans et n'est jamais redescendue sous la barre symbolique du zéro après ça. Dans l'éventualité d'une survenance de sinistre de même envergure, le score de ce sociétaire resterait positif.

Les sociétaires avec les meilleurs scores sont composés de beaucoup de retraités. Ces sociétaires sont plus âgés et sont présents dans le portefeuille depuis plusieurs dizaines d'années. Les trois quarts d'entre eux détiennent quatre ou six contrats.

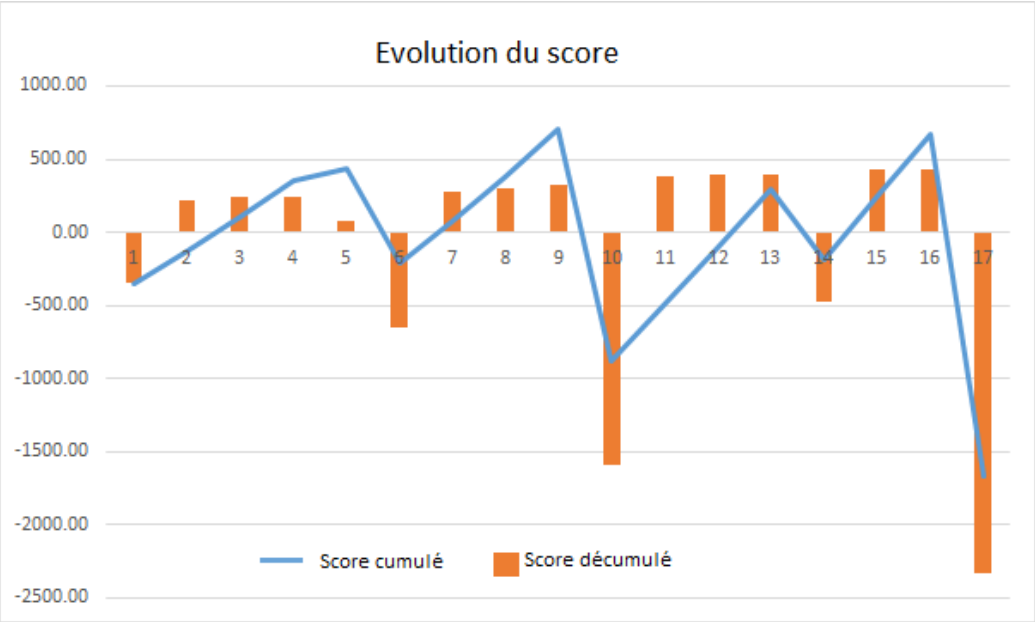

FIGURE 4.9 – Évolution du score cumulé et décumulé d'un mauvais risque

Avec une sinistralité coûteuse tous les trois ans en moyenne, ce sociétaire détient un score cumulé en dents de scie. Il lui faut parfois plusieurs années pour redevenir rentable avant de revenir une nouvelle fois dans le négatif. Si le coût de ses sinistres n'est pas extrême, leur fréquence est très inquiétante pour l'assureur. Il est difficile de prévoir si, à l'avenir, le sociétaire peut maintenir un score positif mais le comportement de ce score laisse supposer que non. Cet individu peut être qualifié de mauvais risque au titre de la branche MGAR. Les indicateurs sont une nouvelle fois une bonne façon de résumer la situation.

Année	Maturité	Score	TjrPos	TjrCroiss	NB_An_negatif	Chgt_Signe
2019	17	-1665,11	0	0	8	9

TABLE 4.3 – Dernière situation connue du score du mauvais risque

Sur les 17 années étudiées, le score a changé neuf fois de signe, il est donc passé cinq fois dans le négatif. De plus, le score était inférieur à zéro près de la moitié du temps.

Les sociétaires avec les pires scores sont pour la plupart des employés. Typiquement, ils sont présents dans le portefeuille depuis moins de 15 ans et sont plus jeunes que les autres. Enfin, la majorité de ces sociétaires détiennent un seul contrat, un quart en possède trois et un cinquième quatre.

4.2.4 Utilisations des scores

Bien qu'il apporte énormément d'informations intéressantes sur les sociétaires et la vie passée de leurs contrats, le score est conçu pour répondre à des situations concrètes sur le temps présent. À l'aide de ces informations et à partir du score mis en place, plusieurs applications sont possibles.

La défense de portefeuille est une situation dans laquelle l'assureur est prêt à effectuer un geste commercial envers un sociétaire qui risque de résilier son ou ses contrats, dans le but de le conserver dans le portefeuille. Cependant tous les sociétaires ne sont pas éligibles. Défendre son portefeuille à un coût et il est important de savoir quels sociétaires valent le coup d'être retenus et lesquels non. Pour choisir les sociétaires éligibles avant que la situation ne l'impose, deux méthodes ont été étudiés sur la branche MRSQ selon deux schémas différents. Les scores cumulés utilisés ont été calculés sur la fenêtre d'observation 2003 2019. Les deux méthodes présentées ici sont également applicables sur toutes les fenêtres d'observations, branches et combinaisons de branches.

Schéma 1 :

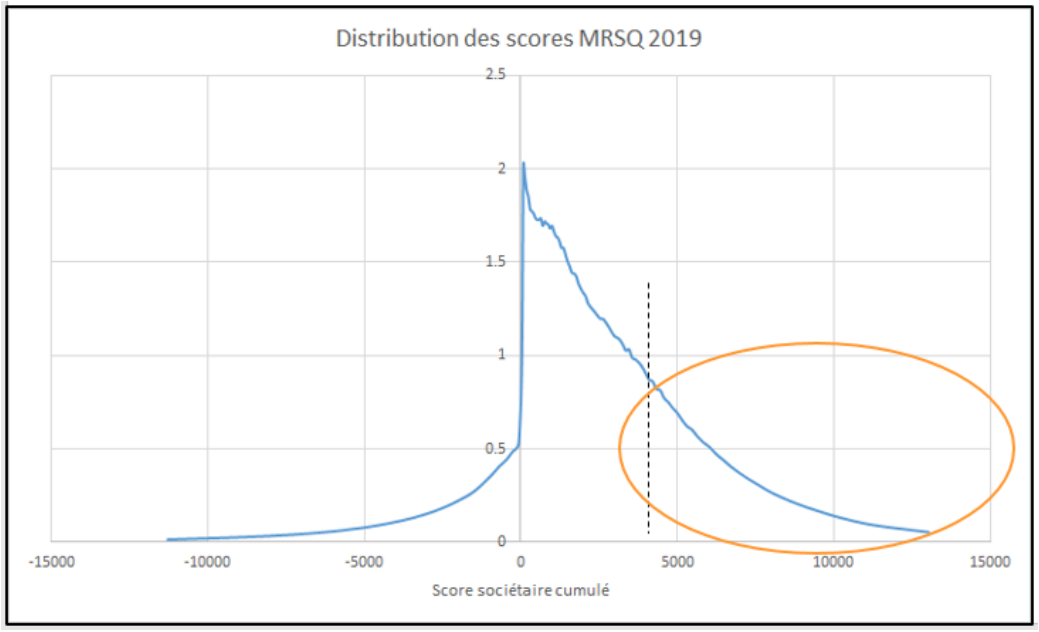

FIGURE 4.10 – Identifications des meilleurs scores de 2019

Avec cette méthode, la base des scores vus à fin 2019 est utilisée et les sociétaires dont le score est supérieur au troisième quartile sont sélectionnés. Le choix du quantile est arbitraire et dépend des moyens alloués à la défense du portefeuille. Cette approche favorise les sociétaires avec le plus d'ancienneté. Par conséquent une population plus âgée que celle de départ et une proportion plus importante de retraités et moins importante d'employés d'entreprises privés est retrouvée. Cela s'explique simplement du fait qu'il est plus facile d'obtenir un meilleur score cumulé que son voisin lorsque son nombre d'année de présence est plus grand. C'est pour ça que 68% des sociétaires retenus ont la maturité maximale possible à ce jour contre 35% sur la totalité de la base et que moins de 4% ont une maturité strictement inférieure à huit ans. Il faut noter qu'un lien existe entre la maturité du score et l'âge d'un sociétaire. En effet, la maturité atteinte par les sociétaires de moins de 25 ans est inférieure à huit ans si

l'âge minimal pour souscrire un contrat MRSQ est de 18 ans. De plus, les jeunes conducteurs sont réputés pour être des individus plus à risques que les conducteurs expérimentés.

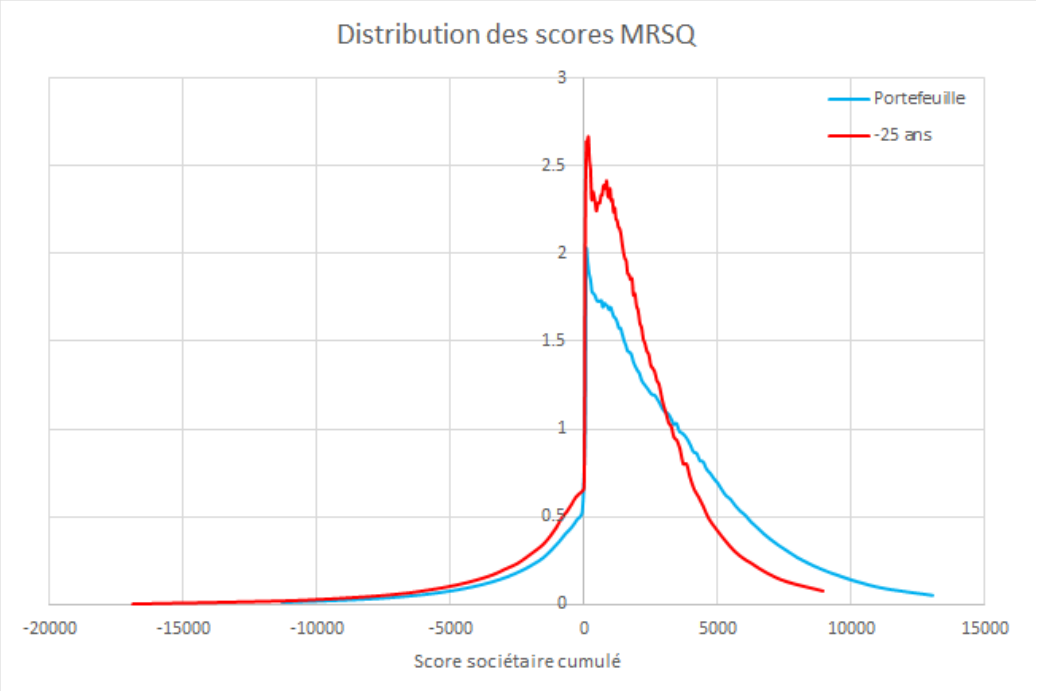

FIGURE 4.11 – Comparaisons de la distribution des scores entre le portefeuille et les souscripteurs de moins de 25 ans

Enfin, le nombre de contrat toutes branches confondues détenus par les sociétaires diffère. Plus d'un tiers de la sous-population possède quatre contrats ou plus, contre un cinquième pour la population totale.

Cette approche revient donc à cantonner la défense de portefeuille sur les sociétaires les plus anciens et déjà multi-équipés.

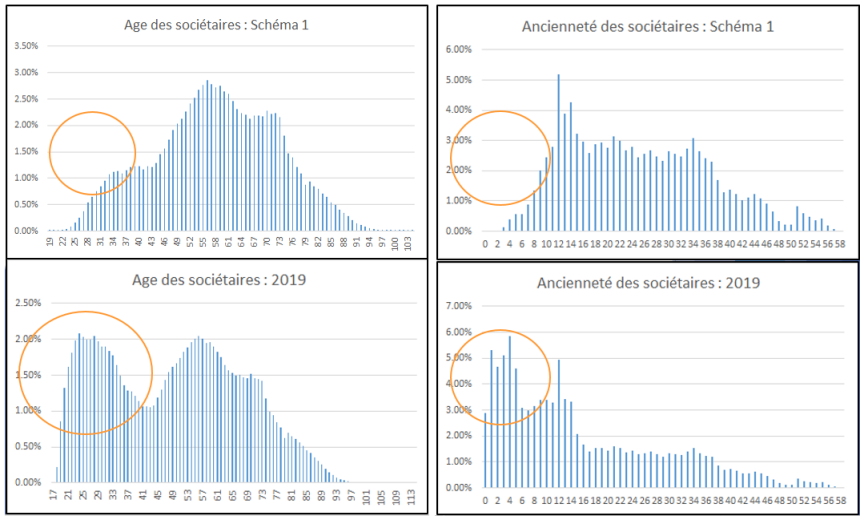

FIGURE 4.12 – Comparaisons de l'âge et de l'ancienneté entre la population sélectionnée et la population totale

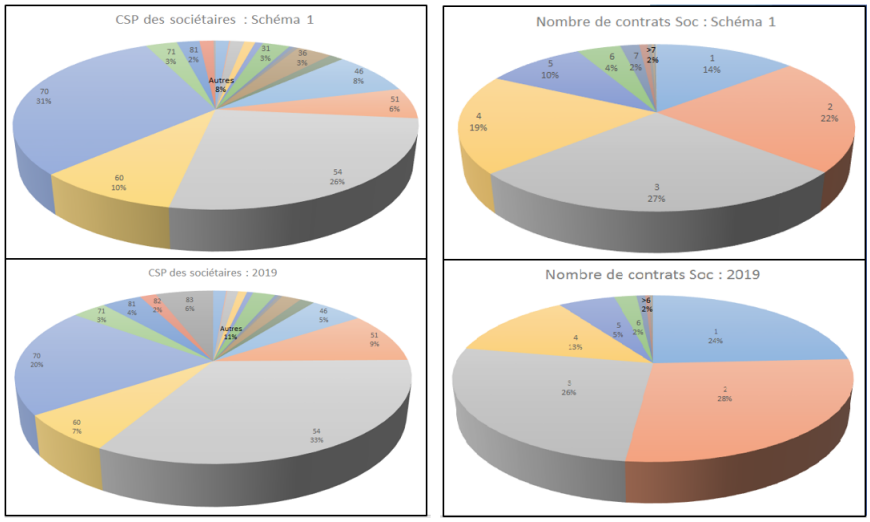

FIGURE 4.13 – Comparaisons des CSP et du nombre de contrats souscrits entre la population sélectionnée et la population totale

Schéma 2 :

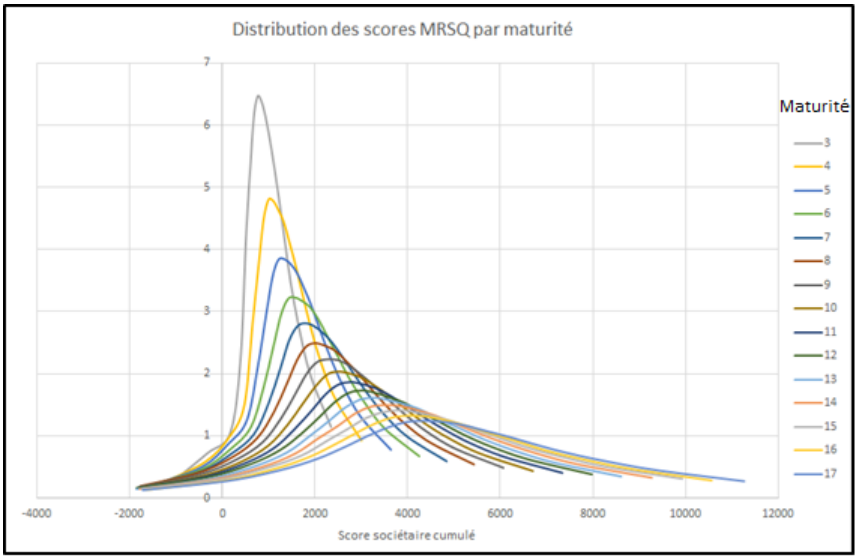

FIGURE 4.14 – Distributions par maturité des scores cumulés MRSQ

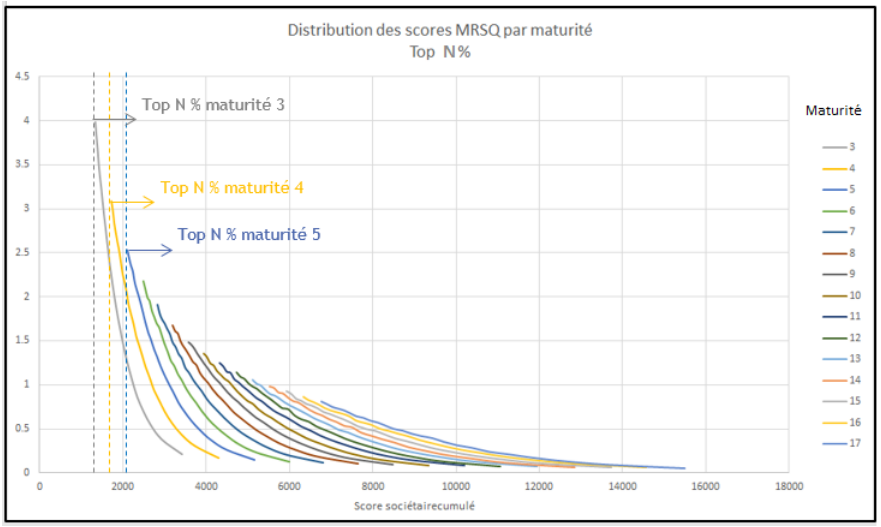

FIGURE 4.15 – Sélection des meilleurs scores cumulés MRSQ par maturité

Pour cette méthode, dans chaque base de scores par maturité les sociétaires dont le score est supérieur au quantile N sont sélectionnés. Pour supprimer les résiliés, les doublons ou les individus qui n'ont plus un bon score, ne sont retenues parmi l'ensemble des observations que celles dont l'année correspondante est la dernière année observée, ici 2019. L'objectif est de choisir N de telle sorte que les sociétaires retenus représentent un quart du portefeuille, cette dernière proportion étant choisie arbitrairement. Enfin, tous les sociétaires qui possèdent sur leur score une maturité supérieure ou égale à trois sont conservés après sélection puisque c'est à partir de ce temps que le classement semble se stabiliser.

Avec cette approche, la sous-population est plus représentative de la population du portefeuille. À l'exception des individus avec moins de trois ans d'ancienneté, la répartition de l'ancienneté est similaire, de même pour la CSP et la maturité. La distribution des âges s'est toutefois décalée légèrement vers la gauche avec une concentration plus importante des 45 – 65 ans. Mais la plus grande différence entre les bons scores et le portefeuille réside dans le nombre de contrat toutes branches confondues détenus par les sociétaires. En effet, comme avec la première méthode plus d'un tiers de la sous-population possède quatre contrats ou plus, en revanche les sociétaires mono-équipés sont plus nombreux de 5% avec cette méthode.

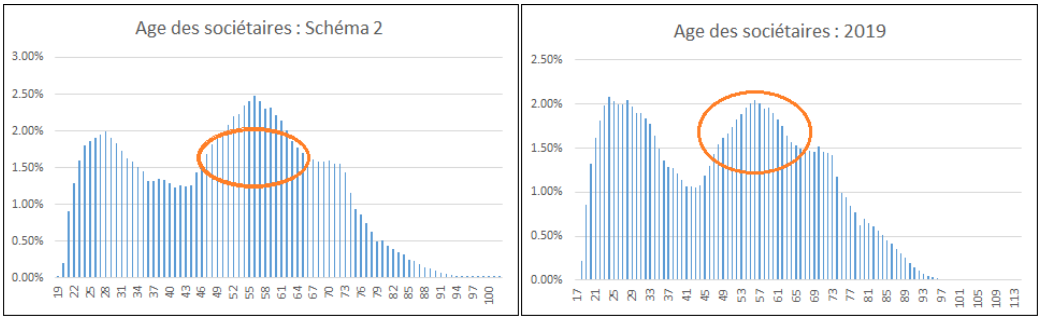

FIGURE 4.16 – Comparaison de l'âge de la population sélectionnée face à la population totale

Cette approche permet à la fois de conserver des sociétaires anciens générant beaucoup de rentabilité mais aussi de retenir des sociétaires plus récents à fort potentiel de développement.

Le score est conçu pour être polyvalent, la défense de portefeuille n'est pas le seul cas d'usage pratique. Le score est également un indicateur essentiel pour la surveillance de portefeuille.

La surveillance de portefeuille est une évaluation globale des sociétaires afin d'identifier les mauvais risques dont l'assureur veut se séparer. Pour cela, plusieurs variables sont prises en compte comme l'ancienneté ou le nombre de contrats souscrits mais l'information principale réside dans la comparaison entre le coût des derniers sinistres et les dernières cotisations perçues.

Aujourd'hui, l'historique des sinistres pris en compte par les collaborateurs lors de l'évaluation d'un sociétaire est limité, de plus ce sont les dernières cotisations de chaque contrat qui font effet. Le problème que cela pose est qu'un mauvais risque avec peu de sinistres ces dernières années et des cotisations récentes fortes est conservé dans le portefeuille alors que le score cumulé est bien négatif lorsque les cotisations sont étudiées en détail et que l'historique des sinistres est étendu.

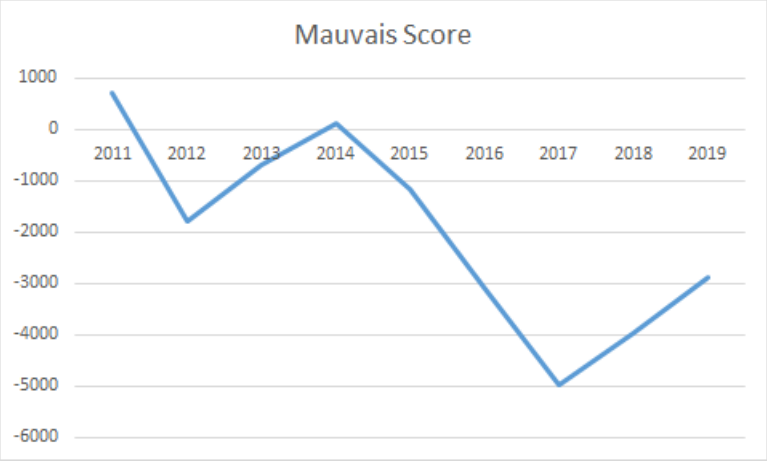

FIGURE 4.17 – Évolution du score cumulé d'un mauvais risque

Dans cet exemple particulier, la présence d'une sinistralité coûteuse est identifiable sans difficulté en 2012 puis en 2015, 2016 et 2017. Le score cumulé de ce sociétaire sur la période de 2011 à 2019 est clairement négatif. Pourtant, le sociétaire n'a pas été résilié par l'assureur. En effet, si les informations disponibles sur la fiche client aujourd'hui sont les seules informations utilisées, la somme des cotisations du sociétaire est majorée et le coût des sinistres avant 2014 inclus est inconnu. Le sociétaire est donc considéré comme rentable pour la même période alors qu'il ne l'est pas dans les faits.

C'est en parti avec ce type d'irrégularité sur le long terme que certains mauvais risques restent présents dans le portefeuille même après plusieurs années. Le score est une solution à ce problème. Il est à la fois plus précis et plus synthétique car il regroupe déjà les branches les plus importantes. Son utilisation reste accessible et adaptée à plusieurs profondeurs d'historique si besoin. Cette solution sera d'autant plus inévitable lorsque sera intégré les autres branches, la dimension foyer et surtout les frais non-considérés jusqu'alors.

Reconnaître le bon moment pour se séparer d'un mauvais risque n'est pas simple. L'analyse de la dynamique du score pourrait permettre de définir une stratégie de résiliation efficace.

Chapitre 5

Projection du score

Connaître la valeur d'un sociétaire aujourd'hui est intéressant mais n'est pas toujours suffisant dans la prise de décision. Il a été vu que certains scores oscillaient autour de zéro, il y a donc une incertitude sur les sociétaires associés. Même si leurs scores sont positifs une année, il existe un risque important que ça ne soit pas le cas l'année suivante alors que le choix de garder ces sociétaires dans le portefeuille a été pris. À l'opposé de cela il y a des sociétaires avec un bon score à qui, malheureusement, un sinistre onéreux survient. Face à ce genre de situations, il est difficile pour l'assureur de choisir le comportement à adopter. Pour faciliter la prise de décision, le score des différents sociétaires peut être projeté dans le temps. Dans ce chapitre, le risque de résiliation total est considéré comme nul.

Les modélisations et les projections sont effectuées sous Python 3.6+ à l'exception du *Grid Searching* qui est effectué sous Python 3.5.2. L'application des matrices de transition est effectuée avec SAS Entreprise Guide 7.1.

5.1 Les piliers de la modélisation

Avant de démarrer les projections à partir du score construit précédemment, il est important de définir certaines notions inévitables dans l'élaboration d'un modèle qui dispose de paramètres exogènes aux données, plus couramment appelés hyperparamètres. Ces notions sont déterminantes à la calibration des modèles employés par la suite.

5.1.1 La validation croisée

Calibrer les paramètres d'un modèle et tester ce modèle sur les mêmes données est une erreur. Avec cette méthode, un modèle considéré comme optimal pour le jeu de donnée serait incapable d'effectuer de bonnes prédictions pour des nouvelles données. C'est ce qui est appelé le surapprentissage.

Pour l'éviter, une partie des données est conservée dans une base de test en attendant une dernière évaluation du modèle, tandis que la base d'apprentissage est utilisée pour calibrer le modèle en amont. Cependant, le risque de surapprentissage existe toujours. Si le modèle ne surapprend pas de la base d'apprentissage, il y a toujours un risque d'avoir un modèle qui correspond de trop à la base de test. Ce qui revient à déplacer le problème de départ.

Une solution est d'utiliser une partie de la base d'apprentissage comme base de validation afin de réduire encore le risque de surapprentissage. En faisant cela, le nombre de données pouvant être utilisé pour calibrer les paramètres est une fois de plus réduit, rendant le modèle plus dépendant à la base d'apprentissage. C'est ici qu'intervient la méthode de validation croisée. La base de test est conservée pour l'évaluation finale mais il n'est plus nécessaire de construire une base de validation intermédiaire au départ.

La méthode consiste à diviser la base d'apprentissage en k parties de même volume, d'utiliser $k-1$ d'entre elles comme une nouvelle base d'apprentissage et la partie restante comme base de validation. L'opération est ensuite répétée de telle sorte que chaque partie soit utilisée comme base de validation. La moyenne des k indicateurs de performances est ensuite utilisée comme valeur de comparaison entre les modèles. Plus k est grand, plus le temps d'apprentissage est long mais le risque de surapprentissage est moindre. Une fois les paramètres calibrés grâce à cette méthode, le modèle peut être évalué sur la base de test.

5.1.2 Le *Grid Searching*

Une des façons de bien calibrer un modèle lorsqu'il y a plusieurs hyperparamètres est la méthode *Grid Search*. Elle consiste en un balayage de toutes les combinaisons d'hyperparamètres possibles. Afin d'éviter de tester une infinité de combinaison, l'ensemble des valeurs à considérer par chaque paramètre est défini au préalable. Pour chaque combinaison testée, une valeur d'évaluation du modèle est renvoyée.

Coupler cette méthode avec la validation croisée est possible. Une fois toutes les combinaisons évaluées, les meilleures sont recherchées pour les utiliser dans le modèle final. Comme la méthode *Grid Search* teste un grand nombre de modèles, le calibrage peut prendre du temps, surtout si la validation croisée est employée.

5.2 Projection par régression logistique

Avec cette première approche, les individus sont classés en fonction d'un certain seuil de score atteint ou non dans les années à venir. Dans une problématique de surveillance du portefeuille, les mauvais risques doivent être identifiés le plus tôt possible afin d'éviter de futurs coûts jugés trop importants pour l'assureur. Sera considéré ici un mauvais risque comme un sociétaire dont le score serait inférieur à zéro dans les prochaines années, ce qui équivaut à dire que son coût en sinistre serait supérieur aux cotisations reçues sur une période. La période d'évaluation est une combinaison de deux autres à savoir :

— La période passée, pour laquelle le score a déjà été calculé.

— La période future, pour laquelle le score est encore inconnu.

Tout d'abord un rappel sur ce que sont les modèles linéaires généralisés sera fait, puis le cas spécifique de la régression logistique sera abordé. Enfin ce modèle sera appliqué pour différentes durées de projection.

5.2.1 Rappel théorique

Les modèles linéaires généralisés sont des modèles qui en plus d'affecter à chaque variable explicative X_i un coefficient β_i comme les modèles linéaires classiques, sont caractérisés par une fonction de lien. Cette fonction de lien est une bijection qui permet de se ramener à une forme linéaire du problème :

$$g(E(Y \mid X)) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

La fonction de lien apporte une relation entre le prédicteur linéaire et la moyenne de la fonction de distribution.

Dans le cas d'une régression logistique, la loi de distribution la plus représentative est une loi de Bernoulli de paramètre q. La fonction de lien doit nécessairement être une bijection de $]0\,;\,1[$ vers R. La fonction logit : $]0;1[\rightarrow\ \mathbb{R}\,;\ q\ \mapsto\ \ln\left(\frac{q}{1-q}\right)$ est couramment utilisée dans cette situation. D'autres fonctions existent cependant comme la fonction probit ou encore la fonction log-log.

Lors de la mise en place du modèle, plusieurs paramètres de régularisation doivent être calibrés afin d'éviter le surapprentissage en pénalisant certaines variables et donc d'obtenir un meilleur modèle de prédiction. Parmi ces paramètres il y a :

— Un paramètre dictant la méthode de régularisation.

— Un paramètre, noté $\frac{1}{c}$, correspondant à la force de cette régularisation.

La première méthode de régularisation utilisable est la régularisation de Ridge. Elle consiste à minimiser les coefficients de la régression, minimisant ainsi la variance pour diminuer l'erreur quadratique moyenne. Cependant, le biais augmente. Cette méthode est particulièrement utile pour atténuer la colinéarité des variables.
La pénalisation associée prend la forme : $\frac{1}{c}\sum_{i=1}^{p}\beta_i^2$

La deuxième méthode est la régularisation Lasso. Si l'idée est la même que pour la régularisation Ridge, la façon de faire est différente. Cette méthode sélectionne les variables jugées importantes pour le modèle et force le coefficient des autres à zéro. Dans le cas où deux variables, ou plus, ont un effet significatif sur le modèle mais que ces variables sont très corrélées, seule une des variables est conservée.
La pénalisation associée prend la forme : $\frac{1}{c}\sum_{i=1}^{p} \mid \beta_i \mid$

La troisième méthode se nomme l'*Elastic-Net*. Il s'agit d'une combinaison des deux méthodes précédentes puisque chacune des deux pénalisations est présente dans le problème de minimisation. Les pénalisations sont pondérées selon un paramètre λ à calibrer pour optimiser cette méthode. Si le paramètre λ est défini à zéro alors cela revient à utiliser la régularisation Ridge, s'il est défini à 1 alors cela revient à utiliser la régularisation Lasso. L'*Elastic-Net* permet une sélection plus souple des variables qu'avec la méthode Lasso tout en réduisant au mieux les coefficients des variables corrélées à l'instar de la méthode Ridge.
La pénalisation associée prend la forme : $\frac{1}{c}\left(\lambda \sum_{i=1}^{p} \mid \beta_i \mid + (1-\lambda) \sum_{i=1}^{p} \beta_i^2\right)$

Pour choisir un modèle parmi plusieurs, il est indispensable d'utiliser un indice de référence commun aux différents modèles testés. Dans le cas des modèles de classification binaire comme les régressions logistiques ou encore les forêts aléatoires de classification qui vont être utilisées par la suite, les courbes ROC sont préférées. Une courbe ROC, pour *Receiver Operating Characteristic*, représente communément le taux de vrais positifs en fonction du taux de faux positifs. Cependant, des valeurs numériques sont plus facilement comparables que des représentations graphiques. C'est pourquoi il existe l'AUC, pour *Area Under Curve*, qui comme son nom l'indique correspond à l'aire sous la courbe. L'AUC prend ses valeurs dans $[0,5;1]$. Les meilleurs modèles imaginables sont évidemment ceux avec un taux de vrais positifs proche de 1 et un taux de faux positifs proche de zéro, ce qui donne une aire sous la courbe ROC associée plus grande qu'un modèle plus médiocre. Au-delà de la comparaison entre plusieurs modèles de classification, l'AUC peut être utilisée pour juger la qualité d'un modèle.

AUC	Qualité du score
= 0.5	Non discriminant
]0.5 ; 0.6[	Peu discriminant
]0.6 ; 0.7[	Médiocre
]0.7 ; 0.8[	Bon
]0.8 ; 0.9[	Excellent
]0.9 ; 1]	Exceptionnel

TABLE 5.1 – Qualité du modèle suivant l'AUC (d'après DELALANDE 2015)

Notons toutefois que l'erreur quadratique moyenne peut aussi être utilisée. Dans un cas comme celui-ci, où la variable de réponse prend uniquement les valeurs 1 et 0, cet indice renvoie tout simplement le taux de mauvaises prédictions. L'avantage de la courbe ROC est qu'elle donne plus de précision sur les erreurs commises par le modèle.

5.2.2 Mise en application des premiers modèles

La première base de données sur laquelle la régression logistique va être appliquée dans ce mémoire est construite comme suit. Pour une maturité donnée, ici quatre ans, le score cumulé toutes branches confondues est récupéré ainsi que tous les indicateurs qui décrivent son évolution. À cela s'ajoute une variable binaire qui indique si le score est positif cinq ans plus tard en utilisant les scores des sociétaires au bout de neuf ans de maturité.

Les sociétaires qui ont résilié leurs contrats entre temps ne sont pas pris en compte. Il y a donc un nombre d'observations qui avoisine 1,9 million. Vient ensuite l'étape de calibrage du modèle sur Python. Les variables explicatives citées ci-dessus sont définies à l'exception de la variable binaire qui est celle à prédire. La proportion de données de test, à savoir 30% du jeu de donnée total, est choisie arbitrairement, soit plus d'un demi-million d'observation. Les données d'entrainement sont ensuite centrées et réduites.

Les hyperparamètres de méthode et de force de régularisation sont déterminés avec la méthode *Grid Search* couplée avec la validation croisée. Il en ressort que la méthode est la régularisation Lasso avec une force de régularisation $\frac{1}{c}$ égale à 100. Avec ces paramètres, l'AUC moyen après validation croisée est d'environ 79,77%. Les données de test sont ensuite transformées de façon identique aux données d'entrainement et le score de prédiction est calculé en appliquant le modèle logistique. L'AUC obtenu est d'environ 68,41%.

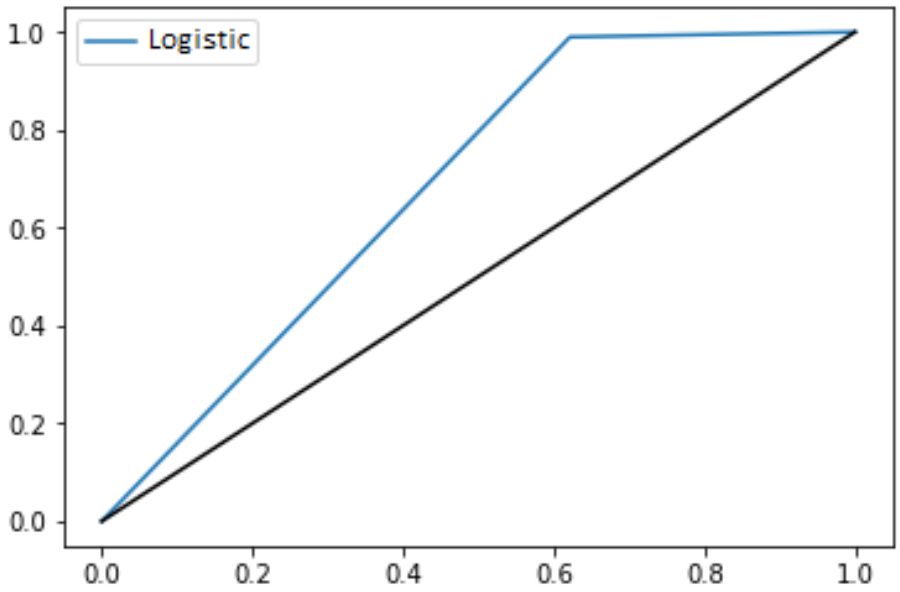

FIGURE 5.1 – Courbe ROC de la régression logistique

Au vue de la durée de la calibration, les hyperparamètres utilisés sont conservés pour les autres régressions logistique. Les coefficients sont réétablis à chaque changement de base. Voici un tableau exhaustif de l'AUC obtenue sur les données de test :

Années projetées \ Maturité	1	2	3	4	5	6	7	8
1	73.85%	63.95%	59.08%	57.26%	56.02%	55.13%	54.36%	54.03%
2	83.01%	73.62%	66.72%	63.28%	61.00%	59.34%	58.22%	
3	86.14%	78.27%	71.95%	67.85%	64.95%	63.08%		
4	88.64%	81.30%	75.64%	71.19%	68.41%			
5	89.22%	82.83%	77.59%	73.87%				
6	90.12%	84.25%	79.86%					
7	90.53%	86.02%						
8	92.77%							

FIGURE 5.2 – Tableau des AUC des regressions logistiques

Chaque ligne correspond à une maturité et chaque colonne à une période de projection. Plus la maturité est élevée, plus l'AUC obtenue est élevée. Au contraire, plus la date de projection voulue est loin de la date de départ, plus l'AUC obtenue décroit.

Ce comportement s'explique par le fait que les scores se stabilisent quand la maturité augmente et qu'il est plus difficile par nature de prévoir, avec les mêmes informations, la situation d'un sociétaire dans cinq ans que dans l'année suivante. De plus, le nombre d'observations dépend avant tout de l'année de maturité d'arrivée. Du fait des résiliations et des sociétaires entrés dans le portefeuille très récemment, ce nombre est plus faible pour une projection à la neuvième année qu'à la troisième année de maturité. Avec moins de données, il est plus difficile d'obtenir un meilleur score de projection car des observations inédites sont plus fréquemment rencontrées.

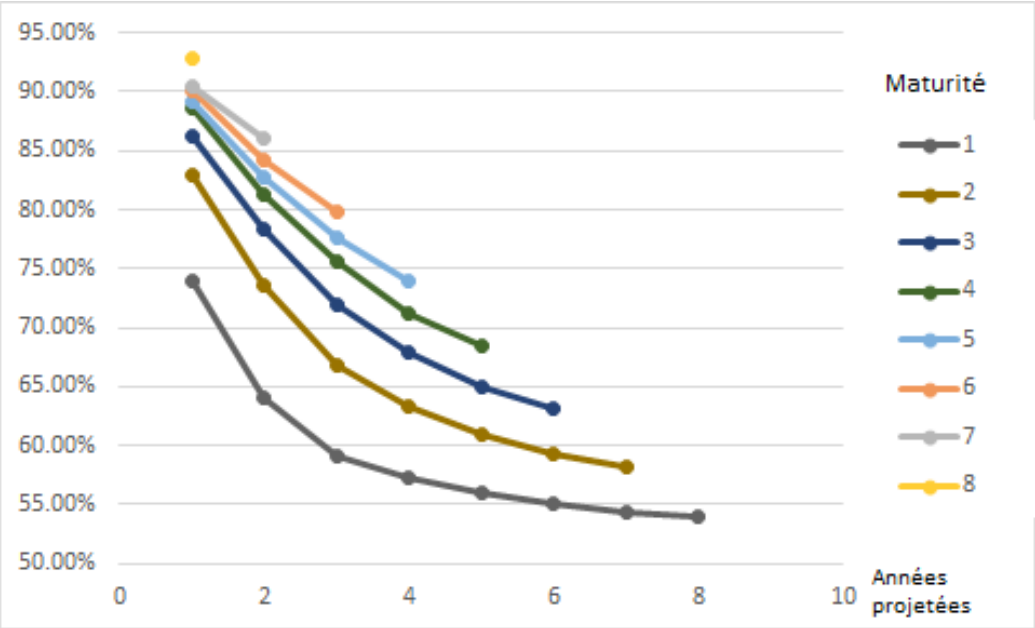

FIGURE 5.3 – AUC des régressions logistiques en fonction de la durée de projection

L'interprétation des résultats est d'autant plus visible sur ce graphique. Plus la maturité augmente, plus les courbes sont proches et chaque courbe est au-dessus de la précédente. Les modèles vont s'enrichir d'année en année avec plus d'observations dans chaque base, des projections toujours plus lointaines et des historiques encore plus profonds.

5.3 Les forêts aléatoires

Avec la croissance du nombre de données disponibles ces dernières années, plusieurs méthodes peu utilisées jusqu'à récemment refont surface. Parmi elles, les algorithmes de *machine learning* se démocratisent. Ils ont comme avantage de traiter énormément d'information en un temps raisonnable et ont démontré d'excellents résultats dans de nombreux domaines.

Comme pour la précédente, cette deuxième approche peut classer les individus en fonction d'un certain seuil de score atteint ou non dans les années à venir. Dans ce cas, des arbres de classification sont utilisés. Cependant il existe aussi des arbres de régression qui, au lieu de donner une sortie binaire, donne directement un score futur de même nature que le score présent. Le choix du type de résultat attendu dépend directement de l'usage donné à ce dernier.

Cette partie se concentrera sur les arbres de classifications d'une façon théorique puis applicative. Dans le même temps, les arbres de régressions seront mentionnés et utilisés dans un exemple afin de mieux discerner les différences entre les deux méthodes. Ensuite, les résultats de la classification par forêt aléatoire seront comparés à ceux obtenus par régression logistique.

5.3.1 Génération des arbres de décision

Un arbre de décision sépare successivement un jeu de donnée en deux jusqu'à l'obtention de feuilles qui donnent en sortie une valeur qui peut prendre plusieurs formes selon la sortie demandée. Chaque séparation se fait selon un critère mettant en scène une variable. Ainsi, si une des variables en entrée est l'année de naissance, il n'est pas impossible qu'un des nœuds sépare les données entre les individus nés avant 1975 et les autres.

La variable sélectionnée pour la séparation est la plus discriminante parmi celles proposées. L'objectif est de minimiser le coefficient de Gini à chaque étape afin d'obtenir les groupes les plus homogènes possibles.

Chaque arbre possède des observations différentes choisies aléatoirement dans la base de départ. Le tirage des observations se fait avec remise, c'est ce qui est appelé le rééchantillonnage. Cette technique permet une diversification sur les arbres de décision. Une autre technique est également utilisée pour diversifier les variables choisies. À chaque nœud de l'arbre, une partie des prédicteurs est tirée aléatoirement parmi tous. Puis l'algorithme

sélectionne le meilleur candidat pour la segmentation. Le choix des variables disponibles pour séparer le jeu de donnée est limité à $\sqrt{p}$ où p est le nombre de prédicteurs. L'idée à travers cela est d'explorer de nouveaux chemins qui n'auraient pas forcément été empruntés sans cette partie aléatoire.

Une fois l'arbre généré, une prédiction est effectuée à chaque feuille. Cette prédiction est dans le cas d'un arbre de régression, la moyenne des valeurs prises par les observations au sein de chaque groupe. Dans un arbre de classification, la prédiction prend la forme d'une classe à laquelle appartiendrait les observations de la feuille. Lorsque toute la forêt est générée, la valeur prédite par le modèle est la moyenne des valeurs prédites par chaque arbre pour les modèles de régressions et la classe attribuée en majorité par la forêt pour les modèles de classification.

De manière générale, plus le nombre d'arbre qui compose la forêt est grand, plus le modèle est précis. En revanche, le temps de calcul croît naturellement avec le nombre d'arbres générés.

5.3.2 Forêts de classification et de régression

Dans un souci de comparaison, la classification par forêt aléatoire est appliquée, dans ce mémoire, sur la même base que celle utilisée en exemple dans la régression logistique. À savoir, la base des scores de maturité 4, enrichie des indicateurs, dont le score en maturité 9 est connu. De même, la variable binaire qui permet de savoir si le score est positif cinq ans plus tard est présente dans la base. La même base d'apprentissage et la même base de test que celles utilisées dans la régression logistique le sont également ici, la base de test correspond donc à 30% du jeu de données total. Le modèle de classification par forêt aléatoire demande plusieurs hyperparamètres à déterminer. Toujours avec la méthode *Grid Search* couplée avec la validation croisée, voici les paramètres choisis :

— Le n_estimators, qui est le nombre d'arbres construits pour la classification, a été fixé à 20.

— La min_samples_leaf, qui est la quantité d'observations minimale dans une feuille, a été fixé à 2% du nombre total d'observations dans la base d'entrée.

— Le max_samples, qui est le nombre d'observations dans la base d'entrée, a été fixé à 100% du nombre total d'observations dans la base d'apprentissage.

Certains paramètres sont laissés par défaut comme le paramètre max_features qui définit combien de variables doivent être testées à chaque nœud. De base, le nombre de variables testées est inférieur à la racine carrée du nombre de variables disponibles. Toutefois, si aucune des variables testées n'est considérée comme suffisamment discriminante, d'autres variables non-tirées aléatoirement sont étudiées par l'algorithme.

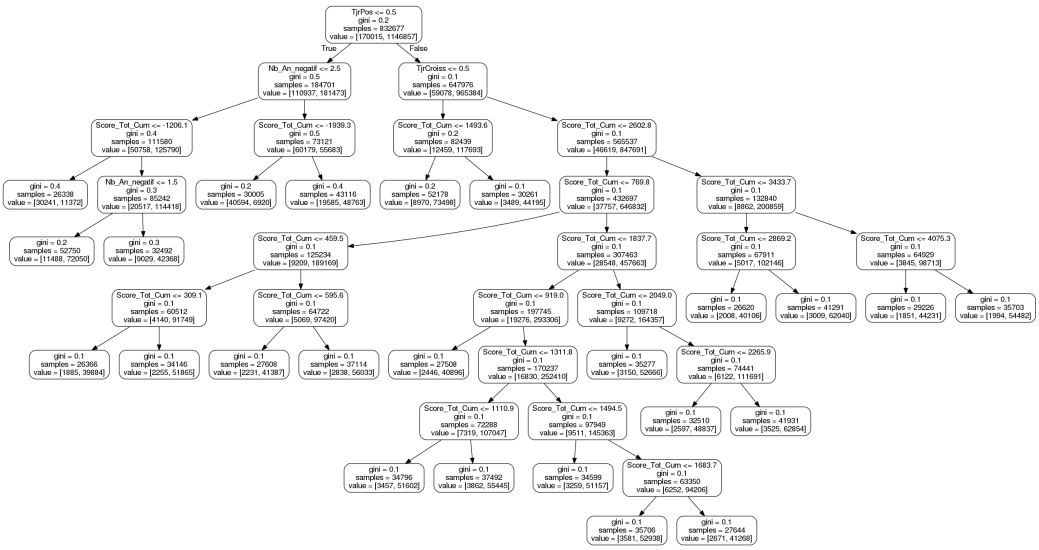

FIGURE 5.4 – Exemple d'arbre de classification

Avec cette méthode, il est possible de connaître aisément quelles variables sont les plus discriminantes et donc lesquelles sont le plus souvent utilisées lors de la séparation des données à chaque nœud.

Variable	Importance
Score_Tot_Cum	0,57
Chgt_Signe	0,17
Nb_An_negatif	0,15
TjrPos	0,11
TjrCroiss	0,01
Annee	0,0
AnneeDepSoc	0,0

TABLE 5.2 – Importance des Variables pour la classification par forêt aléatoire de maturité 4 à maturité 9

Toujours dans l'exemple, la valeur actuelle du score cumulé explique 57% de la séparation des données dans la forêt aléatoire conçue. Mais également que l'année de ce score, et donc l'année d'entrée du sociétaire dans l'étude qui sont deux variables directement liées, ne sont pas prises en compte pour discriminer le jeu de donnée. Ce qui est normal puisque les seuls sociétaires avec un score de neuf ans de maturité sont uniquement ceux qui sont entrés dans l'étude en 2011 et qui n'ont pas quitté le portefeuille depuis.

En revanche, dans le cas de la projection sur trois années des scores avec trois années de maturité, l'importance est presque surement nulle pour ces deux variables. Ce résultat peut s'expliquer ainsi :

Tout d'abord, même si pour toutes les observations les scores n'ont pas été calculés la même année, plus de 87% l'ont été en 2013 limitant les possibilités pour un nœud qui utilise cette variable d'engendrer deux sous-ensembles homogènes en soi. De plus, rassembler les scores avec la même maturité avait justement pour but d'effacer la segmentation par année civile entre les sociétaires. Ce résultat laisse penser que cette démarche a atteint son but.

Variable	Importance
Score_Tot_Cum	0,43
Chgt_Signe	0,21
Nb_An_negatif	0,2
TjrPos	0,11
TjrCroiss	0,05
Annee	0,0
AnneeDepSoc	0,0

TABLE 5.3 – Importance des Variables pour la classification par forêt aléatoire de maturité 3 à maturité 6

Dans ces conditions, l'AUC moyen après validation croisée est d'environ 80,65%. Le modèle de classification est appliqué aux données de test puis le score de prédiction est calculé. L'AUC obtenu est d'environ 71,49%.

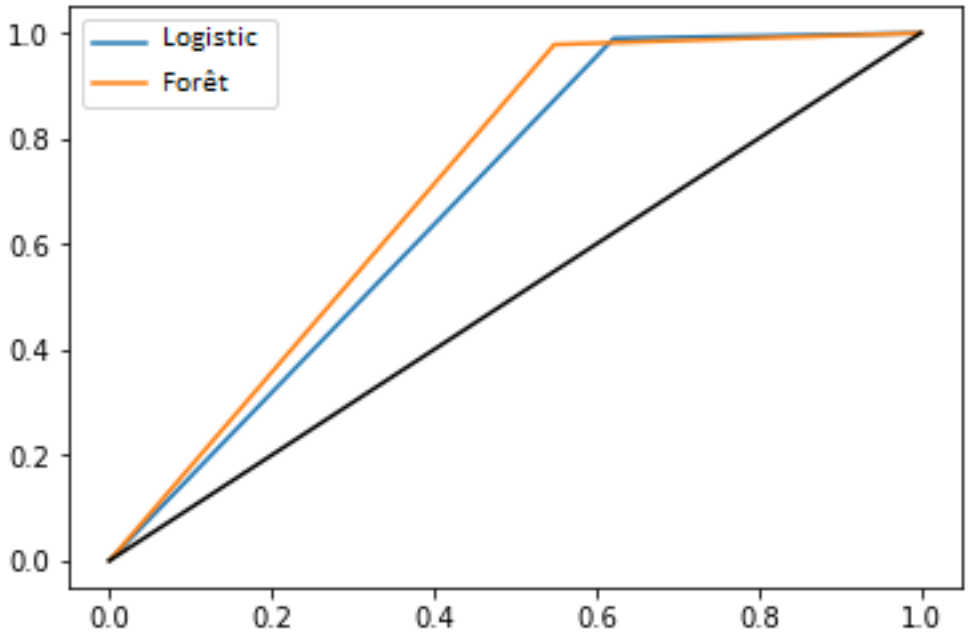

FIGURE 5.5 – Comparaison des courbes ROC des méthodes de classification

L'aire sous la courbe ROC du modèle par forêt aléatoire est visiblement supérieure à celle du modèle logistique pour cette base de test.

En utilisant les mêmes hyperparamètres et en créant une forêt aléatoire de classification par base, l'AUC est calculable sur les prédictions de chaque base de test. Les résultats sont rassemblés dans le tableau suivant.

Années projetées \ Maturité	1	2	3	4	5	6	7	8
1	72.1678%	63.9886%	60.6406%	59.0380%	57.4649%	56.9440%	56.0904%	56.2406%
2	81.0431%	72.8685%	68.5925%	66.0225%	63.3956%	61.0744%	60.9063%	
3	85.1733%	77.9473%	73.4818%	70.6191%	65.6855%	65.7631%		
4	88.4812%	81.5992%	77.1308%	73.3917%	71.4936%			
5	89.7088%	83.6541%	79.1736%	76.2246%				
6	91.0752%	85.0805%	81.6661%					
7	91.4982%	87.3736%						
8	93.6537%							

FIGURE 5.6 – Tableau des AUC des classifications par forêts aléatoires

Comme pour la régression logistique, l'AUC augmente avec la maturité de départ et diminue quand la projection se fait de plus en plus loin. Dans certains cas, les régressions logistiques correspondantes donnent de meilleurs résultats même s'il est clair qu'utiliser les forêts aléatoires est un choix plus judicieux dans l'ensemble.

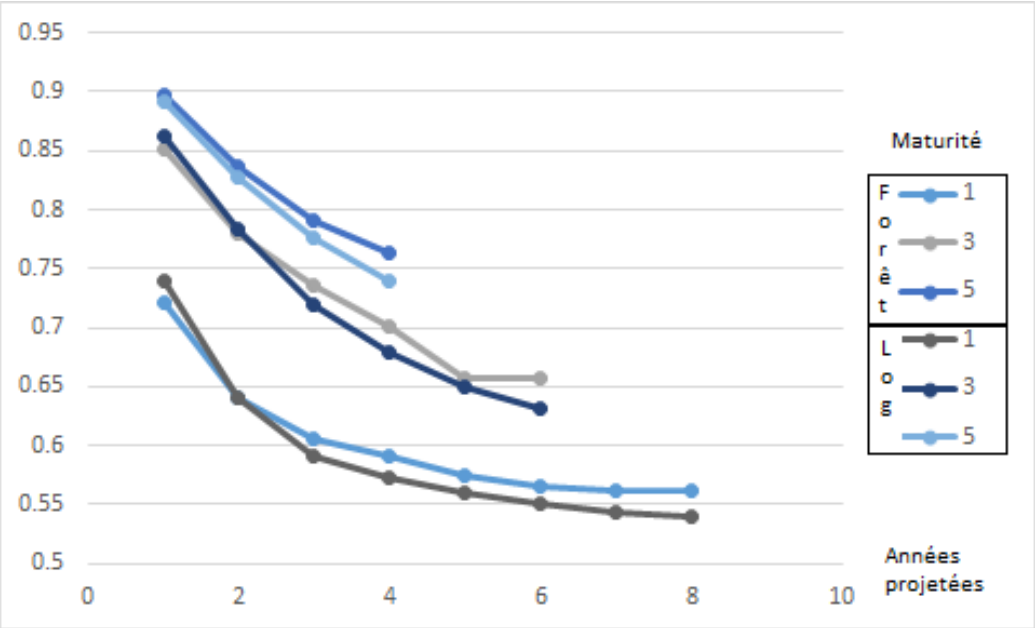

FIGURE 5.7 – Comparaison en courbes des AUC des méthodes de classification

Les forêts aléatoires de classification répondent bien à la problématique de surveillance de portefeuille mais n'apportent pas plus de précision sur la valeur du score dans le futur, pour cela les forêts aléatoires de régression sont plus adaptées. L'objectif est alors différent, les scores projetés sont une anticipation incertaine de la valeur actuelle attribuée au sociétaire. Une appréciation complémentaire du sociétaire qui inclut cette nouvelle valeur est alors concevable.

Pour ces modèles, la racine de l'erreur quadratique moyenne est préférée comme indice de qualité. Plus cet indice est bas, meilleur est le modèle.

La maturité et la durée de projection choisies pour l'application à suivre sont les mêmes que dans les exemples précédents. Seule la variable binaire a été remplacée par le véritable score à maturité 9 pour chaque sociétaire. La proportion de donnée de test est aussi identique, elle est de 30%.

La méthode *Grid Search* couplée à la validation croisée suggère la génération de 150 arbres. En revanche, il n'y a aucune différence selon la profondeur maximale des arbres, les valeurs discutées étant 10, 20 et 30. Par sécurité la profondeur maximale des arbres est fixée à 20. Le développement de chaque arbre s'arrête de sorte que chaque feuille représente plus de 2% de la base d'apprentissage. La racine de l'erreur quadratique moyenne de projection de ce modèle sur les données d'apprentissage est d'environ 14616,32, tandis que sur les données de test elle est de 16622,81.

Une méthode alternative à celle des arbres de classification est de se servir des scores projetés et de les comparer au seuil fixé pour classer les sociétaires. L'AUC alors obtenu sur la base de test est d'environ 71,95%. Il se trouve que dans l'ensemble, cette méthode est meilleure à la fois que la régression logistique et que la classification directe par forêt aléatoire.

Une fois ordonnés, les scores projetés peuvent aussi être source de condition pour la défense de portefeuille. Néanmoins, il existe d'autres méthodes que les forêts aléatoires de régression pour obtenir ces scores.

5.4 Une piste empirique avec des matrices de transition

À partir des bases par maturité, la comparaison des scores cumulés d'une même base est pertinente. Les sociétaires peuvent alors être ordonnés et un rang peut leur être attribué. Ce rang dépend du dernier vingtile dépassé. Ainsi, un sociétaire dont le score se situe entre le troisième et le quatrième vingtile aura un rang de 3. Les rangs vont donc de 0 pour les scores les plus bas à 19 pour les meilleurs scores et évoluent comme le score avec la maturité. Dans cette section, le rang d'un sociétaire de maturité m sera noté X_m.

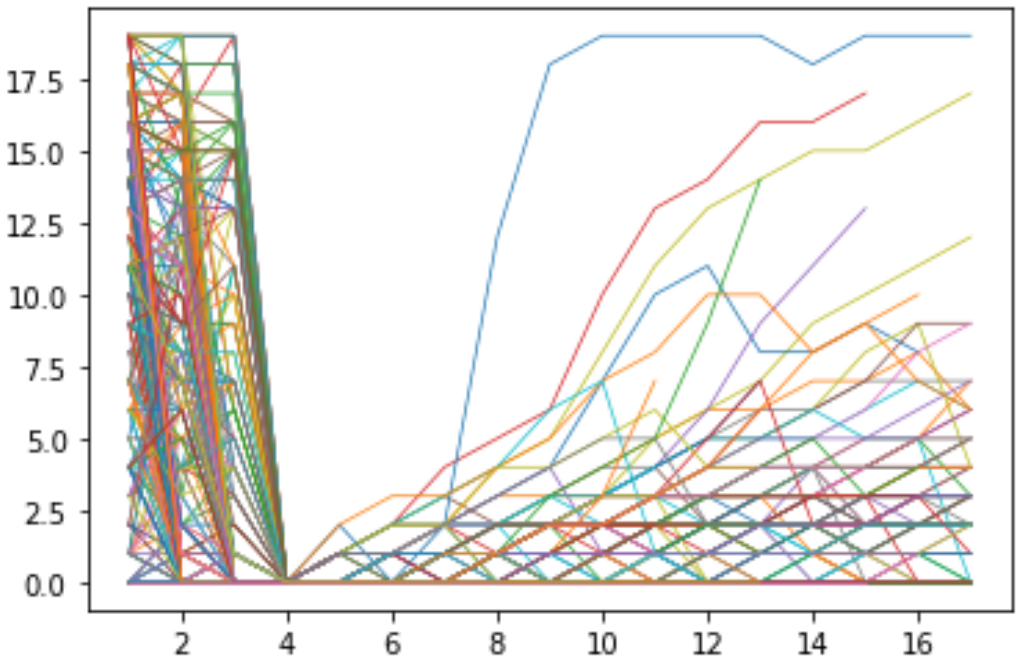

FIGURE 5.8 – Rang en fonction de la maturité avec $X_4 = 0$

Après avoir définis l'aspect théorique nécessaire à leurs compréhension, plusieurs tentatives de projection de scores seront illustrées avant d'être comparées à la méthode de régression par forêt aléatoire.

5.4.1 Exploration théorique

En connaissant le rang de nombreux sociétaires à travers le temps, des matrices de transition empiriques peuvent être construites. $P_{m1,m2}$ est la matrice de transition de la maturité $m1$ à la maturité $m2$ telle que $P_{m1,m2}(i,j) = \mathbb{P}(X_{m2} = j \mid X_{m1} = i)$ pour tout $(i,j) \in [\![0; 19]\!]^2$.

Par construction ces matrices sont stochastiques, il sera important de vérifier en premier lieu si le processus d'évolution du rang dans le temps est

un processus markovien, c'est-à-dire si :
$\mathbb{P}(X_{n+1} \mid X_n) = \mathbb{P}(X_{n+1} \mid X_n, X_{n-1}, ...)$

Les matrices de transition serviront notamment d'outils pour projeter le score des sociétaires, soit en utilisant le rang moyen soit en calculant directement le score moyen, en fonction du rang présent et de la maturité de départ.

En effet, soit $f_{m1,m2}(X_{m1}, X_{m2})$ la fonction qui renvoie le score moyen d'un sociétaire à maturité $m2$ sachant X_{m1} et X_{m2} ; avec $P_{m1,m2}$ connu, $\lfloor E(X_{m2} \mid X_{m1}) \rfloor$ est estimable donc $f_{m1,m2}(X_{m1}, \lfloor E(X_{m2} \mid X_{m1}) \rfloor)$ aussi.

Une autre projection du score possible est $\hat{E}(f_{m1,m2}(X_{m1}, X_{m2}) \mid X_{m1})$.

5.4.2 Application

À l'aide de la commande Proc Freq de SAS, les matrices de transition du rang d'une année de maturité à l'autre sont obtenues.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	67.43%	30.17%	1.82%	0.19%	0.08%	0.07%	0.05%	0.04%	0.03%	0.03%	0.01%	0.01%	0.01%	0.01%	0.01%	0.00%	0.01%	0.01%	0.00%	0.01%
2	6.94%	35.17%	38.51%	7.09%	3.39%	2.54%	1.95%	1.33%	0.91%	0.68%	0.49%	0.33%	0.22%	0.15%	0.10%	0.09%	0.06%	0.03%	0.02%	0.02%
3	3.26%	6.70%	27.39%	18.76%	10.46%	8.72%	7.36%	5.23%	3.87%	2.69%	1.95%	1.31%	0.88%	0.56%	0.34%	0.22%	0.15%	0.08%	0.04%	0.02%
4	1.20%	2.15%	4.60%	58.54%	9.07%	5.82%	5.56%	4.14%	3.08%	2.04%	1.36%	0.89%	0.59%	0.38%	0.27%	0.14%	0.10%	0.05%	0.02%	0.01%
5	1.01%	1.89%	3.08%	14.84%	49.73%	8.93%	5.35%	4.64%	3.58%	2.50%	1.65%	0.99%	0.71%	0.42%	0.30%	0.18%	0.10%	0.04%	0.02%	0.01%
6	0.84%	1.56%	2.72%	1.67%	26.72%	42.54%	7.85%	4.44%	4.01%	2.78%	1.76%	1.17%	0.73%	0.47%	0.29%	0.23%	0.11%	0.07%	0.02%	0.01%
7	1.00%	1.66%	2.82%	1.53%	2.46%	24.59%	40.93%	8.28%	5.23%	4.18%	2.70%	1.76%	1.05%	0.77%	0.44%	0.28%	0.18%	0.09%	0.04%	0.01%
8	1.08%	1.75%	2.70%	1.38%	1.21%	3.54%	19.97%	41.92%	9.81%	5.81%	4.23%	2.50%	1.59%	1.02%	0.68%	0.43%	0.21%	0.13%	0.05%	0.01%
9	1.21%	1.85%	2.58%	1.29%	0.99%	1.32%	4.94%	17.35%	39.17%	11.76%	6.73%	4.33%	2.58%	1.62%	1.00%	0.66%	0.35%	0.19%	0.07%	0.01%
10	1.35%	2.02%	2.31%	1.39%	0.93%	1.09%	1.67%	5.40%	15.77%	36.50%	13.51%	7.29%	4.60%	2.65%	1.62%	0.94%	0.56%	0.27%	0.10%	0.02%
11	1.50%	2.05%	1.92%	1.30%	1.04%	1.06%	1.22%	2.03%	6.04%	15.15%	33.67%	14.97%	7.88%	4.51%	2.60%	1.55%	0.86%	0.46%	0.15%	0.03%
12	1.53%	1.98%	1.89%	0.90%	0.94%	1.12%	1.20%	1.28%	2.59%	6.70%	15.26%	32.11%	15.87%	7.70%	4.29%	2.43%	1.31%	0.62%	0.22%	0.05%
13	1.50%	1.91%	1.84%	0.80%	0.63%	0.84%	1.26%	1.28%	1.45%	3.19%	7.18%	15.25%	30.99%	16.57%	7.79%	4.02%	2.10%	0.98%	0.37%	0.06%
14	1.52%	1.81%	1.62%	0.67%	0.55%	0.67%	0.91%	1.20%	1.50%	1.65%	3.70%	7.43%	15.27%	31.20%	16.93%	7.44%	3.64%	1.57%	0.59%	0.11%
15	1.48%	1.72%	1.52%	0.58%	0.49%	0.51%	0.73%	0.83%	1.21%	1.56%	1.77%	3.82%	7.47%	15.45%	32.07%	17.32%	7.29%	2.97%	1.02%	0.18%
16	1.60%	1.60%	1.36%	0.55%	0.40%	0.48%	0.57%	0.63%	0.85%	1.16%	1.53%	2.03%	3.91%	7.44%	15.52%	33.61%	17.82%	6.58%	2.02%	0.33%
17	1.56%	1.39%	1.13%	0.49%	0.37%	0.43%	0.50%	0.59%	0.71%	0.91%	1.07%	1.44%	1.99%	3.77%	7.45%	15.96%	35.89%	18.58%	5.02%	0.75%
18	1.48%	1.22%	0.93%	0.40%	0.29%	0.31%	0.42%	0.47%	0.55%	0.63%	0.77%	0.94%	1.25%	1.80%	3.37%	7.17%	16.76%	39.95%	19.16%	2.14%
19	1.46%	1.07%	0.77%	0.30%	0.23%	0.26%	0.30%	0.32%	0.39%	0.40%	0.55%	0.61%	0.81%	1.02%	1.42%	2.54%	6.11%	17.66%	48.77%	15.00%
20	1.53%	0.76%	0.50%	0.18%	0.13%	0.16%	0.19%	0.18%	0.21%	0.23%	0.30%	0.30%	0.39%	0.50%	0.59%	0.87%	1.45%	3.57%	15.03%	72.93%

FIGURE 5.9 – $P_{4,6}$

Par construction cette matrice est stochastique, malheureusement le processus d'évolution du rang dans le temps n'est pas un processus markovien, la caractérisation
$\mathbb{P}(X_{n+1} = x \mid X_n) = \mathbb{P}((X_{n+1} = x \mid X_n, X_{n-1}, ...)$ n'est pas respectée. C'est

vérifiable en conditionnant la matrice précédente selon le rang du sociétaire à sa deuxième année de cotisation.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	68.46%	29.46%	1.59%	0.17%	0.07%	0.05%	0.05%	0.03%	0.02%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%	0.00%	0.01%	0.01%	0.00%	0.01%
2	6.49%	37.81%	39.19%	6.20%	2.93%	2.15%	1.64%	1.09%	0.75%	0.53%	0.40%	0.28%	0.18%	0.12%	0.08%	0.06%	0.05%	0.03%	0.02%	0.01%
3	2.85%	6.37%	32.96%	20.78%	9.90%	7.48%	6.10%	4.22%	3.04%	2.10%	1.49%	1.01%	0.70%	0.41%	0.23%	0.15%	0.12%	0.07%	0.02%	0.01%
4	0.92%	1.77%	4.40%	64.86%	9.37%	5.25%	4.29%	3.06%	2.24%	1.41%	0.87%	0.58%	0.39%	0.23%	0.16%	0.09%	0.05%	0.03%	0.01%	0.01%
5	0.78%	1.60%	2.81%	16.05%	54.50%	9.16%	4.57%	3.48%	2.55%	1.69%	1.10%	0.63%	0.44%	0.26%	0.16%	0.11%	0.06%	0.02%	0.01%	0.01%
6	0.65%	1.34%	2.49%	1.59%	28.84%	45.89%	7.71%	3.63%	2.88%	1.86%	1.19%	0.78%	0.46%	0.27%	0.17%	0.13%	0.07%	0.04%	0.01%	0.00%
7	0.78%	1.40%	2.67%	1.47%	2.52%	26.62%	44.65%	7.86%	4.18%	3.02%	1.88%	1.17%	0.67%	0.48%	0.27%	0.18%	0.10%	0.05%	0.02%	0.00%
8	0.87%	1.50%	2.56%	1.31%	1.16%	3.22%	21.30%	46.73%	9.06%	4.76%	3.14%	1.75%	1.03%	0.66%	0.44%	0.27%	0.13%	0.07%	0.02%	0.01%
9	1.02%	1.66%	2.47%	1.21%	0.95%	1.29%	4.06%	17.81%	44.99%	11.05%	5.65%	3.33%	1.88%	1.11%	0.68%	0.43%	0.22%	0.11%	0.05%	0.01%
10	1.39%	2.12%	2.40%	1.40%	0.99%	1.17%	1.61%	4.08%	14.41%	36.42%	15.07%	8.18%	4.82%	2.67%	1.56%	0.86%	0.51%	0.25%	0.07%	0.02%
11	2.12%	2.79%	2.18%	1.29%	1.03%	1.22%	1.46%	1.73%	3.97%	7.83%	20.14%	22.29%	14.51%	7.97%	4.51%	2.60%	1.37%	0.74%	0.22%	0.03%
12	2.25%	2.69%	2.35%	0.99%	0.93%	1.05%	1.47%	1.53%	2.17%	4.41%	7.23%	14.64%	21.89%	16.57%	9.51%	5.51%	2.88%	1.29%	0.51%	0.12%
13	2.45%	2.60%	2.25%	1.01%	0.66%	0.82%	1.18%	1.46%	1.71%	2.52%	4.29%	6.69%	11.41%	20.89%	19.23%	11.07%	6.01%	2.72%	0.87%	0.17%
14	2.47%	2.26%	2.05%	0.81%	0.71%	0.91%	0.99%	1.07%	1.45%	1.84%	3.08%	4.68%	6.94%	10.51%	19.88%	20.43%	12.00%	5.60%	1.90%	0.42%
15	2.47%	2.23%	1.91%	0.71%	0.66%	0.61%	0.95%	0.73%	1.19%	1.36%	1.63%	2.79%	4.81%	7.98%	10.66%	19.82%	22.86%	11.86%	4.03%	0.75%
16	2.51%	2.25%	2.01%	0.58%	0.55%	0.58%	0.81%	0.75%	1.10%	1.15%	1.54%	1.68%	3.06%	5.60%	7.66%	11.11%	21.55%	24.01%	9.81%	1.68%
17	2.75%	1.72%	1.32%	0.53%	0.34%	0.46%	0.53%	0.84%	0.76%	1.01%	1.16%	1.22%	1.79%	3.03%	6.01%	8.45%	13.68%	26.78%	23.54%	4.06%
18	2.50%	1.63%	1.35%	0.59%	0.37%	0.39%	0.31%	0.73%	0.51%	0.51%	0.99%	1.07%	1.72%	1.66%	3.46%	6.45%	11.34%	16.72%	33.66%	14.04%
19	2.68%	1.37%	0.56%	0.66%	0.20%	0.25%	0.30%	0.35%	0.61%	0.81%	0.66%	0.81%	1.11%	1.06%	1.72%	2.43%	6.68%	14.78%	24.61%	38.33%
20	4.23%	1.41%	1.06%	0.00%	0.00%	0.35%	0.35%	0.23%	0.59%	0.59%	0.23%	0.70%	0.12%	0.70%	0.94%	1.17%	1.53%	5.75%	14.55%	65.49%

FIGURE 5.10 – $P_{4,6}$ sachant $X_2 < 10$

Ces matrices sont différentes. En particulier :

$$\mathbb{P}\left(X_6 = 14 \mid X_4 = 12)\right) \neq \mathbb{P}\left(X_6 = 14 \mid X_4 = 12, X_2 < 10\right)$$

Le processus d'évolution du rang n'est alors pas un processus markovien, les matrices $P_{n,n+1}$ sont donc insuffisantes pour construire chacun des chemins possibles et la probabilité qu'ils se produisent.

Une première façon de se servir des matrices de transitions pour projeter le score futur est de calculer dans un premier temps le rang moyen atteint par les sociétaires avec un rang identique pour la maturité de départ étudiée, puis d'attribuer le score moyen obtenu par les sociétaires avec un rang identique pour la maturité d'arrivée fixée. Pour chaque couple de maturité $(m1, m2)$ il existe 20 sorties possibles.

X_4	$\lfloor \hat{E}(X_9 \mid X_4) \rfloor$	$f_{4,9}\left(X_4, \lfloor \hat{E}(X_9 \mid X_4) \rfloor\right)$
0	1	-2133,15
1	3	513,70
2	5	1185,58
3	4	884,73
4	4	841,99
5	5	1142,07
6	5	1195.86
7	6	1516,80
8	7	1831,64
9	8	2152,37
10	9	2482,98
11	9	2478,32
12	10	2821,91
13	11	3174,38
14	11	3169,54
15	12	3546,31
16	13	3949,84
17	14	4404,12
18	15	4923,06
19	17	6407,21

TABLE 5.4 – Tableau des scores moyens de l'estimation des espérances de X_9 sachant X_4

Les résultats sont obtenus empiriquement, les espérances du rang futur calculées ne sont que des estimations. Heureusement, c'est la partie entière qui est déterminante. Pour vérifier la bonne qualité de cette estimation, un test de Student est réalisé sur nos données avant de calculer les bornes de l'intervalle de confiance à 95% de la moyenne empirique du rang futur. Dans le cas d'une maturité de départ étudiée de quatre ans et une maturité future fixée à 9, l'intervalle de confiance est toujours inclus dans un intervalle de la forme $[r; r+1[$ avec $r \in \mathbb{N}$ à l'exception du cas où $X_4 = 1$, alors l'intervalle de confiance est borné par approximativement 3,966 et 4,015. Après vérification, les scores sont projetés selon le rang de chaque sociétaire. La racine de l'erreur

quadratique moyenne obtenue est de 15415,72.

Dans les cas très rares où l'intervalle de confiance serait à cheval entre deux rangs décisifs, la décision finale reviendrait au collaborateur concerné. Néanmoins, les bases comporteront de plus en plus d'observations avec l'arrivée de nouveaux sociétaires ainsi qu'avec les sociétaires restant dans le portefeuille une année supplémentaire. La moyenne empirique du rang sera alors de plus en plus fiable si les lois sous-jacentes sont considérées comme stables dans le temps.

La seconde manière de projeter le score avec les matrices de transition est de d'abord calculé les scores moyens pour chaque couple (X_{m1}, X_{m2}) avant de calculer la moyenne empirique des scores moyens connaissant X_{m1}. Pour chaque couple de maturité $(m1, m2)$ il existe de nouveau 20 sorties possibles.

X_4	$\hat{E}\left(f_{4,9}\left(X_4, X_9\right) \mid X_4\right)$
0	-9602,93
1	-282,48
2	717,02
3	747,45
4	930,42
5	1037,65
6	1238,00
7	1527,02
8	1745,60
9	1962,70
10	2237,94
11	2423,63
12	2733,28
13	3025,48
14	3355,98
15	3717,65
16	4249,11
17	4846,04
18	5810,05
19	8955,96

TABLE 5.5 – Tableau des estimations de l'espérances du score moyen à maturité 9 sachant X_4

Les scores projetés sont différents de ceux calculés avec la méthode précédente. Après projection sur la même base de score, la racine de l'erreur quadratique moyenne obtenue est de 15308,22.

Dans cet exemple, le meilleur modèle de projection est celui qui utilise cette méthode. Or ce n'est pas le cas dans l'ensemble, aucun modèle ne se détache réellement des autres. Chacune des trois méthodes est sensiblement meilleure que les deux autres pour au moins une base de scores à projeter. Dans tous les cas, la génération de la forêt aléatoire prend plus de temps que la génération des matrices.

Le fait de n'avoir que 20 sorties n'est pas toujours idéal. Cependant, il est possible d'augmenter le nombre de sorties en augmentant le nombre de rang en amont. En faisant cela, le nombre de données pour chaque rang est diminué, ce qui affaiblit les estimations. Une solution serait de rechercher un nombre de rang tel que la précision d'un des deux modèles soit optimale.

Une autre façon d'utiliser ces matrices peut être d'observer de manière plus synthétique les probabilités pour un sociétaire de dépasser un certain seuil de rang dans le futur. On rejoint ici la problématique de défense du portefeuille mais avec une vision projetée et non actuelle même si dans les faits, les meilleurs scores d'aujourd'hui sont les plus à même de devenir les meilleurs scores à l'avenir.

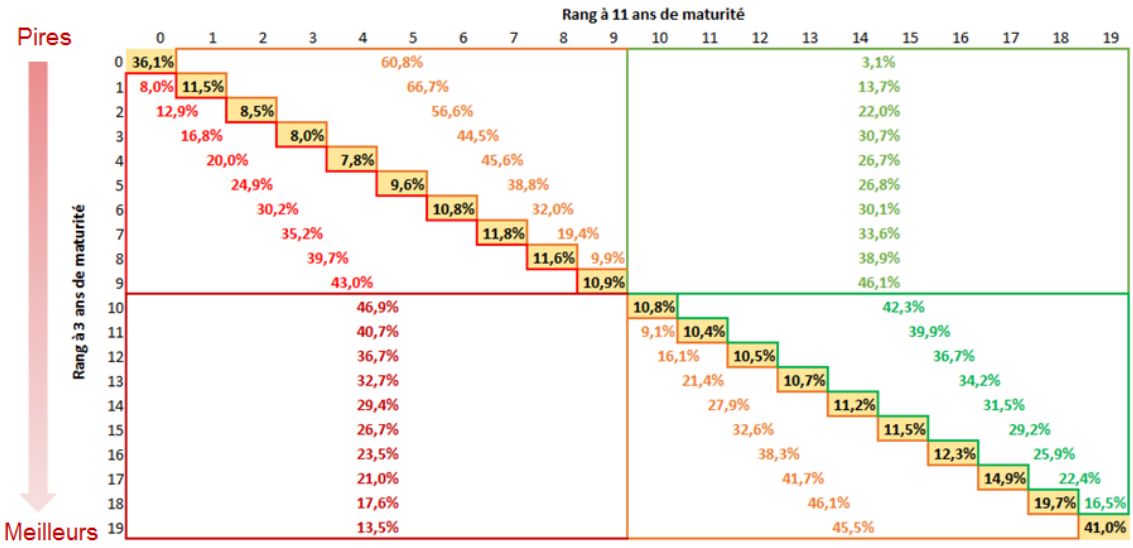

FIGURE 5.11 – Synthèse de $P_{3,11}$ avec un seuil au rang 10

Chapitre 6

Conclusion

Manipuler beaucoup de données est à double tranchant. D'un côté, le champ des possibles à explorer est colossal et permet d'effectuer un grand nombre d'études sans inquiétude. De l'autre, chaque action appliquée à ces données est gourmande en ressources ce qui oblige le partitionnement en de multiples bases et ralentit les différentes analyses.

Le score et les autres outils mis en place pour son étude ou son utilisation ont été conçus en tenant compte des volontés de départ à savoir qu'ils sont relativement accessibles à la compréhension, l'utilisation et l'interprétation. La seconde demande était d'établir un outil polyvalent et flexible dans son utilisation ce qui a été respecté puisque le choix de la période et des branches étudiées est possible dans la limite de l'historique et des branches présents aujourd'hui.

Même pris seul, le score s'est révélé être une très bonne information de synthèse à partir de laquelle énormément d'informations sont déductibles. Connaître le chemin déjà parcouru par le sociétaire est essentiel pour savoir dans quelle direction se dirige le score. Avec la quantité de données disponible, des classifications par forêts aléatoires de bonne qualité ont été réalisables en addition des méthodes de régressions logistiques considérés

comme plus classiques. Dans certaines conditions, ces GLM ont d'ailleurs été plus performantes que les algorithmes de *machine learning* utilisés. Toutefois ces derniers ont su montrer leur supériorité pour des projections à plus long terme.

En ce qui concerne la projection d'une valeur du score, la méthode de projection par estimation de l'espérance du score moyen apporte de meilleurs résultats sur l'exemple étudié que la méthode par forêt aléatoire. En revanche, la régression par forêt aléatoire apporte plus de continuité dans les scores projetés. Par contre le temps de calcul est plus important. Aucune des trois méthodes n'est souveraine, tout dépend de la maturité de départ et d'arrivée.

Il est probable que le score devienne une information incontournable pour l'assureur à l'avenir. L'outil mis en place va déjà faciliter les prises de décisions et a pour vocation de se développer avec les années. Tout d'abord en intégrant les frais et la dimension foyer. Ces ajouts procureront un point de vue plus proche de la réalité vis-à-vis de la dynamique des sociétaires et de la vie de leurs contrats à travers un score de plus en plus précis.

L'addition d'autres branches comme l'assurance scolaire et la protection juridique se fera progressivement pour tendre vers un périmètre plus exhaustif. Pour la même raison, étendre l'historique est prévue. L'ajout d'un modèle de résiliation et de souscription est aussi envisageable dans les années à venir.

Le score conçu à pour but de mêler à terme une vision rétrospective mais aussi prospective afin d'anticiper des comportements futurs dans la vie des différents contrats. De plus, il s'agirait d'un des premiers score mis en place par un organisme d'assurance qui ne se limite pas à une vision par contrat mais qui propose également une vision par sociétaire et même par foyer de sociétaire dans un futur proche.

L'outil présenté dans ce mémoire reste un aperçu d'un projet très vaste et par conséquent inachevé. Les frais ne sont pas encore pris en compte et l'impact de la réassurance dans le calcul du score est ignoré. De nouveaux modules pourront s'ajouter à l'outil dans le futur tel que des modèles annexes d'inflation des sinistres et des cotisations, notamment dans le cadre de la vision prospective.

Aussi il est difficile de projeter la rentabilité d'un client sur un horizon de plusieurs années du fait de l'incertitude relative aux hypothèses nécessaires que ce soit l'inflation, les politiques tarifaires à venir, mais également le comportement des assurés et les évolutions technologiques ou sociétales de ces prochaines années compliquées à imaginer et à retranscrire sous la forme d'une formule mathématique ou d'un programme informatique.

Enfin, bien que très polyvalent dans ses usages, le score seul n'est pas la solution universelle à toutes les problématiques rencontrées en assurance. Selon le cas d'usage mis en face, cet outil doit être complété par d'autres instruments plus spécifiques comme un modèle de fréquence de sinistre pour la surveillance de portefeuille puisqu'on cherchera ici à sanctionner des comportements sinistrogènes.

La démarche présentée dans ce mémoire sur la création d'un score mêlant valeurs Actuelle et Future résonne à la célèbre citation : "Celui qui ne sait pas d'où il vient ne peut savoir où il va"[1]

1. Otto de Habsbourg, Le Nouveau défi européen, 2007

Table des figures

Liste des tableaux

Bibliographie

CARIA, A (2012). "«Contrat de fidélité bancassurance» Modélisation de la valeur d'un client bancassurance à l'aide d'un score et proposition d'un tarif adapté sur le produit d'appel qu'est l'automobile en cas de souscription d'un contrat de fidélité". mémoire d'actuariat. CEA.

CHOQUET, C (2011). "Structuration d'une offre pour les jeunes conducteurs". mémoire d'actuariat. Université Paris Dauphine.

DELALANDE, J (2015). "Rétention au changement de véhicule". mémoire d'actuariat. ENSAE.

DURAND, T (2016). "Évaluation et optimisation de la rentabilité d'un portefeuille automobile". mémoire d'actuariat. EURIA.

VANNEAUX, J-C (2010). "Modélisation de la valeur contrat (PNPV) pour le pilotage du portefeuille automobile particulier". mémoire d'actuariat. ISFA.

Annexe A

Arbre de décision

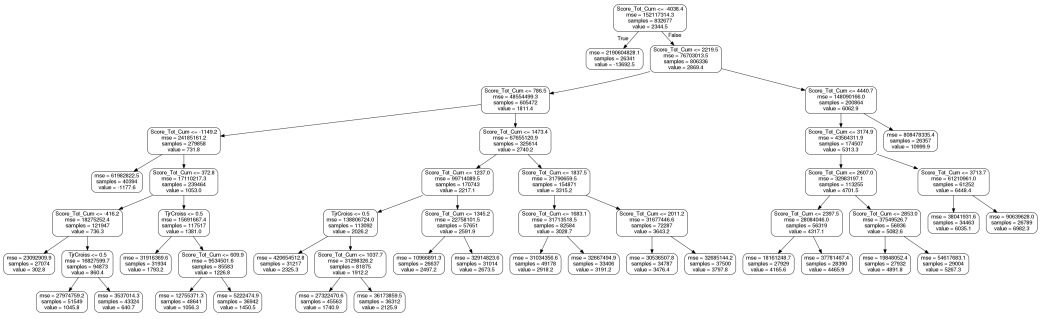

FIGURE A.1 – Exemple d'arbre de régression

Annexe B

Qualités de projection

Années projetées \ Maturité	1	2	3	4	5	6	7	8
1	0.7243759	0.658568	0.624751	0.607855	0.574957	0.569455	0.563106	0.562621801
2	0.830147	0.749021	0.691356	0.669497	0.652869	0.638526	0.592258	
3	0.860276	0.795417	0.755536	0.709721	0.69304	0.684239		
4	0.8970862	0.830099	0.791905	0.757679	0.719471			
5	0.895847	0.848724	0.809778	0.756969				
6	0.9047815	0.864654	0.837327					
7	0.9229884	0.888322						
8	0.9403941							

FIGURE B.1 – Tableau des AUC par la méthode des forêts aléatoires de régression

Années projetées \ Maturité	1	2	3	4	5	6	7	8
1	12585.38	7047.33	18781.28	22371.91	10895.18	23160.65	12869.6	20258.3
2	6971.85	7910.75	16016.83	18898.37	17941.57	15751.31	13200.3	
3	19358.15	16450.68	11389.09	12502.67	14387.02	21353.53		
4	15895.77	21864.69	16065.99	12441.22	16622.81			
5	9567.206	12049.04	18656.51	19732.62				
6	23157.3	18664.28	14873.8					
7	12792.79	13444.15						
8	19800.47							

FIGURE B.2 – Tableau des racines des erreurs quadratiques moyennes par la méthode des forêts aléatoires de régression

Années projetées \ Maturité	1	2	3	4	5	6	7	8
1	8936.781	13397.36	14891.92	16001.71	17507.87	14607.64	15461.7	15595.43
2	13326.44	14860.53	15959.34	17463.87	18344.76	15404.49	15542.9	
3	14754.53	15906.7	17446.2	18310.38	15344.63	15471.59		
4	15792.55	17345.29	18232.47	15288.25	15415.72			
5	17227.74	18121	15159.76	15297.94				
6	17997.26	15007.37	15155.87					
7	14607.64	15023.47						
8	14912.97							

FIGURE B.3 – Tableau des racines des erreurs quadratiques moyennes par la méthode par moyenne du score des estimations de l'espérance du rang

Années projetées \ Maturité	1	2	3	4	5	6	7	8
1	8904.319	13375.85	14868.87	15978.32	17485.62	18352.12	15430.69	15564.96
2	13309.145	14813.94	15923.95	17434.52	18298.49	15346.41	15474.38	
3	14740.972	15869.92	17383.53	18245.56	15274.79	15405.17		
4	15780.197	17309.96	18176.42	15187.82	15308.22			
5	17214.045	18091.79	15090.8	15209.81				
6	17981.516	14969.86	15087.19					
7	14584.637	14991.27						
8	14888.227							

FIGURE B.4 – Tableau des racines des erreurs quadratiques moyennes par la méthode par estimation de l'espérance du score moyen