Mémoire présenté le : 14 juin 2022

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

Par : Laura RESSOT

Titre Détection automatique de populations à risque au sein d'un portefeuille d'assurance non-vie basée sur un indicateur de rentabilité et optimisation tarifaire

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

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

Membre présents du jury de l'Institut des Actuaires signature

François-Xavier NEGRI

Guillaume BENETEAU

Membres présents du jury de l'ISFA

Pierre RIBEREAU

Entreprise :

Nom : addactis® Software

Signature :

Directeur de mémoire en entreprise :

Nom : Michael CASALINUOVO

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Remerciements

Je tiens tout d'abord à remercier mon tuteur en entreprise, Michael Casalinuovo, pour sa pédagogie, son aide précieuse ainsi que le temps qu'il m'a accordé. Ce fut un réel plaisir de travailler à ses côtés. Je remercie également Bilal Sadou qui m'a accompagné durant mes débuts.

Je remercie l'ensemble des membres du groupe addactis® que j'ai pu côtoyer, tout d'abord pour leur confiance ainsi que pour leurs précieux conseils.

Je remercie de façon générale l'ensemble de l'équipe pédagogique de l'ISFA pour la formation complète que j'ai reçue. J'adresse plus particulièrement mes remerciements à Denis Clot, pour sa disponibilité et son aide.

Enfin, je remercie mes proches pour leur soutien tout au long de l'élaboration de ce mémoire.

Résumé

Le cycle de production du secteur de l'assurance étant inversé, la rentabilité d'un contrat d'assurance ne peut être facilement déterminée et notamment lors de sa tarification. La prime doit donc être régulièrement réévaluée afin que les tarifs demeurent adaptés. En effet, les majorations à la date d'anniversaire sont une pratique courante pour les assureurs mais leur efficacité et pertinence peuvent être remises en question. Le secteur de l'assurance étant fortement concurentiel, il est de plus en plus courant pour les assurés de changer d'assureur et cette perte de parts de marché peut s'avérer infructueuse. Une analyse plus fine du portefeuille d'assurance et la détection automatique des populations les moins rentables peut se révéler être un réel atout stratégique.

Ce mémoire répond ainsi à deux objectifs principaux : le premier est la mise en place de processus automatisés permettant d'analyser la rentabilité d'un portefeuille d'assurance et de détecter des populations à risque pour l'assureur. Le second objectif de cette étude est, compte-tenu des populations détectées, d'optimiser la rentabilité globale du portefeuille en déterminant la prime optimale.

Pour cela, un indicateur de pilotage de rentabilité, l'Expected Loss Ratio (ELR), sera utilisé. Nous étudierons cet indicateur et, par le biais de l'algorithme Patient Rule Induction Method (*PRIM*), nous détecterons des segments (sous-ensembles du portefeuille de clients) sur lesquels l'ELR dévie fortement de sa moyenne, que ce soit à la hausse ou à la baisse.

Ensuite, nous étudierons le comportement client des assurés de notre portefeuille concernant l'acte de résiliation. Nous modéliserons également l'impact du tarif sur le comportement client à l'aide d'un indicateur : l'élasticité.

Enfin, une approche plus globale sera entreprise : nous chercherons à optimiser le profit de l'assureur compte-tenu de ses contraintes de volume de portefeuille. Pour cela, des méthodes mathématiques d'optimisation seront utilisées pour lesquelles la variable d'influence sera la prime commerciale payée par l'assuré.

Mots-clés : Assurance automobile, Modèles Linéaires Généralisés (GLM), résiliation, portefeuille d'assurance, Expected Loss Ratio (ELR), algorithme PRIM (Patient Rule Induction Method)

Abstract

As the production cycle of the insurance sector is reversed, the profitability of an insurance contract cannot be achieved easily, particularly during its pricing. The premium must therefore be regularly reassessed so that insurance rates remain adapted. Indeed, increases on the anniversary date are a common practice for insurers but their effectiveness and relevance can be called into question. The insurance sector being highly competitive, it is increasingly common for policyholders to change insurer and this loss of market share may prove unsuccessful. A more detailed analysis of the insurance portfolio and the automatic detection of the least profitable populations may be a truly strategic asset.

This thesis thus meets two major objectives : the first is the implementation of automated processes to analyze the profitability of an insurance portfolio and to detect populations at risk for the insurer. The second objective of this study is, taking into account the populations detected, to optimize the overall profitability of the portfolio by determining the optimal premium.

To achieve this, a profitability management indicator, the Expected Loss Ratio (ELR), will be used. This indicator will be studied and, through the *PRIM* algorithm, for Patient Rule Induction Method, segments (subsets of the client portfolio) on which the ELR strongly deviates from its average will be detected, whether the deviation is upward or downward.

Then, the customer behavior of the policyholders in our portfolio regarding the act of cancellation will be studied. This step will focus on the impact of the price on customer behavior using an indicator : elasticity.

Finally, a more global approach will be undertaken : this last step aims at optimizing the insurer's profit given its portfolio volume constraints. Mathematical optimization methods will be used where the influencing variable will be the commercial premium paid by the policyholder.

Keywords : Motor insurance, Generalized Linear Models (GLM), termination of an insurance contract, insurance portfolio, Expected Loss Ratio (ELR), PRIM (Patient Rule Induction Method) algorithm

Table des matières

Remerciements **1**

Résumé **2**

Abstract **4**

Table des matières **6**

Introduction **9**

1 Contexte et problématique de l'étude **11**
1.1 Le marché de l'assurance automobile en Europe 11
1.2 L'étude d'un portefeuille d'assurance 11
1.2.1 Enjeux 12
1.2.2 Les indicateurs de rentabilité 12
1.3 La décomposition de la prime commerciale d'assurance 13
1.3.1 La prime pure 13
1.3.2 Le passage de la prime pure à la prime commerciale 14
1.4 Les différents statuts d'un contrat 16
1.4.1 La transformation d'un devis 16
1.4.2 La résiliation d'un contrat 16

2 Mise en place d'un suivi de portefeuille **17**
2.1 La base de données 17
2.1.1 Les variables explicatives 18
2.1.2 Premières statistiques univariées 19
2.2 Analyse de la rentabilité du portefeuille 21
2.3 Analyse de la résiliation 25
2.4 Modélisation du taux de résiliation 26
2.4.1 Implémentation du modèle 29
2.4.2 Validation du modèle 30

2.4.3 Premiers résultats du modèle 34

3 Automatisation de la détection de segments de portefeuille non rentables 38

3.1 L'algorithme PRIM 38
3.1.1 Introduction 38
3.1.2 Principe et implémentation 38
3.1.3 Construction des boîtes individuelles 42
3.1.4 Hold-out validation 46
3.1.5 Trajectoire 46
3.1.6 Le modèle PRIM face aux modèles "classiques" 48
3.2 Analyse des boîtes obtenues 49
3.2.1 Diagrammes de fréquences relatives 49
3.2.2 Dissimilarités entre boîtes 49
3.3 Adaptation de l'algorithme à notre cas d'usage 50
3.4 Application de l'algorithme sur la base de données 51
3.4.1 Paramétrage des modèles implémentés 51
3.4.2 Résultats - Segments déviants sous-tarifés 52
3.4.3 Résultats - Segments déviants sur-tarifés 58
3.5 Conclusions 61

4 Modélisation de l'élasticité des taux de résiliation au prix 62

4.1 L'élasticité au prix 62
4.2 Modélisation de l'élasticité 63
4.2.1 Modèle GLM 65
4.2.2 Élasticité globale du portefeuille 65
4.2.3 Élasticité au prix selon les variables explicatives 66
4.3 Étude des élasticités des segments déviants identifiés par l'algorithme PRIM 67

5 Optimisation tarifaire 70

5.1 Objectifs de l'optimisation tarifaire 70
5.2 Présentation du problème 71
5.2.1 Programme d'optimisation 71
5.2.2 Adaptation du problème d'optimisation à notre cas d'usage . . 72
5.3 Le module Price Simulation du logiciel addactis® Pricing 73
5.3.1 Méthode *Simulation* 73
5.3.2 Méthode *Standard* 74
5.4 Cadre théorique de l'approche *Standard* du logiciel - Méthodes de résolution de programmes d'optimisation 75
5.4.1 Lagrangien 75

5.4.2 Conditions de qualification des contraintes 75
5.4.3 Conditions nécessaires d'optimalité de Karush, Kuhn et Tucker 76
5.4.4 Application au cadre quadratique 76
5.5 Résultats sur la base de données 79
5.5.1 Import des données et modèles 79
5.5.2 Définition des indicateurs 80
5.5.3 Méthode *Simulation* 80
5.5.4 Méthode *Standard* 84

6 Limites opérationnelles des données et modèles implémentés 89

Conclusion 92

Bibliographie 94

A Annexe : Le logiciel addactis® Pricing 95

B Annexe : Résultats du modèle de résiliation 97

C Annexe : Exemple d'utilisation de l'algorithme PRIM 99

D Annexe : Résultats du modèle d'élasticité 105

E Annexe : Passage du problème primal au problème dual 107

Liste des abréviations 110

Table des figures 111

Liste des tableaux 113

Introduction

Le monde de l'assurance se trouve actuellement confronté à des réglementations en constante évolution et à l'émergence de nouveaux acteurs compétitifs. Afin de conserver leurs parts de marché ou même d'en gagner, les compagnies d'assurances recherchent de nouveaux leviers de croissance. L'utilisation des données devient ainsi une source d'information et de valeur sur laquelle bon nombre d'entreprises du secteur se penchent.

Les données en assurance, du fait de leur caractère personnel sont un sujet sensible. De même que pour être valorisées, ces dernières doivent faire l'objet d'un important processus de traitement afin de s'assurer de leur qualité, exhaustivité et cohérence. Le portefeuille client d'une entreprise d'assurance est un exemple de données pouvant receler de précieuses informations.

L'objectif de cette étude a donc été de comprendre comment les données disponibles au sein d'un portefeuille d'assurance peuvent être utilisées afin d'optimiser la rentabilité d'une compagnie d'assurance.

Des outils permettant aux assureurs de mieux analyser leur portefeuille d'assurés afin d'en tirer un meilleur profit et d'adapter leurs stratégies ont ainsi été mis en place au sein de cette étude. Pour cela, plusieurs indicateurs peuvent être décisifs afin d'analyser un portefeuille et c'est notamment le cas de la rentabilité. Cette dernière peut être perçue comme le rapport entre ce que coûte le contrat à l'assureur et ce qu'il lui rapporte. Ainsi, lorsque ce rapport sera supérieur à un, le contrat sera déficitaire pour l'assureur et à l'inverse, un indicateur inférieur à un sera révélateur d'un contrat rentable. A l'aide d'un algorithme : l'algorithme *Patient Rule Induction Method*, abrégé en *PRIM*, nous détecterons des segments sur lesquels la rentabilité dévie, c'est-à-dire qu'elle s'éloigne fortement de la moyenne : elle peut s'éloigner à la hausse ou au contraire à la baisse.

Dans un second temps et parce qu'une attention tout aussi particulière peut être portée à la résiliation des contrats en portefeuille, nous étudierons le comportement

client en terme de résiliation. En effet, l'assureur doit s'assurer de rester compétitif par rapport au marché et, dans ce cadre, analyser les contrats résiliés peut s'avérer être un processus clé. Un modèle *GLM* sera implémenté.

A la suite de cela, nous nous intéresserons à l'impact d'une hausse de tarif sur la résiliation à l'échelle de l'assuré. Le concept d'élasticité sera ainsi introduit et approfondi sur certains segments du portefeuille tels que les segments déviants définis par l'algorithme *PRIM*.

Enfin, nous adopterons une approche plus globale en nous plaçant à l'échelle du portefeuille. Nous mettrons en place un programme d'optimisation théorique permettant de maximiser le gain de l'assureur tout en conservant ses parts de marché. Deux méthodes de résolution de ce programme seront proposées. Ces méthodes permettront de déterminer l'ajustement optimal des primes d'assurance.

1 Contexte et problématique de l'étude

1.1 Le marché de l'assurance automobile en Europe

L'assurance automobile en responsabilité civile est une assurance obligatoire dans la totalité des pays de l'Union Européenne. Ce type d'assurance est donc hautement stratégique : bien souvent, l'assurance automobile s'avère être l'entrée en relation d'un jeune client avec son assureur.

De nombreux acteurs sont présents sur le marché de l'assurance automobile. Les acteurs historiques, assureurs notamment, ont vu depuis quelques années la concurrence être exacerbée par de nouveaux entrants sur le marché tels que les bancassureurs ou plus récemment les assurtechs. Bien que la structure du marché européen soit très différente d'un pays à l'autre, notamment avec l'ajustement des canaux de distribution selon la culture et les besoins de chaque population, tous font donc face à une concurrence accrue. L'innovation est au coeur de l'assurance avec notamment la naissance de produits innovants et de plus en plus adaptés aux besoins des clients, basés sur des données nouvelles.

1.2 L'étude d'un portefeuille d'assurance

Selon l'article de Charpentier, Denuit et Elie [3], l'assurance repose sur l'idée que la mutualisation des risques entre des assurés est possible. C'est ainsi que la notion de portefeuille d'assurance prend son sens. Afin d'assurer la pérennité de l'organisme d'assurance, une étude fine et détaillée du portefeuille doit etre réalisée.

1.2.1 Enjeux

Dans le contexte actuel que l'on connaît, les entreprises d'assurance ont tout intérêt à suivre de près leur portefeuille de clients. Ce suivi peut en revanche s'avérer bien plus complexe qu'attendu, du fait notamment du très grand nombre de données que cela fait intervenir.

En effet, un suivi peut être mené sur de nombreux indicateurs tels que la rentabilité des contrats en portefeuille, le taux de transformation des devis complétés, le taux de résiliation des contrats en portefeuille ou même encore la rentabilité des offres commerciales.

Pour pouvoir tirer profit de son activité, il est donc nécessaire à l'entreprise d'assurance de mieux connaître son portefeuille, ses spécificités, profils et principaux changements. Pour répondre à ce besoin, de nombreux indicateurs clés et outils ont été développés afin de collecter régulièrement des données chiffrées. La périodicité de cette collecte peut être annuelle, pour, par exemple, connaître la rentabilité sur un an d'un nouveau produit, mais elle peut aussi être mensuelle afin d'étudier, par exemple, les impacts d'une offre promotionnelle. Des données chiffrées quotidiennes peuvent aussi être intéressantes pour un membre de l'équipe commerciale ou de direction qui souhaiterait suivre, en temps réel, son portefeuille.

De nombreuses autres analyses telles que celles des sinistres (taux d'incidence, coût moyen,...) peuvent aussi être menées, se focalisant alors sur les sinistres ayant été déclarés par les individus présents dans le portefeuille d'assurance.

1.2.2 Les indicateurs de rentabilité

De nombreux indicateurs peuvent être utilisés afin d'analyser la rentabilité de l'ensemble d'un portefeuille d'assurance, d'une sous-population précise ou bien même d'un assuré en particulier. Les deux principaux indicateurs de rentabilité sont les suivants :

— Le Loss Ratio aussi noté S/P est un indicateur clé qui représente le rapport entre le montant des sinistres et celui des primes acquises sur une même période.

$$Loss\ Ratio = \frac{Sinistres}{Primes\ acquises} \tag{1.1}$$

Il a ainsi vocation à mesurer la rentabilité. En effet, un contrat sera rentable dès lors que les primes acquises pour ce contrat sont suffisantes pour payer les sinistres survenus au titre de ce dernier. Cela signifie donc que le contrat sera rentable lorsque notre Loss Ratio sera inférieur à 1. A l'inverse, un Loss

Ratio supérieur à 1 sera le signe d'un contrat non rentable.
Une solution à cette problématique de rentabilité pourrait donc être d'augmenter les cotisations de tous les assurés. Mais cette réponse est loin d'être optimale puisque cette hausse des prix entraînerait une diminution plus ou moins considérable du volume de notre portefeuille. Le pilotage des contrats par le Loss Ratio doit donc être réalisé à une maille beaucoup plus fine et ainsi de façon segmentée. Ce dernier doit faire l'objet d'une constante surveillance.

— L'Expected Loss Ratio aussi noté ELR représente le ratio entre la prime pure prédite et la prime acquise. Or, la prime pure prédite se calcule comme l'espérance des sinistres futurs et la prime acquise n'est rien d'autre que la prime commerciale. L'Expected Loss Ratio représente donc également le ratio entre l'espérance des sinistres futurs et la prime commerciale.

$$ELR = \frac{Prime\ pure\ prédite}{Prime\ acquise} \tag{1.2}$$

Bien que le Loss Ratio soit un indicateur très facile à implémenter il présente certaines limites. En effet, c'est la sinistralité passée et donc observée qui est utilisée pour son calcul. Il ne sera donc pas possible de connaître la rentabilité future d'un contrat ou même d'une affaire nouvelle à l'aide de cet indicateur.

L'Expected Loss Ratio permet alors de palier certaines des limites du Loss Ratio. En effet, ce n'est pas la sinistralité observée qui est utilisée mais bien la sinistralité prédite. En estimant, par un modèle de prime, la sinistralité future d'un contrat en portefeuille ou d'une affaire nouvelle, nous obtiendrons donc un ELR pour ces contrats.

1.3 La décomposition de la prime commerciale d'assurance

La prime commerciale d'assurance représente la prime payée par l'assuré afin de transférer son risque à l'assureur. Cette prime dépend de paramètres techniques, commerciaux ainsi que réglementaires et peut donc être décomposée en plusieurs montants que nous allons voir dans cette partie.

1.3.1 La prime pure

La prime pure ou prime technique (PP) s'obtient à l'aide d'une modélisation. En effet, l'inversion du cycle de production en assurance ne permet pas de connaître à l'avance le coût d'un contrat pour l'assureur, il doit donc être estimé par un modèle.

L'assureur cherchera à estimer l'espérance de ses pertes ou en d'autres termes, l'espérance des sinistres futurs. Cette somme représentera le montant du sinistre moyen auquel devra faire face l'assureur pour le risque assuré.
Si on note S la somme des sinistres du portefeuille, on aura :

$$S = \sum_{i=1}^{N} X_i \tag{1.3}$$

où N représente le nombre de sinistres et X_i le coût du sinistre i. Lorsque l'on fait l'hypothèse selon laquelle les coûts des sinistres X_i sont indépendants et suivent une même loi et qu'on suppose également que le coût d'un sinistre est indépendant de la fréquence, on peut écrire :

$$\mathbb{E}(S) = \mathbb{E}\left(\sum_{i=1}^{N} X_i\right) \tag{1.4}$$

$$= \mathbb{E}\left(\mathbb{E}\left(\sum_{i=1}^{N} X_i | N\right)\right) \tag{1.5}$$

$$= \mathbb{E}\left(N\mathbb{E}(X)\right) \tag{1.6}$$

puisque les coûts de sinistres X_i sont indépendants et identiquement distribués. Enfin, grâce à l'indépendance de N et de X, on obtient :

$$PP = \mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(X) \tag{1.7}$$

où $\mathbb{E}(N)$ représente la fréquence de sinistre et $\mathbb{E}(X)$ le coût moyen.
La plupart du temps, ce sera donc un modèle de type $Fréquence \times Coût\ moyen$ qui sera implementé afin d'estimer la prime pure.

1.3.2 Le passage de la prime pure à la prime commerciale

Les chargements et taxes

Le Chargement de Sécurité (CS) est un montant permettant à l'assureur de faire face à une sur-sinistralité, c'est-à-dire à un écart entre la charge réelle des sinistres et la charge estimée. En effet, il existe une volatilité autour de la prime pure. Cette volatilité est d'une part liée aux incertitudes relatives à la modélisation et d'autre part à l'aléa de la sinistralité. Il est donc nécessaire pour l'assurance d'immobiliser un capital afin de pouvoir faire face à ce risque. Cette immobilisation a un coût qui sera donc ajouté à la prime pure. On obtient ainsi la prime de risque :

$$Prime\ de\ risque = PP + CS \tag{1.8}$$

Mais un assureur est aussi soumis à de nombreux frais, qu'ils soient directement liés aux sinistres de ses clients ou plus généraux :

— Il est tout d'abord tenu à la gestion des sinistres. Cette gestion entraîne des frais de suivi et de gestion qui doivent être ajoutés à la prime de risque.
— De plus et comme toute entreprise, l'organisme d'assurance est soumis à des coûts qui ne sont pas directement liés aux sinistres mais qui doivent être pris en compte dans la prime d'assurance. Ces coûts peuvent être liés à la gestion des contrats, à l'administration (services comptables, actuariat), au fonctionnement de la société ou encore à l'activité commerciale (commission des intermédiaires, publicité).

L'ensemble de ces coûts sont appelés des chargements pour frais (CH) et s'ajoutent à la prime de risque.

Enfin, les cotisations d'assurance ne sont pas soumises à la TVA mais elles incluent en revanche une taxe fiscale, réglementée par le code des assurances, et dont le taux légal varie selon la nature des contrats d'assurance. Les sociétés d'assurances reversent alors les montants des taxes collectées au Trésor Public. La cotisation d'assurance peut aussi inclure une contribution reversée à divers fonds tels que le fonds de garantie des victimes des actes de terrorisme et d'autres infractions (FGTI) ou encore le fonds de garantie des assurances obligatoires de dommages (FGAO) par exemple. Ces coûts sont appelés des taxes et sont eux aussi additionnés à la prime pure.

La prime commerciale

Une fois que l'on a défini la prime pure ainsi que les différents frais inhérents à l'activité de l'entreprise, au contexte fiscal ainsi qu'au risque de sur-sinistralité, il est possible de déterminer la prime commerciale de chaque assuré. Cette prime commerciale (PC) s'écrit alors :

$$PC = PP + CH + T + CS \tag{1.9}$$

Il est à noter que les frais ne sont pas fixes et dépendent du profil de risque de l'assuré. Il sera donc intéressant de voir quelles peuvent être les stratégies mises en place par les organismes d'assurance quant au passage de la prime technique à la prime commerciale pour les différents profils d'assurés.

1.4 Les différents statuts d'un contrat

La vie d'un contrat d'assurance implique le passage par différents statuts prédéfinis tels que ceux d'affaire nouvelle, de contrat en portefeuille ou encore de contrat résilié. Nous allons ici nous intéresser à deux d'entre eux.

1.4.1 La transformation d'un devis

Dans un marché concurrentiel comme le nôtre, le nombre de devis effectués est croissant et cet effet n'est qu'amplifié par la possibilité de réaliser des devis sur internet. En effet, de nombreux prospects sont en quête du meilleur tarif d'assurance, leur conférant les meilleures garanties au plus bas prix. Chacun d'eux est donc prêt à réaliser de nombreux devis auprès d'organismes différents.

La transformation sera caractérisée par l'acceptation d'un devis. On considèrera qu'un prospect qui réalise un devis devient véritablement client lorsqu'il paye sa première échéance et donc que sa garantie prend effet. C'est à partir de ce moment là que le prospect est en affaire nouvelle.

L'information contenue par la transformation ou non de chaque devis peut donc se révéler être un atout majeur pour mieux comprendre son portefeuille et optimiser son tarif d'assurance.

1.4.2 La résiliation d'un contrat

Plusieurs études sur les différentes populations européennes montrent que l'amélioration de l'accès à l'information couplée au développement des canaux numériques permettent aux assurés de comparer et ainsi de changer plus aisément de fournisseur d'assurance.

De ce fait, il devient primordial pour les assureurs de conserver leurs clients et pour cela, l'information des profils les plus à même de résilier est capitale.

En revanche, on s'intéressera ici uniquement aux contrats résiliés par l'assuré et non pas ceux dont la résiliation est demandée par l'assureur pour non-paiement des cotisations par exemple.

2 Mise en place d'un suivi de portefeuille

Pour suivre un portefeuille d'assurés, l'approche graphique peut constituer une première étape. En effet, l'émergence du Business Intelligence (BI) ou Informatique Décisionnelle en français, qui désigne l'ensemble des technologies permettant aux entreprises d'analyser leurs données dans le but de faciliter les prises de décisions a instauré une nouvelle ère dans laquelle la donnée fait partie intégrante du quotidien des décisionnaires. Des tableaux de synthèse, aussi appelés dashboards, optimisent les processus internes gagnant ainsi en efficacité.

Dans ce cadre, l'attention a été tout particulièrement portée sur les indicateurs de résiliation et sur l'Expected Loss Ratio du portefeuille utilisé.

2.1 La base de données

Les données d'un portefeuille d'assurance présentent un caractère hautement sensible pour plusieurs raisons. En effet, la CNIL (Commission nationale de l'informatique et des libertés) définit les données personnelles comme : « toute information se rapportant à une personne physique identifiée ou identifiable ». Pour préciser ce propos, toute personne peut être identifiée : directement (avec son nom et prénom par exemple) ou indirectement (par un identifiant (numéro client), un numéro de téléphone, une donnée biométrique ou à l'aide de plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale par exemple). Pour encadrer les pratiques autour des données personnelles, la réglementation RGPD, qui signifie « Règlement Général sur la Protection des Données », a été mise en place, encadrant le traitement des données personnelles sur le territoire de l'Union Européenne. Bien évidemment, l'ensemble des entreprises

du secteur de l'assurance est soumis à cette réglementation.
De telles données peuvent également représenter un risque en cas d'exposition auprès d'un concurrent ou du grand public qui pourraient détenir des informations importantes quant aux enjeux financiers ou commerciaux liés au portefeuille d'assurance.

C'est dans ce cadre que, malgré la forte demande d'étude des données d'assurance, il est très peu commun que les assureurs partagent ces données en externe. Ainsi, pour mener à bien ce mémoire et pouvoir lui attacher une illustration concrète, une base de données a dû être trouvée au sein de l' « Open Data » ou, autrement dit, des données dont l'accès et l'usage sont laissés libres.

La base de données utilisée est une base en libre accès puisqu'elle fait partie de la librairie *R* intitulée *CASDatasets* [5]. Celle-ci, intitulée *eudirectlapse* est une base de données sur laquelle nous n'avons que des informations partielles puisque nous ne connaissons pas l'année ni le pays à laquelle elle correspond (il s'agit d'un pays appartenant à l'Union Européene). En revanche, ces données se basent sur une année et sont issues d'un portefeuille automobile comportant 23 060 polices d'assurance.

2.1.1 Les variables explicatives

Dix-neuf variables sont présentes au sein de cette base :

- Certaines sont catégorielles : Le genre de l'assuré, l'évolution de son coefficient bonus-malus (ou CRM : Clause de Réduction-Majoration), la différence de statut entre l'assuré et le conducteur, le secteur d'activité de l'assuré (médical ou non), l'usage de la voiture (professionnel ou non), le moyen de garage du véhicule, sa puissance (en kW) et la région de vie de l'assuré.
- D'autres sont numériques : L'âge de l'assuré, l'âge de la police d'assurance, le nombre de contrats d'assurance souscrits par l'assuré, la prime proposée à l'assuré pour l'année suivante, la prime que l'assuré payait tout au long de l'année, la prime actuarielle ou prime pure, la prime du marché ainsi que la fréquence de paiement, l'âge du véhicule ainsi que ce même âge à l'achat du véhicule.
- Et enfin, une variable est booléenne : La variable *Lapse* qui vaut 1 si l'assuré a résilié sa police d'assurance et 0 sinon.

En revanche, de nombreuses autres informations auraient également pu nous être utiles si celles-ci avaient été renseignées au sein de la base de données :

- Le risque couvert : responsabilité civile, bris de glace, vol, incendie, etc. Cette donnée aurait permis d'obtenir d'importantes informations dans le but de mettre en perspective les résultats des modèles implémentés dans la suite de ce mémoire. En effet, que ce soit la sinistralité ou encore le comportement

d'un client, tous deux diffèrent selon le risque assuré ainsi que l'obligation légale ou non d'assurer ce risque.
— La sinistralité réelle de l'assuré aurait permis de définir le Loss Ratio de chacun des contrats, indicateur de rentabilité clé.
— La sinistralité passée : la vision de la sinistralité sur une année n'apporte qu'une information partielle sur la rentabilité d'un contrat. Connaître les antécédents en termes de sinistralité du contrat sur plusieurs années permet de mieux apprécier le risque supporté par l'assureur. Il apparaîtrait en effet peu judicieux de majorer un client n'ayant eu qu'un seul sinistre au cours des dix dernières années.
— L'historique des majorations ou minorations appliquées à chaque contrat.
— La temporalité du contrat : sa date d'effet ainsi que la date de résiliation du contrat.

2.1.2 Premières statistiques univariées

Avant d'analyser les indicateurs utiles à la suite de notre étude et d'élaborer nos modèles, il est primordial de mieux connaître la population observée afin de détecter de premières informations importantes ou tendances.

La première figure (figure 2.1) présentée ci-dessous explicite le profil moyen présent au sein de notre portefeuille :

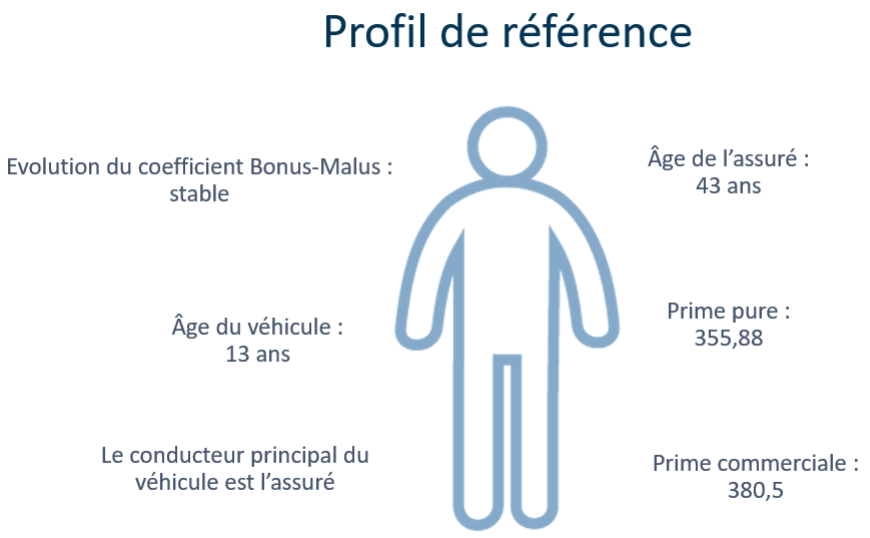

FIGURE 2.1 – Profil de référence du portefeuille étudié

Elle nous indique, qu'en moyenne, les individus du portefeuille ont 43 ans et possèdent un véhicule âgé de 13 ans. Toujours en moyenne, leur cotisation s'élève à 380 (nous ne connaissons pas la monnaie en question). La majorité des assurés présente une évolution du coefficient bonus-malus stable et ces assurés sont eux-mêmes les conducteurs principaux du véhicule assuré.

Si l'on s'intéresse plus précisément à la répartition des âges des individus au sein de notre portefeuille (figure 2.2), on remarque que ces assurés ont entre 20 et 85 ans et que la grande majorité de ces derniers présente un âge compris entre 25 et 60 ans.

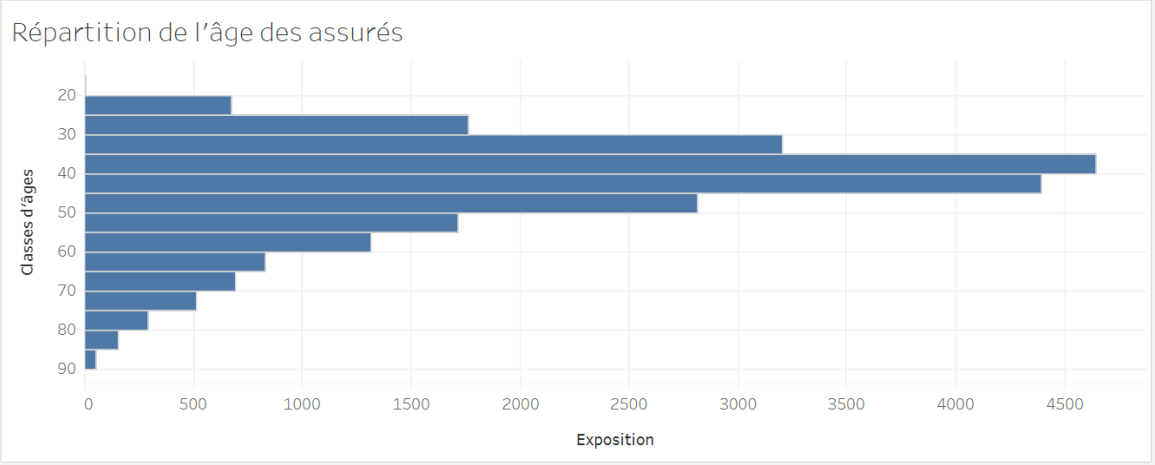

FIGURE 2.2 – Distribution de l'âge des assurés du portefeuille

Comme nous le savons également, la personne désignée comme conducteur principal joue un rôle important quant à la tarification du contrat d'assurance. En effet, bien qu'il n'existe pas de définition légale de ce qu'est un jeune conducteur, que ce soit dans le Code de la route ou dans le Code des assurances, il est généralement considéré qu'il s'agit des automobilistes titulaires du permis de conduire depuis moins de 3 ans, soit la durée du permis de conduire probatoire. Pour ces jeunes conducteurs, porteurs d'un risque accru du fait de leur manque d'expérience, l'assurance automobile peut s'avérer coûteuse. Pour réduire les frais, le jeune conducteur peut par exemple s'assurer en tant que conducteur secondaire sur le contrat d'assurance de ses parents.
Nous remarquons (figure 2.3) qu'au sein du portefeuille à notre disposition, quatre modalités sont prépondérantes : la modalité « same » qui signifie que le conducteur principal est l'assuré lui-même, la modalité « only partner » qui signifie que c'est cette fois-ci le conjoint. Suivent les modalités « young drivers » : c'est un jeune conducteur puis la modalité « all drivers > 24 » : un conducteur âgé de plus de 24

ans. Les modalités « learner 17 » : conducteur apprenti, « commercial » et « unknown » sont clairement sous-représentées et feront l'objet de regroupements pour les modélisations futures.

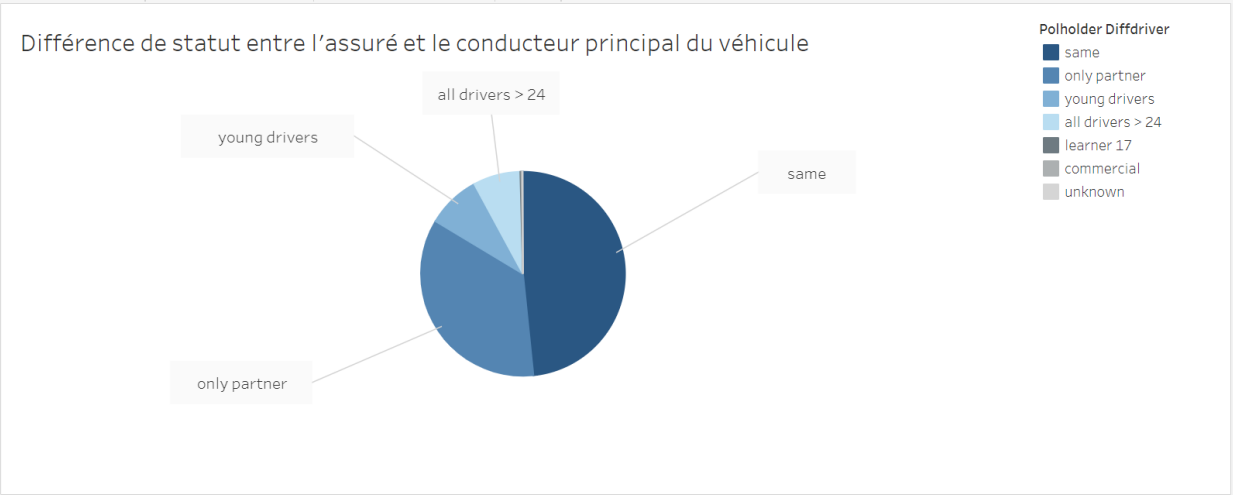

FIGURE 2.3 – Différence de statut entre l'assuré et le conducteur principal du véhicule

2.2 Analyse de la rentabilité du portefeuille

L'analyse de la rentabilité du portefeuille d'assurés se base sur un indicateur : l'ELR. Cet ELR est défini, avec notre jeu de données, par :

$$ELR = \frac{Prem_Pure}{Prem_Last} \tag{2.1}$$

qui revient à définir le ratio entre la prime pure et donc actuarielle et la prime commerciale demandée à l'assuré.

Une première étude globale du portefeuille indique un ELR global proche de 93,53%. Ce dernier se distingue de l'ELR moyen, proche de 95,99%. Cet indicateur étant dans les deux cas inférieur à 100%, on remarque donc que le portefeuille est globalement rentable puisque les primes reçues par l'assureur permettront de couvrir les sinistres.

Une analyse variable par variable peut alors être menée et, dans un premier temps, une analyse graphique permet de mieux comprendre les données à notre disposition. Pour cela, un outil BI, le logiciel "Tableau" a été utilisé permettant une

représentation facilitée des KPI (Key Performance Indicators).

Nous allons dans un premier temps nous intéresser à l'ELR en fonction du coefficient bonus-malus de l'assuré :

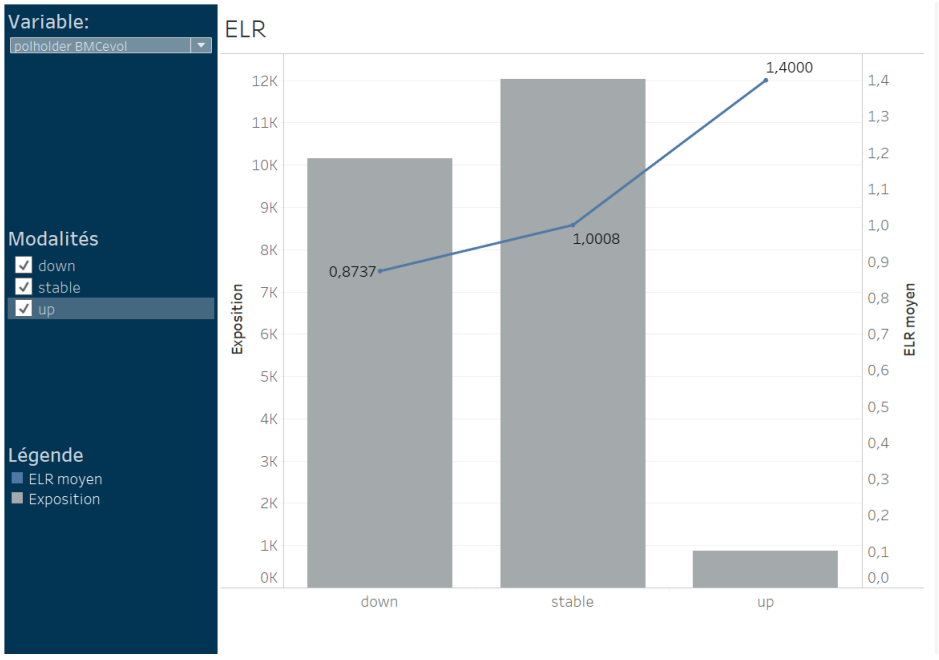

FIGURE 2.4 – Analyse de l'ELR en fonction de l'évolution du CRM

On remarque ici une très nette différence d'ELR selon l'évolution du CRM. En effet, lorsque ce coefficient est en hausse, l'Expected Loss Ratio est en moyenne, pour l'ensemble de ces contrats, égal à 1,4. Cela signifie que ces contrats sont globalement sous-tarifés et donc déficitaires : les primes acquises ne permettent pas de régler les sinistres prédits pour ces contrats. A l'inverse, les contrats pour lesquels le CRM est en baisse ont un ELR moyen proche de 0,87. Ces contrats sont sur-tarifés par rapport à leur sinistralité. On remarque aussi que leur exposition cumulée est bien plus faible : il semblerait donc que ces contrats ne restent pas en portefeuille ou ne transforment pas leur devis. Enfin, pour les individus dont le CRM est stable, on remarque un équilibre entre prime pure et prime acquise.

Ces résultats apparaissent ici cohérents. En effet, lorsque le CRM d'un individu est en hausse, cela signifie que ce dernier est responsable d'un accident. Ce sinistre a donc très certainement entraîné un coût pour l'assureur. La faible exposition peut être le signe d'une hausse tarifaire brutale visant à faire résilier l'assuré sinistré. A contrario, un CRM en baisse est le signe d'un conducteur non sinistré et donc qui s'avère être un bon risque pour l'assureur.

Nous pouvons maintenant nous intéresser à l'ELR en fonction de l'âge de la police d'assurance :

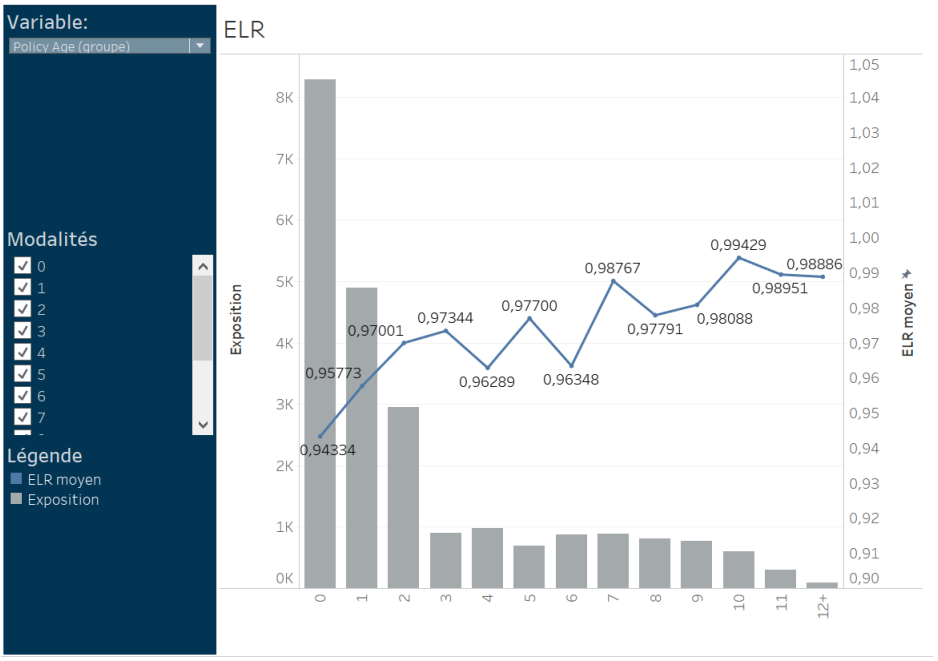

FIGURE 2.5 – Analyse de l'ELR en fonction de l'âge de la police d'assurance

On remarque aisément une tendance croissante de l'ELR. Ceci est révélateur d'une stratégie souvent en vigueur visant à attirer de nouveaux clients en leur proposant un tarif attractif impactant immédiatement l'ELR puisque la marge entre prime pure et prime commerciale est fortement réduite.

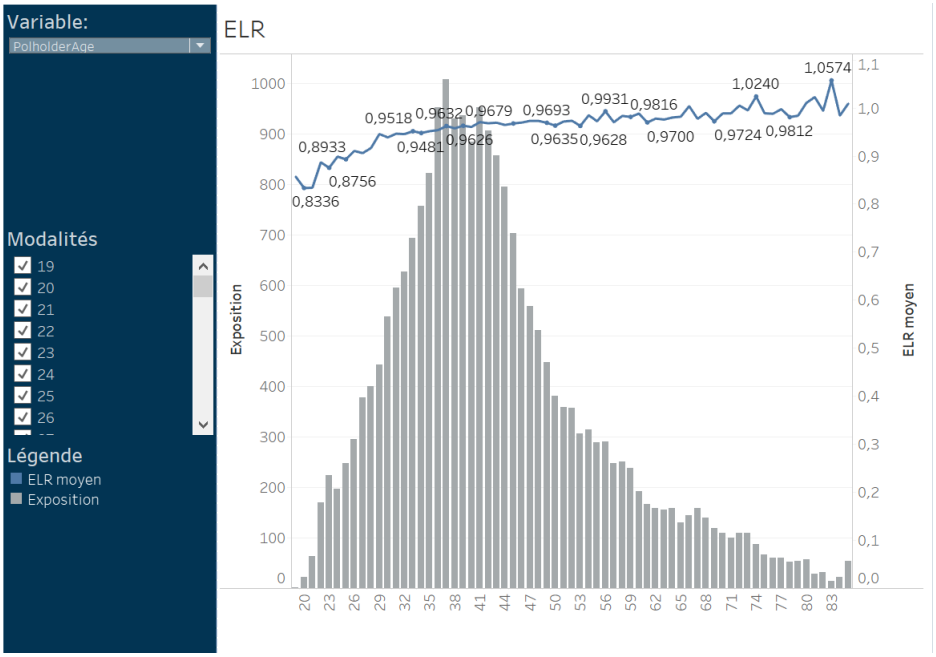

FIGURE 2.6 – Analyse de l'ELR en fonction de l'âge de l'assuré

Enfin, l'ELR en fonction de l'âge de l'assuré permet à nouveau de percevoir une tendance croissante : globalement, plus l'assuré vieillit, plus son ELR augmente et donc son contrat est sous-tarifé. On remarque donc un prix très attractif chez les seniors contrairement aux jeunes assurés.

Limites de l'ELR :

Il est à noter que l'Expected Loss Ratio est un indicateur qui présente des limites dans le cadre de ce mémoire. En effet et comme nous l'avons vu dans le paragraphe 1.3.2, le passage de la prime pure à la prime commerciale dépend également des taxes et chargements appliqués à chaque contrat ainsi que de la marge souhaitée par l'assureur. Les taux de taxes ou de chargements peuvent différer selon la nature du risque assuré ou encore selon l'ancienneté du contrat par exemple. Ils impacteront directement l'ELR. Quant à la marge, elle peut dépendre de la période et du contexte dans lequel est souscrit le contrat mais également de la stratégie commerciale : si l'assureur cherche à attirer une population qu'il considère rentable, il peut tenter de réduire sa marge pour être compétitif face à ses concurrents.
Une seconde limite réside dans la construction même de cet indicateur : la prime pure est une estimation, elle provient d'un modèle. Un contrat dont l'ELR sera supérieur à 1 sera donc probablement déficitaire au sens du modèle mais on ne peut en avoir la certitude. Il faudrait faire l'hypothèse très forte selon laquelle les primes

pures sont parfaitement calibrées pour affirmer que le contrat sera déficitaire. Mais cela peut également signifier que la cotisation d'assurance d'un tel contrat a évolué depuis sa tarification initiale. Cette évolution peut être liée à de nombreuses raisons : un coefficient bonus-malus en baisse, une offre commerciale ou remise à l'initiative du distributeur qui connaît son client et son risque. Cette diminution de la prime commerciale d'assurance qui pourra devenir, au fil des années, inférieure à la prime pure ne signifie par pour autant que le contrat est ou sera déficitaire.

2.3 Analyse de la résiliation

Ensuite, un processus similaire peut être réalisé sur les données de résiliation. Cette fois encore, l'analyse graphique a été effectuée grâce à l'application "Tableau".

On peut dans un premier temps s'intéresser à la résiliation selon l'évolution du CRM de l'assuré :

FIGURE 2.7 – Analyse de la résiliation en fonction de l'évolution du CRM de l'assuré

A travers cette figure, on remarque que la variable indiquant l'évolution du CRM de l'assuré a un impact fort sur la résiliation. En effet, un assuré dont le CRM est en baisse résiliera en moyenne son contrat auto avec deux fois plus de chance qu'un assuré ayant un CRM en hausse (16% de chance de résilier contre 9%). Ceci a en

effet du sens : les tarifs d'assurance sont rarement revus à la baisse et une baisse du CRM est un élément permettant de faire jouer la concurrence pour trouver un tarif moins cher ailleurs. En revanche, un assuré ayant commis un sinistre responsable se verra proposer un tarif d'assurance très élevé chez un assureur concurrent.

Il peut aussi apparaître pertinent d'analyser la résiliation des assurés en fonction de leur âge :

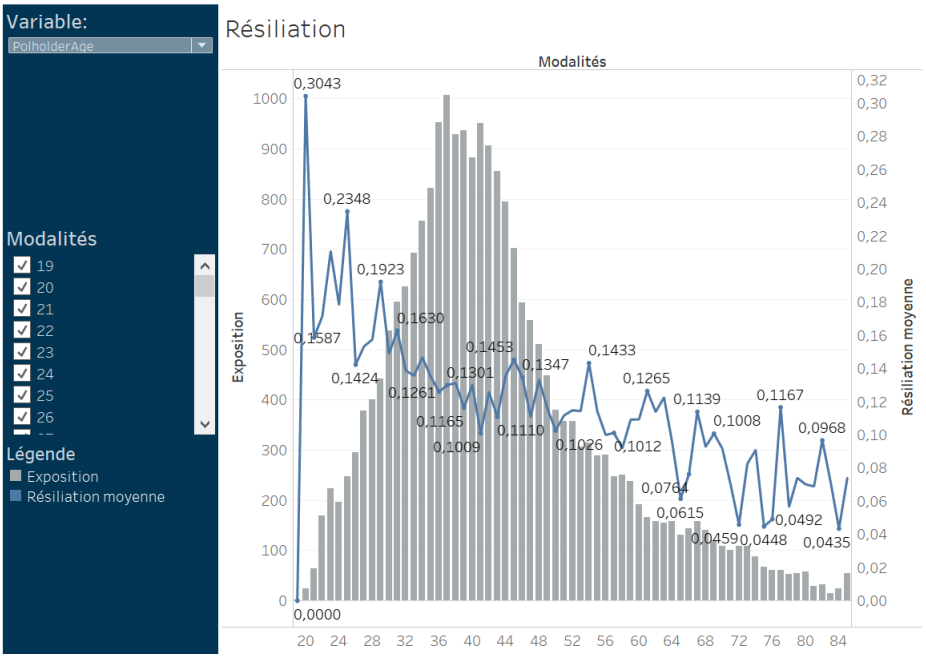

FIGURE 2.8 – Analyse de la résiliation en fonction de l'âge de l'assuré

On remarque ici une très nette tendance à la baisse : plus un assuré vieillit, moins il a tendance à résilier son contrat d'assurance. Plusieurs facteurs peuvent l'expliquer : une relation de confiance avec son assureur, un moins bon accès aux comparateurs d'assurance...

2.4 Modélisation du taux de résiliation

Dans cette partie, l'objectif est de modéliser le taux de résiliation des assurés. Plusieurs variables peuvent permettre de définir cet indicateur. Au sein de la base de données « eudirectlapse », il s'agit d'une variable booléenne : la variable "Lapse" qui vaut un si l'assuré a résilié son contrat et zéro si ce n'est pas le cas. Mais, lorsque

la date de résiliation du contrat est disponible, des modèles de durée peuvent également être implémentés, apportant une information supplémentaire.
Le modèle de Cox est l'un de ces modèles. Il permet de modéliser la durée de vie d'un contrat d'assurance à partir de la date d'effet de ce dernier, connue. Le modèle des risques proportionnels de Cox, n'impose aucune forme spécifique à la fonction de survie, permettant de modéliser les données censurées de façon très flexible. C'est en effet nécessaire puisque les données de résiliation sont censurées à droite : la date d'entrée en vigueur du contrat est toujours connue mais la date de résiliation ne l'est pas toujours.

Ainsi, le modèle de Cox permet de tirer profit de l'information des temps de survie et de censure tandis que le modèle logistique, ne tient compte que d'une variable binaire. En analyse de survie et donc avec le modèle de Cox, le résultat du modèle est un délai avant manifestation (de la résiliation dans notre cas) tandis que pour la régression logistique, la sortie du modèle est une variable binaire qui indique la survenance ou non de l'acte de résiliation (ou sa probabilité).
Il est également à noter que des méthodes alternatives au modèle de Cox existent : losque l'hypothèse de proportionnalité des hasards n'est pas vérifiée, des modèles alternatifs peuvent être utilisés tels que les modèles stratifiés (lorsque l'hypothèse de hasard proportionnel n'est vérifiée que par strates), les modèles à coefficients dépendants du temps ou encore le modèle additif de Aalen.

Dans le cadre des données que nous avons à notre disposition, seule la régression par modèle GLM est réalisable et sera implémentée au sein de ce mémoire. Pour cela, un modèle basé sur la variable "Lapse" qui est une variable booléenne valant un si l'assuré a résilié son contrat et zéro si ce n'est pas le cas, a été implémenté. Ce modèle est donc un modèle appliqué à un indicateur de résiliation reflétant le comportement des assurés. De nombreux types de modèles de regression existent et peuvent être utilisés à l'image des modèles GLM mais aussi CART, Random Forest et XGBoost. En effet et contrairement à la tarification, l'aspect réglementaire est ici nettement moins présent. Dans notre cas, ce sont des modèles GLM qui ont été implémentés permettant une meilleure interprétabilité.

Un modèle de régression logistique sera alors utilisé. Ce modèle est en réalité un modèle linéaire généralisé (GLM) pour lequel la fonction de lien est une fonction logit. En effet, la variable à modéliser est la probabilité de résiliation de notre portefeuille d'assurance et les observations ne peuvent prendre que deux valeurs : 0 ou 1. La fonction de lien logit pour un GLM est donc particulièrement adaptée dans notre cas puisqu'elle permet de conserver notre intervalle de valeurs $[0, 1]$.

La fonction logit est définie par :

$$logit(\pi) = ln\left(\frac{\pi}{1-\pi}\right) \tag{2.2}$$

On pourra dans la suite écrire la probabilité de résiliation de l'assuré i par :

$$\pi_i = \mathbb{P}(Y_i = 1|X_1 = x_1, ..., X_n = x_n) \tag{2.3}$$

où les variables $X_1, ..., X_n$ sont les variables explicatives du modèle et x_j la modalité de la variable X_j prise par l'assuré i.

On rappelle que classiquement, un modèle linéaire généralisé s'écrit :

$$g(\mathbb{E}(Y|X_1, ..., X_n)) = X\beta \tag{2.4}$$

avec :
- Y la variable à expliquer qui admet une distribution issue de la famille exponentielle ;
- $X_1, ..., X_n$ les variables explicatives ;
- g la fonction de lien ;
- β le paramètre de régression à estimer.

Ainsi, en appliquant la fonction de lien logit définie précédemment à cette expression, on obtient :

$$ln\left(\frac{\pi}{1-\pi}\right) = X\beta \tag{2.5}$$

ce qui nous permet finalement d'obtenir :

$$\pi = \frac{\exp(X\beta)}{1+\exp(X\beta)} \tag{2.6}$$

<u>Modèle retenu :</u>

Le modèle retenu a été implémenté à l'aide du logiciel addactis® Pricing et optimisé à partir de la procédure Stepwise. Cette procédure allie les méthodes forward et backward :
- Dans la méthode forward, on débute avec un modèle sans variable et l'algorithme rajoute successivement et une à une les variables explicatives significatives tout en vérifiant à chaque étape que la règle d'arrêt n'est pas vérifiée.

— A l'inverse, dans la méthode backward, l'algorithme est initialisé avec un modèle incluant toutes les variables explicatives disponibles puis retire, les unes après les autres, les variables non significatives.

La méthode Stepwise va ainsi tout d'abord procéder à une méthode forward lui permettant d'inclure des facteurs significatifs (la significativité est jugée à l'aide d'un seuil choisi à 5% et d'un test de type III sur chacun des facteurs disponibles). Une fois la procédure forward réalisée, une procédure backward est implémentée cherchant alors à retirer successivement les potentiels facteurs les moins significatifs du modèle en cours, la significativité étant à nouveau jugée à l'aide d'un seuil (choisi à 5% et d'un test de type III également).

2.4.1 Implémentation du modèle

Les variables explicatives retenues pour le modèle final sont alors :

— L'âge de l'assuré : *polholder_age*
— La différence de statut entre l'assuré et le conducteur : *polholder_diffdriver*
— Le genre de l'assuré : *polholder_gender*
— Son travail (milieu médical ou non) : *polholder_job*
— L'ancienneté du contrat : *policy_age*
— L'âge du véhicule : *vehicl_age*
— L'âge du véhicule à l'achat : *vehicl_agepurchase*
— Le type de garage du véhicule : *vehicl_garage*
— La région de vie de l'assuré :*vehicl_region*
— La prime commerciale (ou cotisation) payée par l'assuré : *prem_last*

Dans ce modèle, des regroupements de modalités ont été effectués :

— La différence de statut entre l'assuré et le conducteur :
  — Les jeunes conducteurs ont été regroupés avec les apprentis conducteurs
  — Les individus pour lesquels nous n'avons pas d'information sont regroupés avec les individus qui sont à la fois assuré et conducteur (modalité la plus représentée)
— Le type de garage du véhicule : les garages privés sont regroupés avec les abris de voiture et les individus se garant dans leur résidence privée.
— Certaines régions ont été regroupées en tenant compte de leur attitude face au risque (détails disponibles en annexe B). En effet, la base de données ne permet pas de regrouper les régions par similarité géographique ou démographique, le choix des regroupements peut donc uniquement se baser sur leur comportement face à la résiliation.

Ceci a notamment permis de regrouper des modalités sous-représentées ou encore

non-significatives.

Les résultats de ce modèle ainsi que les regroupements effectués sont détaillés et disponibles en annexe B.

2.4.2 Validation du modèle

Une attention particulière a été apportée à la validation de ce modèle ainsi qu'aux différents critères permettant de valider la qualité prédictive de ce dernier. Ils sont présentés dans cette partie.

Significativité des paramètres :

A travers les tableaux de résultats disponibles en Annexe (figure B.1), on peut remarquer que la très grande majorité des modalités utilisées dans ce modèle sont significatives. Plusieurs indicateurs nous permettent en effet d'en juger :

— En premier lieu, les colonnes "Wald Chi-2" et "Pr > Chi-2" nous informent sur la significativité des coefficients β. Pour cela, la valeur de la colonne "Wald Chi-2" correspond à $\left(\frac{Value}{Standarderror}\right)^2$. Ensuite, la colonne "Pr > Chi-2" indique la probabilité qu'une variable aléatoire suivant une loi du Khi-2 à 1 degré de liberté soit supérieure à la valeur "Wald Chi-2". Ainsi, si, pour un paramètre β, le résultat "Pr > Chi-2" est proche de 0, on peut juger le paramètre significatif (différent de 0) dans le modèle à l'étude.

— Un second outil est à notre disposition dans le logiciel addactis® Pricing : l'intervalle de confiance représenté par les colonnes "Lower conf. limit" ainsi que "Upper conf. limit". Si l'intervalle de confiance pour un paramètre β ne contient pas 0, on peut de même juger le paramètre significatif.

Significativité globale des variables du modèle :

Afin de vérifier la significativité globale des variables, un test de type III a été mis en place. Les résultats sont les suivants (figure 2.9) :

Graph | Results | Type III | Performance graphs | Stats | Exclusions | Consistency

Criteria	Degrees of freedom	Chi-2	Pr > Chi-2
vehicl_region	7	76,7446	0,0000
policy_age	1	40,1002	0,0000
polholder_age	3	34,4959	0,0000
vehicl_agepurchase	2	29,2947	0,0000
polholder_diffdriver	3	24,7525	0,0000
polholder_gender	1	15,4065	0,0001
prem_last	1	12,6058	0,0004
vehicl_garage	3	17,6585	0,0005
vehicl_age	1	9,6571	0,0019
polholder_job	1	4,6823	0,0305

FIGURE 2.9 – Test de type III appliqué aux variables du modèle de résiliation

A nouveau, c'est le test statistique du Chi-2 qui est implémenté. Les faibles valeurs présentes dans la colonne "Pr > Chi-2" indiquent donc que la totalité des variables du modèle sont globalement significatives et apportent de l'information (ces valeurs sont toutes inférieures à 5%). Ceci étant en effet prévisible de par la construction même du modèle. En effet, c'est une procédure Stepwise qui a permis de sélectionner les variables se basant ainsi sur la significativité globale.

Qualité d'ajustement du modèle :

Le graphique intitulé "Lift" dont dispose le logiciel, permet de juger de l'adéquation du modèle aux données sur différents intervalles de valeurs prédites. Pour obtenir ce graphique, des intervalles de prédictions (rangées préalablement par ordre croissant) sont créés et deux courbes sont représentées : "Estimated" et "Observed". Elles correspondent respectivement à la moyenne pondérée des valeurs prédites par le modèle et à la moyenne pondérée des valeurs observées. Ce graphique permet donc de juger, sur chaque intervalle, de la qualité d'ajustement du modèle aux données. Plus la courbe "Estimated" se rapproche de la courbe "Observed", meilleur sera le modèle.

On remarque, à l'aide de la figure ci-dessous, (figure 2.10) que les deux courbes sont globalement très proches l'une de l'autre. Le modèle implémenté présente ainsi une bonne qualité d'ajustement.

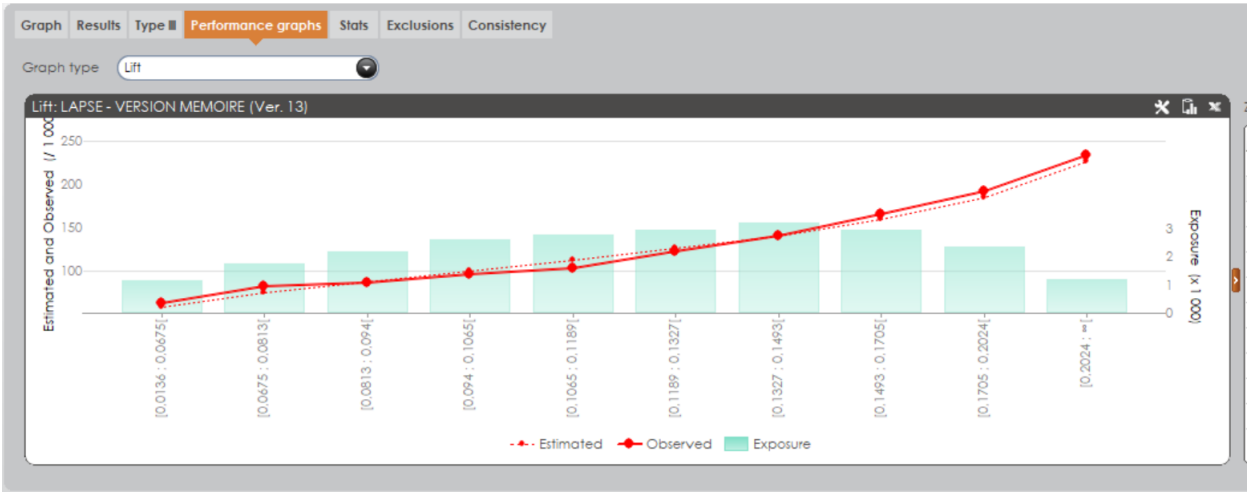

FIGURE 2.10 – Graphique Lift

ROC Curve :

Dans un modèle de propension sur données binaires, le critère de déviance, usuellement utilisé pour valider un modèle (résidus de déviance), n'est pas pertinent et ne nous permet pas d'apprécier le bon ajustement du modèle. Nous allons alors définir la ROC Curve, indicateur qui permettra de valider la qualité d'un modèle ainsi que de comparer plusieurs modèles entre eux.

La ROC (Receiver Operating Characteristic) Curve se définit par plusieurs étapes :

1. Pour chaque observation, le modèle GLM estime sa probabilité de résiliation comprise entre 0 et 1. Un seuil doit alors être défini dans l'intervalle $]0, 1[$. Cela nous permet d'effectuer une prédiction : si la réponse estimée par le modèle est supérieure au seuil, on la décide égale à 1, sinon, à 0.
2. A la suite de cela, nous allons pouvoir définir plusieurs indicateurs :
   — La sensibilité ou le Taux de Vrais Positifs (TVP). Il est souhaité proche de 1 et est défini par :
   $$TVP = \frac{VP}{VP + FN}$$
   où VP sont les vrais positifs et FN sont les faux négatifs.
   — La spécificité ou le Taux de Vrais Négatifs (TVN) défini par :
   $$TVN = \frac{VN}{VN + FP}$$
   dans ce cas également, un TVN proche de 1 signifiera que le modèle prédit bien les non résiliés.

Ces scores se définissent notamment à partir de la matrice de confusion suivante (2.11) :

	Valeur prédite	
Valeur observée	1	0
1	Vrais Positifs	Faux Négatifs
0	Faux Positifs	Vrais Négatifs

FIGURE 2.11 – Matrice de confusion

3. On représente sur un graphique le point d'abscisse un moins la spécificité et d'ordonnée la sensibilité ;

4. On réitère la procédure en faisant varier la valeur du seuil dans l'intervalle $]0, 1[$.

On remarque ainsi que si le modèle attribue aléatoirement des valeurs $\{0, 1\}$ (équiprobabilité), il y aura équirépartition dans la matrice de confusion et la sensibilité sera égale à la spécificité pour tous les seuils définis. La courbe ROC sera donc confondue avec la courbe $y = x$. A l'inverse, si la prédiction du modèle est parfaite pour toute valeur du seuil, le taux de vrai positif ou encore sensibilité sera toujours égale à 1 : la courbe ROC sera représentée par le carré supérieur et passera par les points $(0; 1)$ et $(1; 1)$. Ainsi, plus la courbe d'un modèle se situe au-dessus de la bissectrice, meilleur est le modèle.

Puisque la courbe ROC a aussi pour vocation de comparer plusieurs modèles, un indicateur chiffré a été défini afin de permettre une comparaison plus aisée. En effet, la courbe ROC d'un modèle peut être située en dessous de la courbe d'un second modèle dans un intervalle mais au dessus dans un autre intervalle : il n'y a pas de relation d'ordre dans ce cas. Nous allons ainsi calculer l'aire sous la courbe ROC : c'est l'indicateur AUC (Area Under Curve). Par transposition avec l'interprétation des courbes, plus l'aire sous la courbe (AUC) est élevée, meilleur est le modèle.

Nous allons donc maintenant nous intéresser à la courbe ROC ainsi qu'à la valeur de l'AUC de notre modèle de résiliation. Les résultats sont les suivants (figure 2.12), disponibles au sein du logiciel addactis® Pricing :

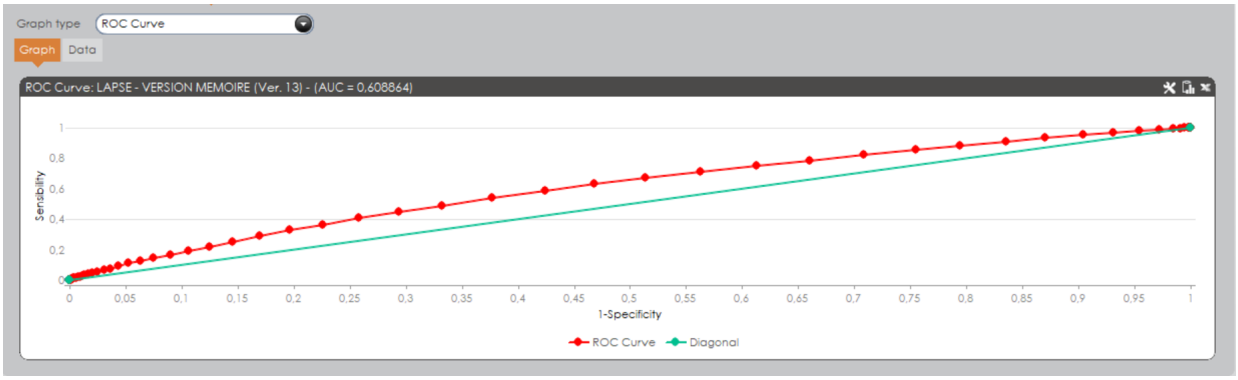

FIGURE 2.12 – Courbe ROC du modèle de résliation implémenté

L'AUC de ce modèle vaut 0,61. C'est, parmi l'ensemble des modèles implémentés, le modèle qui maximise cet indicateur. En revanche et bien que l'AUC soit maximisé au sein de ce modèle, cet indicateur reste faible. Le manque de contrôle sur les données ainsi que le nombre restreint de variables à notre disposition peuvent expliquer cela.

2.4.3 Premiers résultats du modèle

Ce modèle nous permet donc de détecter les populations ayant une forte propension à résilier leur contrat d'assurance. La dernière partie du logiciel addactis® Pricing permet notamment à son utilisateur de réaliser des tableaux croisés dynamiques axés sur les indicateurs de son choix. Ici, la moyenne des estimations de notre modèle par profil est réalisée (tableau 2.1).

		polholder_gender	
polholder_age	polholder_BMCevol	Female	Male
[19 ; 35[	down	0,149075196314	0,172701120286
	stable	0,141859800558	0,166516183938
	up	0,148385168858	0,171675943454
[35 ; 41[	down	0,118065598736	0,137373079949
	stable	0,111913966474	0,132738844724
	up	0,122358093003	0,134203393126
[41 ; 49[	down	0,123264619085	0,134102657176
	stable	0,107651619545	0,124306447904
	up	0,126696420127	0,128452100158
[49 ; ∞[	down	0,101399590338	0,117568512995
	stable	0,08659344049	0,103450576275
	up	0,09991568652	0,110177363606

TABLE 2.1 – Tableau croisé dynamique des résiliations estimées par profil

Ce tableau nous permet de déceler les profils ayant une plus forte appétence à la résiliation. On remarque que globalement, les hommes ont une plus forte probabilité de résiliation que les femmes. Il en est de même pour les assurés jeunes qui résilient leur contrat en moyenne avec une plus grande probabilité que leur aînés. Il semblerait ainsi que les contrats précaires sont plus enclins à être résiliés. On remarque aussi que globalement, plus le CRM d'un individu augmente, moins ce dernier aura tendance à résilier son contrat.

	polholder_gender		
policy_age	Female	Male	Total
[0 ; 1[	0,135895107997	0,14994784381	0,144500311196
[1 ; 2[	0,126223154235	0,140229256013	0,135144966719
[2 ; 3[	0,117308110571	0,133703660404	0,12800134375
[3 ; 4[	0,125878912236	0,139038284596	0,134563804585
[4 ; 5[	0,125960232793	0,138266371761	0,13439337371
[5 ; 10[	0,085317586527	0,102495559667	0,096849056845
[10 ; ∞[	0,065469402865	0,07640837536	0,072272246457
Total	0,119558700084	0,13293933836	0,128100607112

TABLE 2.2 – Tableau croisé dynamique des résiliations estimées par profil

Le tableau croisé (tableau 2.2) représentant en colonne le genre de l'assuré et en ligne l'ancienneté des contrats appuie le propos précédent. Plus un contrat est récent et donc précaire et puis son détenteur a une chance de le résilier élevée.

D'autres outils ou indicateurs existent et nous présenterons en particulier les odds ratio :

Odds Ratio :

L'odds ratio (OR), également intitulé rapport des cotes ou rapport des chances, est une mesure statistique permettant d'évaluer l'effet d'un facteur.
Il se définit par la formule suivante :

$$OR_X(x_A, x_B) = \frac{odd_X(x_A)}{odd_X(x_B)} = \frac{\frac{\pi^X_{x_A}}{1-\pi^X_{x_A}}}{\frac{\pi^X_{x_B}}{1-\pi^X_{x_B}}} \tag{2.7}$$

où A et B sont deux groupes d'individus qui diffèrent uniquement au regard d'une variable explicative X dont ils prennent respectivement la modalité x_A et x_B et où π^X_x désigne $P(Y = 1|X = x)$, la probabilité que le contrat soit résilié sachant que l'individu en question présente la modalité x de la variable X.

Dans le cas de notre modèle, l'odds ratio est le rapport des chances de résilier le contrat d'assurance sachant que la variable X prend la modalité x_A et non x_B.

Nous interpréterons l'odd ratio $OR_X(x_A, x_B)$ de la façon suivante :

— < 1 : cela signifie que l'évènement de résiliation a moins de chances d'être entrepris pour un assuré présentant la modalité x_A que pour un assuré présentant la modalité x_B ;
— > 1 : cela signifie qu'un assuré dont la variable X prend la modalité x_A a plus de chances de résilier son contrat qu'un assuré dont la variable X prend la modalité x_B.

Ces interprétations se font, toutes choses étant égales par ailleurs.

Dans notre cas, nous nous intéresserons à l'odds ratio d'un groupe A par rapport à la modalité de référence utilisée pour la modélisation.

Lorsque la variable à expliquer est binaire et que l'on utilise la fonction logit au sein de la modélisation, la formule analytique de l'odds ratio s'écrit :

$$OR = e^{\beta_i} \tag{2.8}$$

Le tableau ci-dessous (2.3) prend l'exemple de la variable polholder_age (âge du souscripteur). C'est une variable ordinale :

Modalités	Odds Ratio
$[35, 41[$	0,8106
$[41, 49[$	0,7935
$[49, \infty[$	0,6943

TABLE 2.3 – Odds Ratio pour la variable polholder_age

Les rapports des cotes pour la variable âge de l'assuré nous indique que l'acte de résiliation est plus fréquent dans le groupe des assurés ayant entre 19 et 34 ans que dans le reste de la population assurée. Ces odds ratio représentent en effet l'influence de la modalité indiquée dans le tableau (2.3) par rapport à la modalité de référence : $[19, 35[$.

3 Automatisation de la détection de segments de portefeuille non rentables

3.1 L'algorithme PRIM

3.1.1 Introduction

L'algorithme Patient Rule Induction Method (PRIM) est un algorithme proposé par Friedman et Fisher en 1999 [6] qui a connu de nombreuses implémentations dans divers domaines tels que la recherche médicale, l'analyse des risques financiers ou encore les sciences sociales.

L'objectif de cet algorithme est de déterminer des sous-régions de l'espace d'entrée présentant des valeurs de la variable cible relativement élevées ou faibles. Par construction, l'algorithme PRIM cible directement ces régions plutôt qu'indirectement par l'estimation d'une fonction de régression comme pourrait le faire un algorithme CART.

Cette méthode permet ainsi de décrire ces sous-régions par des règles très simples puisque ces dernières représentent l'union de rectangles de l'espace d'entrée. Il s'agit là d'un réel avantage dont bénéficie cet algorithme puisque l'interprétabilité des modèles demeure un véritable enjeu.

3.1.2 Principe et implémentation

On considère dans cette partie Y notre variable d'intérêt, aussi intitulée variable cible. Y peut être binaire, qualitative ou quantitative. Il est à noter que si Y est une variable qualitative, il faudra sous-diviser cette variable en p variables binaires, p étant le nombre de modalités de la variable Y et implémenter l'algorithme PRIM sur chacune de ces sous-variables.
En notant $X_1, ..., X_n$ les variables explicatives de Y qui peuvent être qualitatives ou

quantitatives ainsi que S_j l'ensemble des valeurs que peut prendre la variable X_j, $j = 1, ..., p$, on peut alors définir le domaine d'entrée S par un produit à n dimensions :

$$S = S_1 \times ... \times S_n \tag{3.1}$$

Fonction à optimiser :

L'objectif de l'algorithme PRIM sera alors de trouver une sous-région R de notre espace d'entrée telle que $R \subset S$ et telle que :

$$\overline{Y_R} > \overline{Y} \tag{3.2}$$

où $\overline{Y_R}$ représente la moyenne de la variable cible Y sur les données d'entrée appartenant à la région R et où $\overline{Y}$ représente la moyenne de la variable cible Y sur l'ensemble des données d'entrée. On peut de la même façon s'intéresser à des sous régions R' telle que $R' \subset S$ et telle que :

$$\overline{Y'_R} < \overline{Y} \tag{3.3}$$

Le même algorithme sera alors utilisé en inversant l'ensemble des signes de la variable cible Y. Sans perte de généralité, on s'intéressera pour la partie théorique uniquement au premier cas.

En sortie de l'algorithme PRIM, nous obtiendrons donc une région R définie par :

$$R = \bigcup_{k=1}^{K} B_k \tag{3.4}$$

Cela signifie que la région solution R est l'union d'un ensemble de sous-régions simplement définies. Et c'est là un réel atout de l'algortithme : en analyse descriptive des données, l'interprétabilité devient un enjeu important. Ici, la région solution R est spécifiée par des énoncés simples (conditions logiques) portant sur les valeurs des variables d'entrée individuelles.

Du fait de leur forme rectangulaire en deux dimensions (orthotopes ou hyper-rectangles pour des dimensions supérieures), nous appellerons dans la suite ces sous-régions des boîtes.

Ainsi, une boîte B sera définie par :

$$B = s_1 \times ... \times s_n \tag{3.5}$$

où $s_i \subset S_i$ représente un sous-ensemble des valeurs que peuvent prendre la variable d'entrée X_i. L'observation e appartiendra donc à B si l'ensemble de ses modalités x_i appartiennent respectivement à s_i :

$$e \in B \Leftrightarrow \bigcap_{i=1}^{n} (x_i \in s_i) \tag{3.6}$$

Dans le cas où $s_i = S_i$, c'est-à-dire que la boîte B présente l'ensemble des modalités de la variable X_i, cette dernière n'a donc pas d'influence sur la variable cible Y et peut être retirée de l'écriture. On notera alors :

$$e \in B \Leftrightarrow \bigcap_{s_i \neq S_i} (x_i \in s_i) \tag{3.7}$$

On dira que les variables d'entrée X_i pour lesquelles $s_i \neq S_i$ sont celles qui définissent la boîte B.

Support :

Une propriété importante pour la suite présentée par ces boîtes est leur support, défini par :

$$\beta_B = \frac{1}{N} \sum_{l=1}^{N} \mathbb{1}_{e_l \in B} \tag{3.8}$$

où e_l est l'une des N observations de notre jeu de données.

Ce nombre, compris entre 0 et 1, indique donc la part représentée par les individus de la boîte B par rapport à la population totale.

Stratégie de couverture :

Précédemment, nous avons vu que la procédure d'optimisation avait pour objectif d'induire un ensemble de boîtes à partir des données de départ couvrant la région de l'espace d'entrée où la fonction cible présente des valeurs élevées. Étant donné un algorithme permettant de construire de telles boîtes (présenté dans la partie suivante), cela peut être accompli par une stratégie de "covering". Le même algorithme de construction de boîte est alors appliqué à plusieurs reprises et de manière séquentielle à des sous-ensembles de données. La première boîte B_1 est construite à partir de l'ensemble des données d'entrée. La deuxième, B_2 est elle construite à partir des données d'origine privée des données appartenant à B_1 :

$$B_2 = \{y_i; x_i | x_i \notin B_1\} \tag{3.9}$$

et ainsi de suite. Différents critères d'arrêt peuvent être définis. En effet, cette procédure peut s'arrêter lorsque la moyenne cible de la boîte B_K ($\overline{Y_K}$) devient inférieure à la moyenne globale $\overline{Y}$ ou à un seuil précédemment défini $\overline{Y_0}$. Mais la procédure peut aussi être interrompue lorsque le support de la boîte :

$$\beta_K = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1} \left(x_i \in B_K \ \& \ x_i \notin \bigcup_{k=1}^{K-1} B_k \right) \tag{3.10}$$

devient trop petit. Ici, N représente le nombre d'observations présentes dans notre base de données.

L'output final de l'algorithme sera un ensemble de boîtes $B_1, ..., B_k$ qui pourront être mises en commun en une unique région $R = \bigcup_{i=1}^{k} B_i$ selon le cadre d'utilisation.

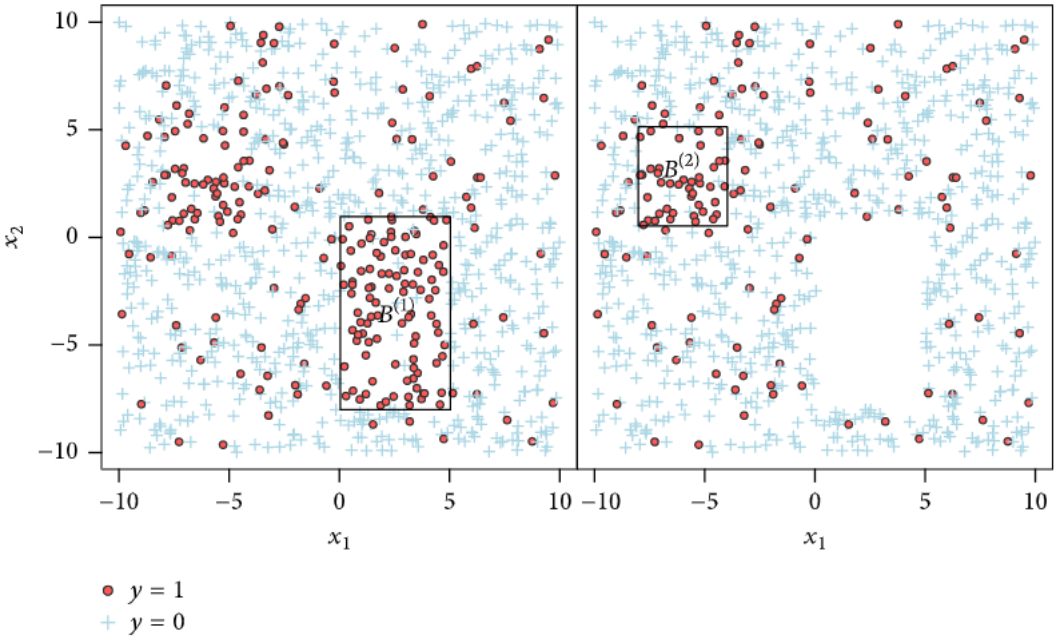

FIGURE 3.1 – Stratégie de couverture de l'algorithme PRIM

A travers la figure ci-dessus tirée de l'article de Ott et Hapfelmeier [12], ce processus est illustré. En effet, la variable cible est ici une variable booléenne qui vaut 0 ou 1 et deux variables explicatives sont utilisées : x_1 et x_2. On peut très rapidement remarquer, à l'oeil nu, que deux zones de l'espace d'entrée semblent être fortement concentrées en réponses positives ($y = 1$). Ainsi, une fois que la boîte B_1 est déterminée par l'algorithme (partie gauche de l'illustration), ce dernier va alors

chercher à déterminer une seconde boîte B_2 sur les données restantes. Ce processus se répéterait alors tant qu'un critère d'arrêt ne serait pas atteint.

3.1.3 Construction des boîtes individuelles

La stratégie utilisée par l'algorithme PRIM quant à la construction des boîtes est celle du raffinement successif descendant, suivi d'une expansion récursive ascendante. Deux étapes sont donc nécessaires à l'obtention d'une boîte :

— Le top-down peeling
— Le bottom-up pasting

Le top-down peeling

Cette première étape débute par une boîte B qui recouvre l'ensemble des données d'entrée S. À chaque étape ou encore itération de l'algorithme, une sous-boîte b^* de la boîte B actuelle est retirée. Cette sous-boîte particulière b^* est choisie de sorte à ce que la moyenne de notre variable cible Y sur notre nouvelle boîte $(B - b^*)$ soit supérieure à la moyenne sur la boîte B initiale. A chaque itération, la sous-boîte b^* que nous retirons ne sera définie que sur une seule variable X_i. Cette procédure descendante est alors répétée jusqu'à atteindre un critère d'arrêt qui sera défini par la suite.
Un pseudo-code peut ainsi être explicité afin de faciliter la compréhension du processus réalisé lors de l'étape du top-down peeling.

Pseudo-code du top-down peeling :

$B \leftarrow$ Ensemble des données d'entrée
Répéter :

1. $\mathscr{C}(b)$: ensemble des sous-boîtes b pouvant être retirées à B
2. $b^* = \arg\max_{b \in \mathscr{C}(b)} \{\overline{y_i} | x_i \in B \setminus b\}$
3. $B \leftarrow B \setminus b^*$
4. $\beta_B \leftarrow$ support de B, c'est-à-dire : $\beta_B = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{x_i \in B}$ avec N le nombre d'observations.

Jusqu'à ce que $\beta_B \leq \beta_0$
Retourner B

Le paramètre β_0 doit donc être défini en amont et correspond à un seuil en dessous duquel la boîte est considérée trop petite par l'algorithme. Sa valeur doit donc

faire l'objet d'une réflexion selon le cadre d'utilisation de cet algorithme.
Ainsi, β_0 définit l'effectif minimum : l'algorithme se termine lorsqu'il est nécessaire d'avoir un effectif inférieur à l'effectif minimum β_0 pour améliorer la moyenne de la dernière boîte obtenue.

Se pose maintenant la question de la façon dont on détermine $\mathscr{C}(b)$ et donc l'ensemble des sous-boîtes b pouvant être retirées à chaque étape de l'algorihme. En effet, à chaque itération, l'algorithme de top-down peeling va déterminer la ou les boîtes éligibles par variable d'entrée X. En revanche, ces dernières seront déterminées différemment selon la nature de la variable explicative en question et c'est ce que nous allons voir maintenant.

Cas d'une variable qualitative :
Si X_i est une variable qualitative, il y a autant de boîtes b éligibles que de modalités de X_i encore présentes dans la boîte B.
On obtient alors : $b_{jm} = \{x|s_j = s_{jm}\}$ où $s_{jm} \subset S_j$. Dans le cas d'une variable qualitative, l'algorithme n'a pas de contrôle sur le nombre d'observations retirées puisque toutes les observations présentant la même modalité sont étudiées en une seule fois.

Cas d'une variable quantitative :
Pour chaque variable quantitative X_i, deux boîtes sont éligibles à chaque itération de l'algorithme :

- $b_{j^-} = \{x|x \leq x_{j(\alpha)}\}$
- $b_{j^+} = \{x|x \geq x_{j(1-\alpha)}\}$

Ici, $x_{j(\alpha)}$ est un quantile de niveau α des x_j présents dans la boîte actuelle. Cette quantité α est un paramètre de l'algorithme souvent choisi assez faible ($\alpha \leq 0,1$). Ce sont donc des intervalles bordant la limite inférieure et la limite supérieure de la boîte courante B qui seront éligibles. Ainsi, une boîte ne pourra être éligible que si son effectif n'est pas trop important pour être retiré et l'algorithme PRIM exerce de ce fait un contrôle sur la proportion d'observations retirées à chaque étape ce qui lui vaut le nom d'algorithme "patient".

Une fois que l'ensemble $\mathscr{C}(b)$ est défini, la deuxième étape de l'algorithme consiste donc à retirer la boîte b (représentant une modalité d'une variable qualitative ou un intervalle d'une variable quantitative) pour laquelle la moyenne de B privée de cette sous-boîte est maximisée.

Afin d'illustrer ces propos, une figure simple tirée de l'article de Ott et Hapfelmeir [12] peut être utilisée :

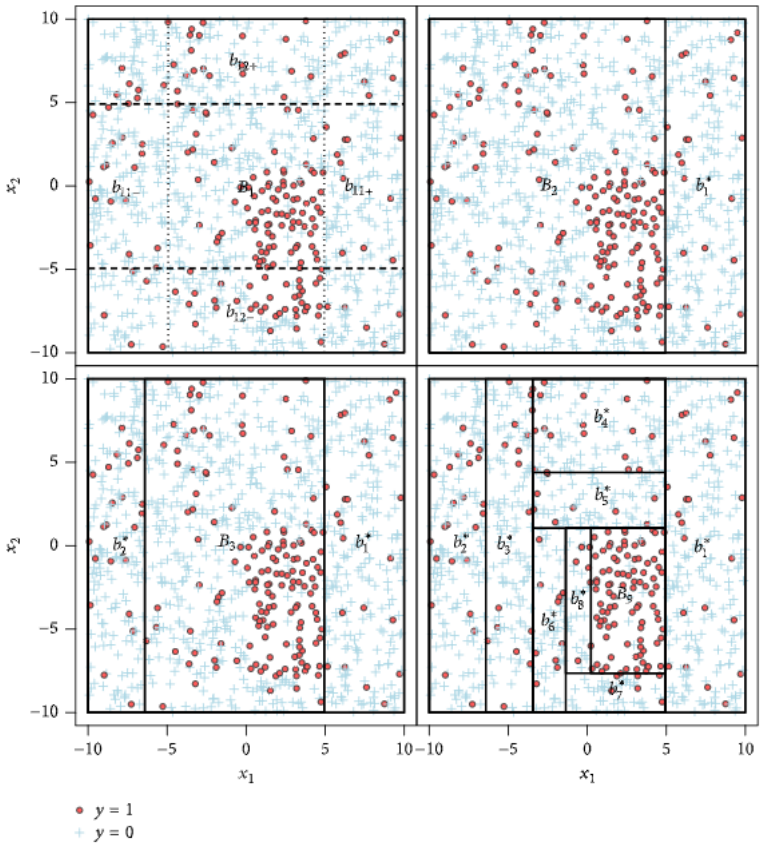

FIGURE 3.2 – Illustration de l'implémentation de la procédure de peeling en deux dimensions

La figure ci-dessus permet d'exemplifier les différentes étapes suivies par l'algorithme de peeling et les différentes boîtes qui en résultent. On se place ici dans un cadre simplifié d'une variable cible binaire et d'un domaine d'entrée à deux dimensions. On considère également que X_1 et X_2 sont deux variables quantitatives. On peut alors rapidement comprendre le processus adopté par l'étape de peeling de l'algorithme PRIM.
Ici, le paramètre de peeling α a été choisi assez grand ($\alpha = 0,25$). Dans l'encadré supérieur gauche, la boîte initiale contenant tous les points ainsi que les quatre boîtes éligibles (ou boîtes candidates) pour la première étape de peeling sont affichées. Les limites en pointillées sont les quantiles de niveau α et $1 - \alpha$. Le deuxième graphique nous montre donc que la boîte candidate b_{11+} est finalement celle retenue pour être retirée et devient b_1^*. L'algorithme continue ainsi et le troisième graphique illustre la deuxième étape de l'algorithme avec les deux sous-boîtes b_1^* et b_2^* qui ont été retirées de la boîte initiale pour former B_3. Le dernier graphique montre le résultat final de

l'algorithme, atteint lorsque le méta-paramètre β_0 (qui vaut ici 0,075) soit atteint. Cela signifie que B_9 contient au moins 7,5% de toutes les observations.
Dans cet exemple, on obtient une régle pour B_9 qui s'exprime en fonction des deux variables X_1 et X_2 comme :

$$B_9 = \begin{cases} 0,23 < X_1 < 4,94 \\ -7,64 < X_2 < 1,05 \end{cases} \tag{3.11}$$

Le bottom-up pasting

L'objectif de l'algorithme top-down peeling est de déterminer une boîte B couvrant une sous-région de l'espace des variables d'entrée dans laquelle la moyenne de notre variable cible est relativement grande. Les limites de cette boîte B ont donc été déterminées par des valeurs particulières des variables X_i qui ont défini les sous-boîtes b^* choisies pour le peeling aux différentes étapes de la procédure. À l'exception de la dernière, les limites finales de la boîte B ont été déterminées aux étapes antérieures de la séquence de peeling sans tenir compte des peelings ultérieurs qui affinaient davantage les limites d'autres variables d'entrée. Il est donc possible que la boîte finale puisse être améliorée en réajustant certaines de ses limites.
Pour palier cette limite présentée par le top-down peeling, un procédé de bottom-up pasting est introduit par l'algorithme PRIM. Le principe de cette deuxième étape est de réaliser le procédé inverse du peeling puisque c'est une approche ascendante qui est utilisée.
En effet, lors du bottom-up pasting, la boîte de départ B est la boîte finale obtenue par le peeling et l'objectif n'est pas cette fois de la rétrecir mais bien de l'élargir toujours en respectant notre objectif initial : déterminer un segment déviant qui se traduit mathématiquement par l'augmentation de la moyenne de notre variable cible.

A nouveau, un processus itératif est mis en place :

Pseudo-code du bottom-up pasting :
$B \leftarrow$ Boîte obtenue par le peeling
Répéter :

1. $\overline{Y} = \{\overline{y_i} | x_i \in B\}$
2. $\mathscr{C}(b)$: ensemble des sous-boîtes b pouvant être ajoutées à B
3. $b^* = \arg\max_{b \in \mathscr{C}(b)} \{\overline{y_i} | x_i \in B \cup b\}$
4. $B \leftarrow B \cup b^*$

Jusqu'à ce que $\overline{Y_{B \cup b^*}} \leq \overline{Y_B}$
Retourner B

L'ensemble des classes éligibles du pasting $\mathscr{C}(b)$ est défini de la même façon que pour le peeling. Pour une variable qualitative, chaque modalité non présente dans la boîte représente une boîte éligible. Pour une variable qualitative, les boîtes éligibles représentent des intervalles visant à étendre les limites inférieures ou supérieures de la boîte actuelle. La largeur de ces intervalles est choisie de sorte à contenir αN_B observations où α est le paramètre de peeling et où N_B représente le nombre d'observations présentes dans la boîte actuelle.
Comme l'indique ce pseudo-code, l'étape du pasting s'arrête lorsque la moyenne de la variable cible Y n'est plus maximisée par l'agrandissement de la boîte obtenue.

Bien que le bottom-up pasting apporte généralement une certaine amélioration, il a rarement un effet considérable.

Une illustration plus concrète du top-down peeling ainsi que du bottom-up pasting est disponible en annexe C.

3.1.4 Hold-out validation

Un risque fréquent auquel sont exposés les algorithmes classiques tels que les GLMs ou encore les algorithmes de Machine Learning s'intitule risque de sur-apprentissage. Une vigilance doit donc, pour l'algorithme PRIM également, être portée à ce problème.

En effet, la boîte déterminée par les étapes de peeling et de pasting peut présenter une moyenne sur notre variable cible très élevée sur les données utilisées mais cela peut être dû aux particularités de l'échantillon. De plus, plus l'effectif de la boîte déterminée est faible, plus le risque de sur-apprentissage est élevé.

De ce fait, une solution peut être proposée en partitionnant les données en une base d'apprentissage et une base de validation. L'algorithme PRIM et donc les étapes de peeling et de pasting sont d'abord entraînées sur l'échantillon d'apprentissage puis les résultats sont appliqués à l'échantillon de validation. On obtient donc une estimation de la moyenne de notre variable cible Y dans chacune des boîtes successives. Les boîtes sélectionnées seront donc celles qui présenteront également une moyenne élevée sur l'échantillon de validation.

3.1.5 Trajectoire

Une illustration graphique des étapes de peeling peut être donnée par la trajectoire. Ce graphique représente la valeur prise par la fonction cible à chaque itération par rapport au support de la boîte correspondante. Ainsi, il est possible de trouver une boîte "optimale" à partir de ce compromis entre le résultat moyen et le support

de la boîte.

Évidemment, au fur et à mesure que l'algorithme se poursuit, le support de nos boîtes diminue et la moyenne de notre cible, dans la plupart des cas, augmente. La trajectoire peut tout de même permettre à l'utilisateur de choisir une ou plusieurs boîtes présentant des propriétés conformes à des exigences spécifiques.

Une illustration de cette trajectoire est la suivante :

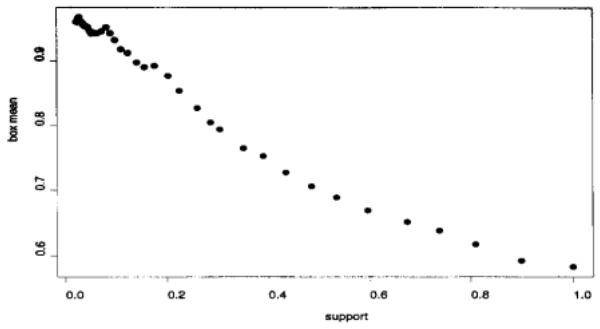

FIGURE 3.3 – Exemple de trajectoire d'un processus de peeling

Trajectoires multiples :

Lorsque l'algorithme top-down peeling est appliqué, une trajectoire de peeling est induite comme mentionné précédemment. Mais cette trajectoire dépend évidemment des données ainsi que du paramétrage de l'algorithme. Ce paramétrage se traduit par le choix des valeurs des méta-paramètres qui contrôlent divers aspects de la procédure. L'un d'eux est le paramètre de peeling α. Bien que la procédure soit relativement stable, différentes valeurs de ce paramètre peuvent induire des trajectoires différentes et la question de la "sélection du modèle" se pose selon les besoins de l'utilisateur.

Ainsi, pour sélectionner le modèle le plus adapté, une stratégie possible consiste à présenter simultanément un grand nombre de trajectoires sur un seul graphique, chaque trajectoire présentant des valeurs de méta-paramètres différents. On obtient ainsi un nuage de points qui représente pour chaque boîte le support et la valeur de la fonction cible.

3.1.6 Le modèle PRIM face aux modèles "classiques"

Le modèle PRIM est un modèle dont l'objectif se rapproche des modèles classiques tels que les modèles CART ou encore Random Forest : la mission de l'ensemble de ces trois modèles peut par exemple s'avérer être la classification de données ou, plus concrètement, le découpage de l'espace des variables explicatives en régions $R_1, ..., R_n$ (les feuilles de l'arbre ou boîtes de l'algorithme PRIM). Plusieurs différences existent tout de même entre l'algorithme PRIM et l'algorithme CART pour ne citer que lui :

— En premier lieu et dans le cas de l'algorithme PRIM, le programme s'assimile à une optimisation (maximisation ou minimisation selon les cas). Ce n'est donc pas nécessairement l'ensemble du domaine d'entrée qui doit être segmenté : l'algorithme PRIM se concentre sur les extrêmes en maximisant la moyenne de la variable d'intérêt. Dans le cas d'un arbre CART, les feuilles sont définies en minimisant la variance intra-groupes. La construction d'une feuille dépend donc de celle des autres : dans notre cas les feuilles déviantes (et donc qui nous intéressent) sont déterminées de sorte à également minimiser la variance intra-groupes des feuilles qui ne nous intéressent pas dans le cadre de notre étude. Les performances obtenues sur les feuilles déviantes peuvent ainsi être altérées.

— Ensuite, PRIM propose une interaction forte avec son utilisateur : le paramètre β_0 permet de déterminer l'effectif minimum de chacune des boîtes détectées. Un paramètre semblable peut être utilisé lors de l'implémentation de l'algorithme CART. C'est notamment le paramètre « minbucket » de la librairie « rpart » classiquement utilisée au sein du logiciel R. En revanche, l'objectif dans notre cas est d'arriver à des segments dont l'effectif est réellement proche de notre paramètre β_0 ou minbucket. Il est important, à des fins commerciales, de contrôler la proportion d'individus sur laquelle nous voudrons par la suite appliquer un changement tarifaire. L'algorithme PRIM, du fait du contrôle qu'il exerce sur le nombre d'observations retirées à chaque étape du top-down peeling (paramètre de peeling alpha), parvient à déterminer une solution qui s'approche grandement de notre effectif minimum β_0 : c'est ce qui lui vaut son caractère « patient ». Un tel contrôle n'est pas disponible au sein de l'algorithme CART. Le paramètre « minbucket » aura ainsi moins d'influence sur l'implémentation de l'arbre CART.

— Enfin, l'algorithme CART « split » ou divise successivement les données en deux selon des variables explicatives et jusqu'à ses feuilles tout en minimisant la variance intra-groupes. Des critères ou nœuds s'appliqueront ainsi à plusieurs feuilles. La procédure est entièrement descendante. A l'inverse, PRIM permet d'utiliser des critères (ou variables discriminantes) différents pour chaque segment détecté. En effet, pour chacun d'entre eux : l'algorithme

PRIM réengage une procédure ascendante en repartant des données restantes à être classifiées. Bien que l'élagage qui est une procédure ascendante (on cherche à simplifier l'arbre en élaguant ses branches de bas en haut) soit une méthode courante pour l'algorithme CART, cela n'atteint pas les mêmes objectifs que la procédure proposée par PRIM.

3.2 Analyse des boîtes obtenues

Une fois l'algorithme implémenté, l'utilisateur a accès à un ensemble de règles définissant une séries de boîtes. Des analyses peuvent alors être effectuées afin de mieux comprendre les résultats de l'algorithme.

3.2.1 Diagrammes de fréquences relatives

Pour chaque variable d'entrée X_i et pour chaque boîte B_k, un diagramme de fréquences relatives peut être construit. Pour chaque modalité s_i de la variable X_i, le calcul sera le suivant :

$$r_{ik}(s_i) = \frac{p_i(x_i|x_i \in B_k \ \& \ x_i = s_i)}{p_i(s_i)} \tag{3.12}$$

où $p_i(x_i|x_i \in B_k \ \ x_i = s_i)$ représente la proportion de lignes ayant la modalité s_i et présentes dans la boîte B_k et où $p_i(s_i)$ représente la proportion de lignes ayant la modalité s_i dans l'ensemble des données d'entrée.

Une distribution uniforme pour $r_{ik}(x_i)$ impliquera que x_i est totalement sans rapport avec la définition de la boîte B_k puisque la fréquence relative de ses valeurs est la même à l'intérieur ou à l'extérieur de la boîte. Les écarts par rapport à l'uniformité indiquent un lien avec la définition de la boîte. Une distribution très élevée de $r_{ik}(x_i)$ sera révélatrice d'une donnée X_i très pertinente pour la définition de B_k, qu'il s'agisse ou non de l'une des variables ayant déterminé les règles de la boîte en question.

3.2.2 Dissimilarités entre boîtes

La procédure de recouvrement de l'algorithme PRIM permet l'obtention d'une séquence de boîtes qui couvrent une sous-région R de l'espace des variables d'entrée. En revanche, aucune contrainte n'est placée sur les relations entre les boîtes. Ces dernières peuvent donc tout à fait être disjointes, se chevaucher ou même être très éloignées les unes des autres, selon la nature de la fonction cible.

Ainsi, il peut paraître intéressant de connaître les relations existantes entre les différentes boîtes obtenues.
Ces relations entre les boîtes peuvent être déduites de leurs dissimilarités par paires. La dissimilarité $D(B_k, B_l)$ entre les boîtes B_k et B_l est définie comme la différence entre le support de la plus petite boîte B_{kl} qui couvre les deux d'entre elles, et le support de l'union des deux boîtes.
Autrement dit,

$$D(B_k, B_l) = \beta(B_{kl}) - \beta(B_k \cup B_l) \tag{3.13}$$

On aura ainsi,

$$x \in B_{kl} \Leftrightarrow \bigcap_{j=1}^{n} [x_j \in s_j(k,l)] \tag{3.14}$$

Le sous-ensemble $s_j(k,l) \in S_j$ est obtenu à partir des sous-ensembles respectifs des deux boîtes : $\{s_{jk}\}_{j=1}^n$ et $\{s_{jl}\}_{j=1}^n$ selon le type de chacune des variables x_j.
Pour une variable x_j continue ou ordinale, $s_j(k,l)$ représentera le plus petit intervalle $[t_j^-(k,l), t_j^+(k,l)]$ qui couvrira les intervalles respectifs des boîtes B_k et B_l pour la variable x_j. Mathématiquement, cela se traduira par :

$$t_j^-(k,l) = min(t_{jk}^-, t_{jl}^-) \qquad t_j^+(k,l) = max(t_{jk}^+, t_{jl}^+) \tag{3.15}$$

Pour une variable catégorielle x_j, le sous-ensemble $s_j(k,l)$ sera l'union des sous-ensembles respectifs de B_k et B_l :

$$s_j(k,l) = s_{jk} \cup s_{jl} \tag{3.16}$$

La dissimilarité entre deux boîtes appartiendra donc à l'intervalle $[0, 1[$. Lorsque la dissimilarité sera faible, les segments contenus dans les deux boîtes seront deux segments similaires. A l'inverse, lorsque la dissimilarité sera élevée, cela indiquera que les individus présents dans les deux différentes boîtes seront deux segments très éloignés.

3.3 Adaptation de l'algorithme à notre cas d'usage

Dans le cadre de ce mémoire, un objectif majeur est de détecter des sous-populations du portefeuille pour lesquelles la rentabilité n'est pas conforme aux attentes. Comme nous avons pu le constater lors des précédentes parties, l'analyse graphique est en mesure de nous fournir la rentabilité d'une partie de la population : les femmes ou encore les habitants de Paris. En revanche, il paraîtrait très difficile et surtout chronophage d'effectuer ce travail sur des combinaisons de modalités, aussi nombreuses qu'elles puissent être. C'est ainsi que l'algorithme PRIM prend tout son

sens dans le cas d'usage de l'assurance. L'automatisation de ce processus peut engendrer un gain en efficacité et en temps considérable notamment sur des portefeuilles très volumineux.

L'algorithme aura pour fonction de détecter des sous-populations pour lesquelles notre indicateur de rentabilité est très élevé. En effet, un Loss Ratio ou encore un Expected Loss Ratio supérieur à 1, indiquera que la sinistralité de l'individu est supérieure à la prime que ce dernier paye à l'assureur pour couvrir son risque. A l'inverse, il peut aussi paraître tout à fait judicieux de s'intéresser aux sous-populations présentant une très bonne rentabilité qui pourraient être la cible d'offres promotionnelles ou d'un intérêt tout particulier. Dans ce cas, on s'intéressera particulièrement aux Loss Ratio ou Expected Loss Ratio très inférieurs à la moyenne du portefeuille.

Ainsi, la variable cible Y retenue sera, tout au long de cette partie, l'Expected Loss Ratio qui représente le ratio entre la prime pure prédite et la prime acquise.
Les variables explicatives $X_1, ..., X_n$ seront les variables explicatives de notre base de données.

3.4 Application de l'algorithme sur la base de données

Dans cette partie, l'objectif va donc être de mettre en pratique les considérations théoriques précédentes. Le but ici est donc de détecter des segments du portefeuille pour lesquels l'Expected Loss Ratio est déviant. L'application a été réalisée à l'aide de la fonction R intitulée *PRIM* et issue du package éponyme [2].

3.4.1 Paramétrage des modèles implémentés

Deux paramètres doivent être renseignés par l'utilisateur du modèle PRIM : le paramètre β_{min} qui représente le support minimum de chaque boîte et donc de chaque segment déviant. Ce paramètre a été fixé à 0,01 tout au long de cette application. Chaque segment déviant représentera donc au moins 1% du portefeuille d'assurés. Une société d'assurance qui souhaiterait mettre en place une offre promotionnelle pour 5% de son portefeuille d'assurés (ses bons risques), pourrait alors fixer ce paramètre à 0,05.
Quant au paramètre de peeling α, la fonction R permet de définir un ensemble de valeurs. A chaque étape de peeling, 30 valeurs de α, comprises entre 0,01 et 0,4 avec un pas de 0,01, seront alors testées. A chaque itération, le paramètre α permettant de détecter la meilleure sous-boite b^* sera retenu.

Les variables explicatives retenues pour cet algorithme sont les suivantes :

- L'âge de l'assuré : *polholder_age*
- L'évolution de son coefficient bonus-malus : *polholder_BMCevol*
- La différence de statut entre l'assuré et le conducteur : *polholder_diffdriver*
- Le genre de l'assuré : *polholder_gender*
- Son travail (milieu médical ou non) : *polholder_job*
- L'ancienneté du contrat : *policy_age*
- L'âge du véhicule : *vehicl_age*
- L'âge du véhicule à l'achat : *vehicl_agepurchase*
- Le type de garage du véhicule : *vehicl_garage*
- La région de vie de l'assuré :*vehicl_region*

Ce sont donc les mêmes variables que celles retenues pour le modèle de résiliation à l'exception de la variable prime commerciale puisque l'ELR est lui-même basé sur cette variable.

3.4.2 Résultats - Segments déviants sous-tarifés

Le premier modèle implémenté a permis de définir 4 segments :

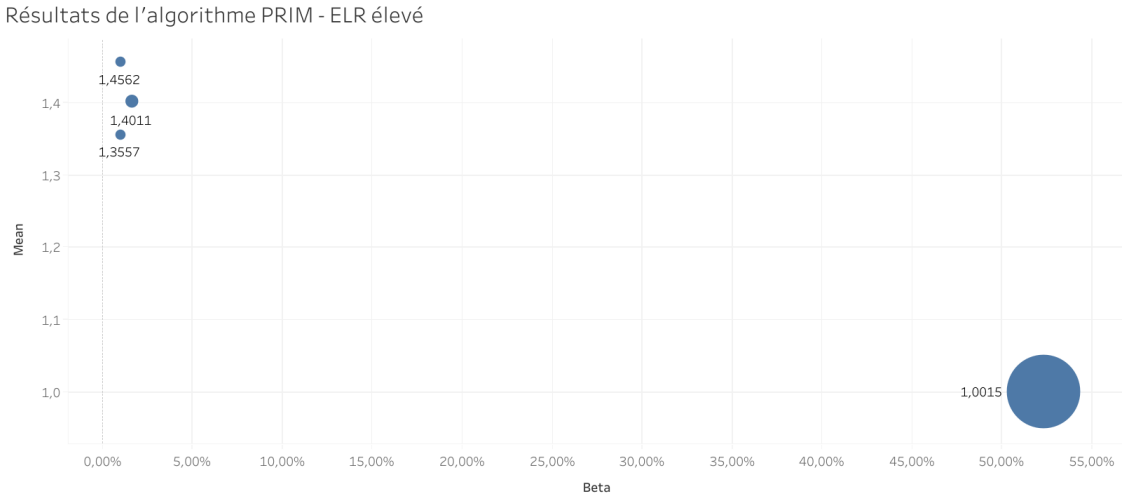

FIGURE 3.4 – Segments déviants obtenus par l'algorithme PRIM

En abscisse de la figure 3.4 se trouve l'exposition du segment en question et en ordonnée est représentée la moyenne des ELR sur ce segment. On peut alors dé-

tailler, segment par segment, les résultats obtenus :

- Le premier segment représente 1,0104% de la population totale et se compose des assurés ayant une ancienneté strictement inférieure à 5 ans, dont le véhicule est âgé de moins de 17 ans, vivant dans les régions 2, 5, 6, 9, 10, 12 ou 13 et dont le CRM est en hausse. Sur ce segment, l'ELR moyen est de 1,4562 et est donc 1,52 fois supérieur à l'ELR moyen du portefeuille.
- Le second segment représente 1,62% du portefeuille et est constitué des assurés de moins de 74 ans, dont le CRM est en hausse. Leur région de vie se situe dans les régions 3, 4, 7, 8, 9. Enfin, le conducteur principal de cette voiture n'est ni un apprenti, ni un jeune conducteur, ni un conducteur sans lien avec l'assuré. Sur ce segment, l'ELR moyen vaut 1,40.
- 1,01% de la population totale compose le 3ème segment. Ce dernier est défini par les assurés ayant un véhicule âgé de plus de 7 ans, dont le CRM est en hausse et n'habitant pas en région 12.
- Enfin, le dernier segment représente une majorité du portefeuille : 52,3%. Il se compose des assurés dont le CRM est stable ou en hausse. On remarque donc que la variable CRM est une variable très discriminante. Mais la modalité "stable" étant une modalité très représentée dans le portefeuille (cf figure 2.4), elle empêche ainsi à l'algorithme d'aller plus loin. Il pourrait donc être intéressant d'observer les résultats d'un algorithme qui ne tiendrait pas compte de cette variable.

En retirant la variable indiquant l'évolution du CRM, ce sont dorénavant 51 segments qui sont détectés par l'algorithme PRIM. Ces segments sont représentés par la figure 3.5 ci-dessous :

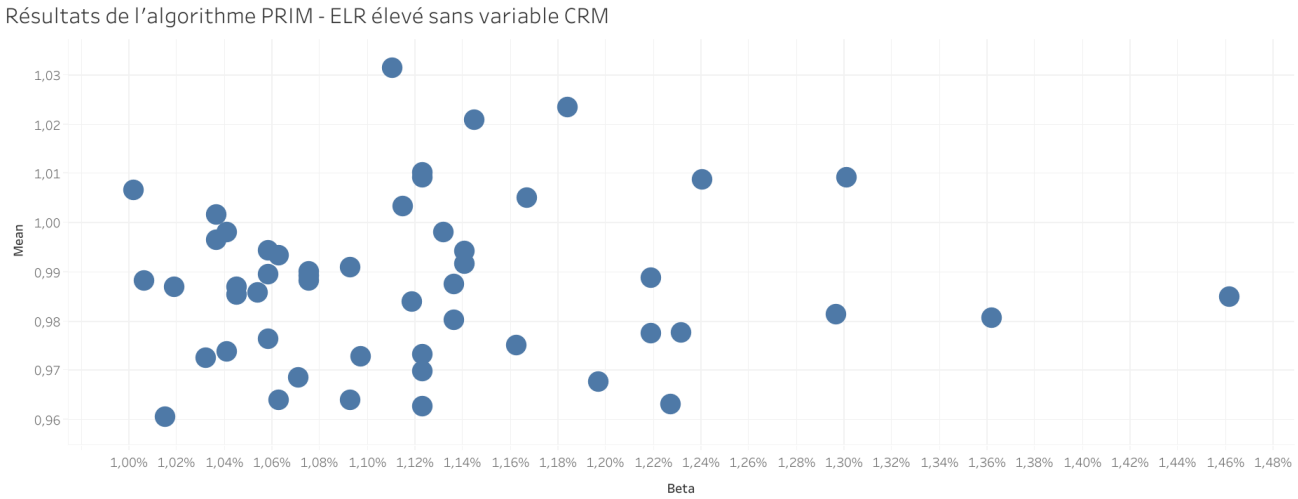

FIGURE 3.5 – Segments déviants obtenus par l'algorithme PRIM sans la variable CRM

Il est ici notable que la performance du modèle est nettement diminuée : l'ELR maximum détecté vaut 1,03 et n'est plus que 1,07 fois supérieur à la moyenne du portefeuille. La variable indiquant l'évolution du CRM est donc fortement significative quant à la détection de segments d'ELR déviants.

Nous allons donc dans la suite nous intéresser aux 4 boîtes obtenues grâce à la première implémentation du modèle.

Diagrammes de fréquences relatives :

L'analyse des diagrammes de fréquences relatives est un moyen de déceler des variables pertinentes dans la définition des différentes boîtes malgré leur absence dans les règles de ces dernières.
On rappelle que ce sont l'ancienneté de la police d'assurance, l'âge du véhicule, la région ainsi que l'évolution du CRM qui déterminent le premier segment déviant et l'étude suivante sera réalisée uniquement pour ce segment.

En s'intéressant au diagramme de fréquences relatives de la variable âge du souscripteur (figure 3.6), on remarque que les souscripteurs jeunes (dont l'âge est inférieur ou égal à 20 ans) sont quatre fois plus représentés dans le sous-segment que nous étudions. En effet, ils représentent 0,4% au sein du premier sous-segment alors qu'ils ne représentaient que 0,1% du portefeuille global.

FIGURE 3.6 – Diagramme de fréquences relatives associé à l'âge du souscripteur

Cela signifie ainsi que la population présente dans la boîte 1 est plus jeune que la population globale du portefeuille d'assurés.

Le diagramme de fréquences relatives associé au statut de l'assuré par rapport au souscripteur (figure 3.7) rejoint notre constatation précédente : les assurés jeunes ou même apprentis sont sur-représentés dans notre boîte d'étude.

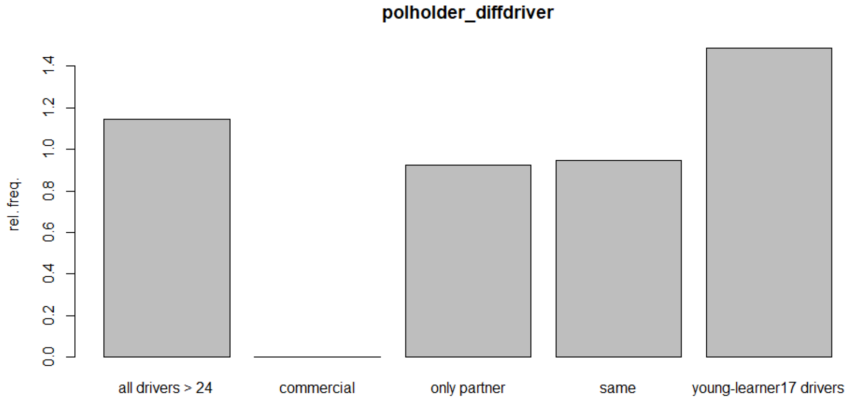

FIGURE 3.7 – Diagramme de fréquences relatives associé au statut de l'assuré par rapport au souscripteur

Ensuite, concernant le genre de l'assuré, ce sont les femmes qui sont les plus représentées dans ce sous-segment (figure 3.8) :

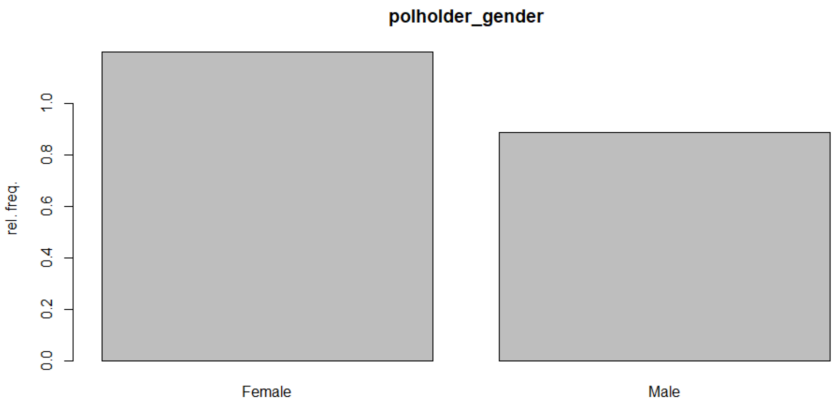

FIGURE 3.8 – Diagramme de fréquences relatives associé au genre du souscripteur

En revanche, aucune tendance ne peut être observée quant au secteur d'activité du souscripteur (figure 3.9). Cela signifie donc que la répartition des travailleurs au sein de notre première boîte est la même que celle au sein de notre jeu de données.

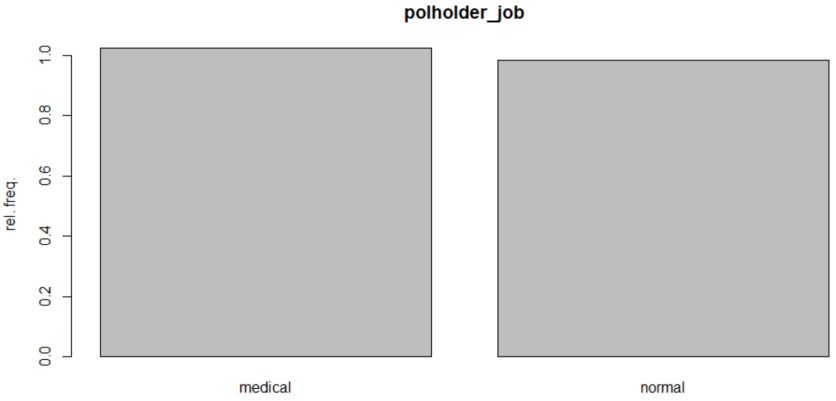

FIGURE 3.9 – Diagramme de fréquences relatives associé au secteur d'activité du souscripteur

Enfin, on peut s'intéresser à notre variable cible : l'ELR avec la figure 3.10.

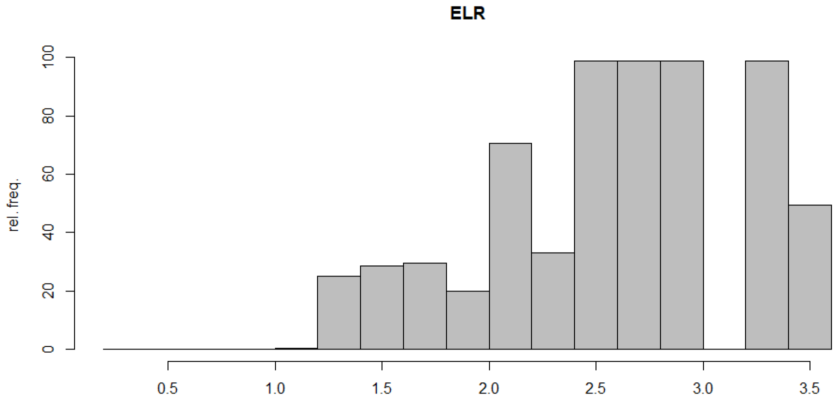

FIGURE 3.10 – Diagramme de fréquences relatives associé à l'ELR de l'assuré

Ici, les ordres de grandeurs sont nettement supérieurs : les contrats dont l'ELR est compris entre 2,5 et 3 sont cent fois plus représentés dans la boîte 1 que dans l'ensemble de la population : notre algorithme permet ainsi de bien cibler ces contrats.

Dissimilarité entre boîtes :
Si on s'intéresse maintenant à l'ensemble des boîtes obtenues à l'aide de notre algorithme, on remarque très nettement à l'aide du tableau 3.1 que les segments obtenus sont proches d'être emboîtés.

Dissimilarité	Segment 1	Segment 2	Segment 3	Segment 4
Segment 1		$9,62706 \times 10^{-3}$	$2,775369 \times 10^{-3}$	0
Segment 2			$1,040763 \times 10^{-3}$	0
Segment 3				0

TABLE 3.1 – Etude des dissimilarités entre les différentes boîtes

En effet, si on prend l'exemple de la dissimilarité entre les boîtes B_1 et B_2, on remarque que cette dernière vaut $9,62706 \times 10^{-3}$. Or, la dissimilarité est calculée comme la différence entre le support de la plus petite boîte B_{12} qui couvrirait les deux boîtes B_1 et B_2 et le support de l'union de ces deux dernières. Ici, on remarque que cette différence est très faible : les boîtes B_1 et B_2 sont similaires. Il en est de même pour les dissimilarités entre les boîtes B_1 et B_3, B_1 et B_4, B_2 et B_3, B_2 et B_4 et enfin B_3 et B_4.

3.4.3 Résultats - Segments déviants sur-tarifés

On s'intéresse maintenant aux segments dont l'ELR est très inférieur à sa moyenne et ce sont cette fois quarante segments détectés comme le montre la figure 3.11 ci-dessous :

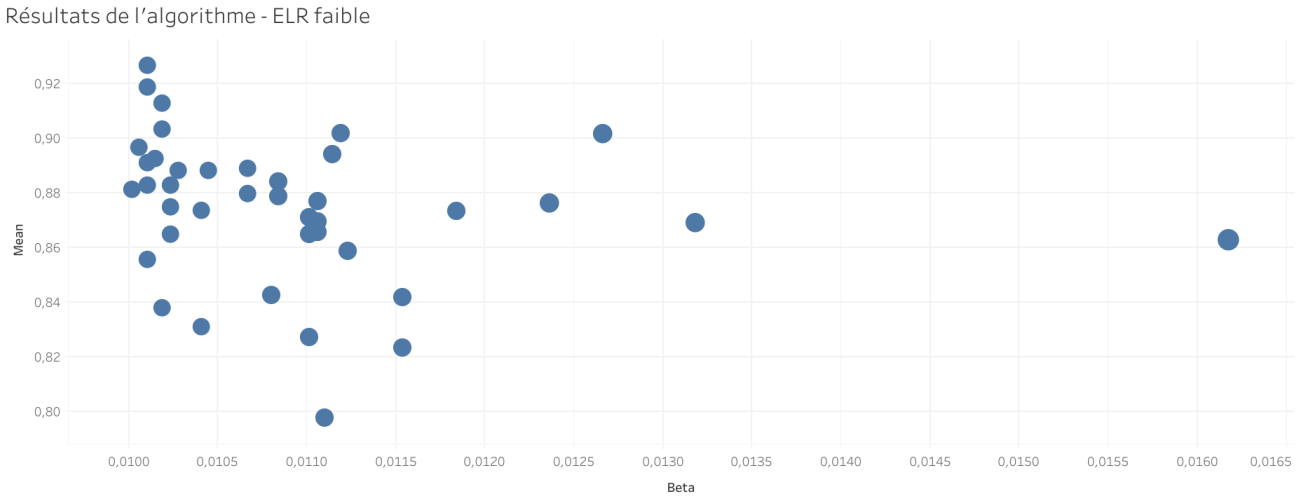

FIGURE 3.11 – Segments déviants obtenus par l'algorithme PRIM

Nous allons ainsi pouvoir expliciter les deux premiers :

— Le premier segment représente les nouveaux contrats dont le CRM est en baisse, souscrits par des assurés de moins de 28 ans, ayant acheté un véhicule de plus de 10 ans et qui aujourd'hui a moins de 18 ans. Ces contrats doivent être souscrits par le conducteur lui-même et dans une région différente des régions 3, 9 et 12. Sur ce segment (1,11% de la population), l'ELR moyen vaut 0,798 et est donc plus de 1,2 fois inférieur à la moyenne globale.

— Le second segment déviant est composé des nouveaux souscripteurs âgés de 28 à 53 ans, détenant un véhicule de moins de 18 ans et acheté à plus de 12 ans. Ces nouveaux souscripteurs doivent avoir un CRM en baisse, travailler dans le secteur "normal", s'assurer pour eux-mêmes, pour un apprenti ou jeune conducteur et vivre dans une région différente des régions 3, 9, 10 et 12. Ici, le segment défini présente un ELR moyen de 0,823.

Diagrammes de fréquences relatives :

A nouveau, on peut s'intéresser aux diagrammes de fréquences relatives du premier sous-segment détecté. On rappelle ici que ce sont l'évolution du CRM, l'âge de

l'assuré, l'âge de la voiture, ainsi que son âge au moment de l'achat, la différence entre le souscripteur et l'assuré et la région de vie qui interviennent dans la définition du segment.

En s'intéressant au diagramme de fréquences relatives associé à la variable représentant l'ancienneté du contrat (figure 3.12), on remarque que les contrats ayant une ancienneté élevée (supérieure à 1 an) ont tendance à être sur-représentés dans notre sous-segment déviant puisque leur fréquence relative est nettement supérieure à 1.

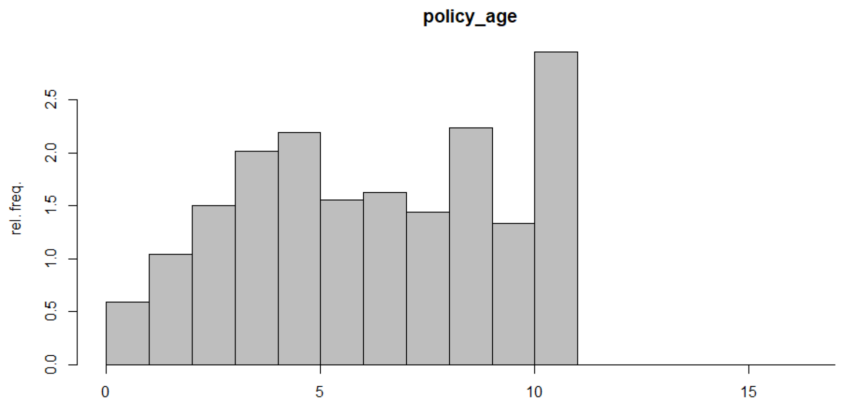

FIGURE 3.12 – Diagramme de fréquences relatives associé à l'ancienneté de la police d'assurance

Le type de garage est lui aussi intéressant puisque le diagramme 3.13 permet de remarquer la sur-représentation des stationnements dans la rue tandis que les stationnements privés sont eux sous-représentés. Cela signifie qu'une majorité des contrats qui déclarent que leur voiture stationne sur la voie publique ont des contrats sur-tarifés. Il apparaît en effet rationnel de penser que ces voitures là seront plus souvent endommagées et cela conduit à une hausse de la prime commerciale. A l'inverse, les assureurs ne redoutent pas autant de sinistres des véhicules bénéficiant d'un stationnement privé et ainsi, leur prime commerciale semble être amoindrie.

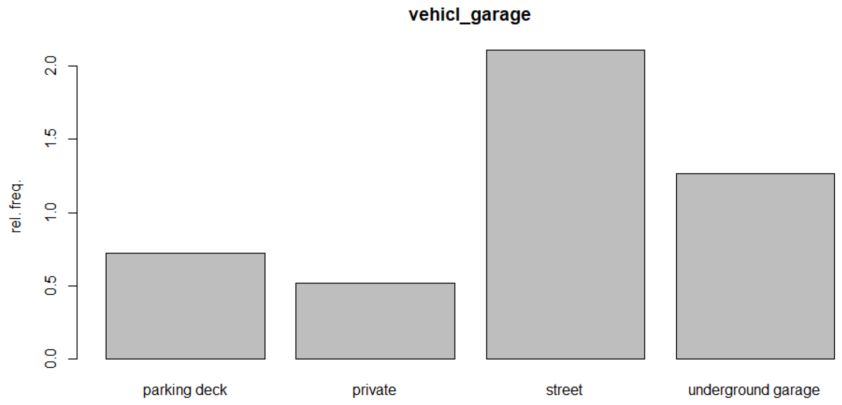

FIGURE 3.13 – Diagramme de fréquences relatives associé au moyen de garage de la voiture assurée

Enfin, à l'aide de la figure 3.14, on remarque qu'aucune femme n'est présente dans le sous-segment détecté bien que cette variable ne fasse pas partie des règles qui déterminent la boîte.

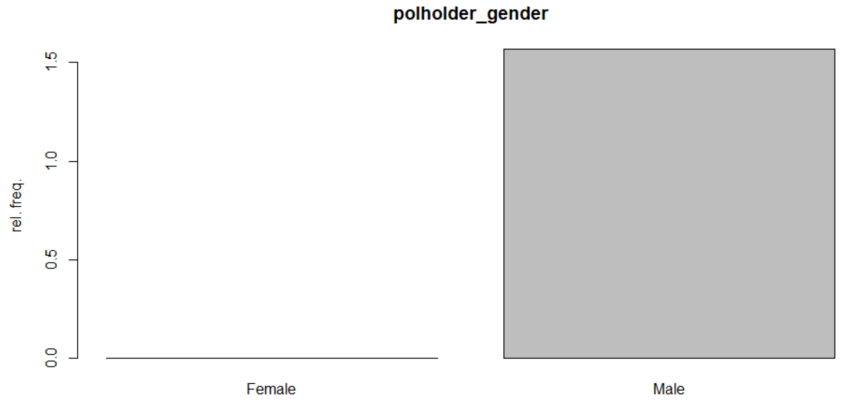

FIGURE 3.14 – Diagramme de fréquences relatives associé au genre de l'assuré

A l'aide du diagramme de fréquences relatives en fonction de l'ELR, on remarque que globalement les ELR des contrats sélectionnés sont inférieurs à 1,75. En revanche, les contrats dont l'ELR est compris entre 2 et 2,5 sont sur-représentés. Ce modèle pourrait donc être amélioré.

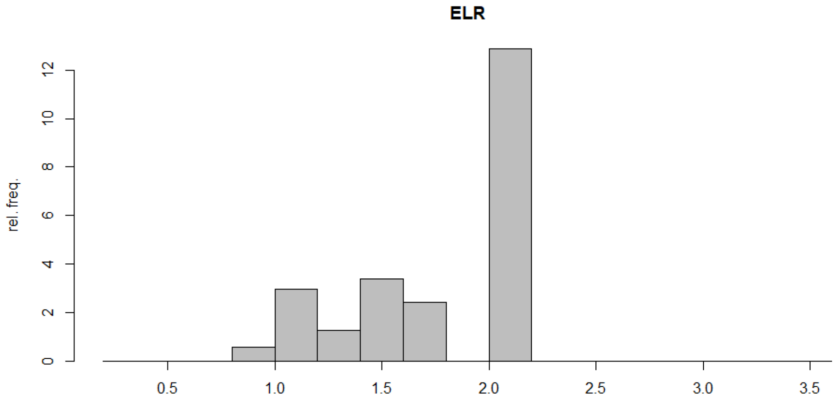

FIGURE 3.15 – Diagramme de fréquences relatives associé à l'ELR de l'assuré

Dissimilarité entre boîtes :

Dissimilarité	Segment 1	Segment 2	Segment 30	Segment 40
Segment 1		$2,146574 \times 10^{-2}$	$2,979618 \times 10^{-1}$	$2,025152 \times 10^{-2}$
Segment 2			$2,37771 \times 10^{-1}$	$3,586297 \times 10^{-2}$
Segment 30				$7,059844 \times 10^{-2}$

TABLE 3.2 – Étude des dissimilarités entre les différentes boîtes

A travers le tableau 3.2, on remarque que les deux premières boîtes obtenues sont proches d'être emboîtées. En revanche, la boîte 1 et la boîte 30 par exemple présentent moins de similarités se traduisant par un coefficient plus élevé.

3.5 Conclusions

L'algorithme PRIM nous a ainsi permis de détecter des sous-populations de notre portefeuille global pour lesquelles une attention toute particulière peut être portée. En effet, ces segments présentent des indicateurs de rentabilité très éloignés de la moyenne et permettent donc de revoir les positions de l'assureur quant à ces contrats. Dans le but d'adapter la stratégie tarifaire de l'assureur, l'analyse de la sensibilité de ces segments au prix sera un outil tout aussi déterminant que nous verrons dans la prochaine partie.

4 Modélisation de l'élasticité des taux de résiliation au prix

Après avoir modélisé et donc prédit la probabilité de résiliation des individus de notre portefeuille, il est intéressant de comprendre comment cette probabilité est impactée par des chocs tarifaires. Ainsi, nous allons dans cette section étudier l'élasticité au prix de la probabilité de résiliation en se concentrant notamment sur les segments déviants détectés lors de la partie précédente.

4.1 L'élasticité au prix

L'élasticité est une notion couramment utilisée en microéconomie de l'assurance ou en économie en général. Elle correspond à la variation relative de la demande d'un bien lorsque le prix de celui-ci varie d'une quantité infinitésimale. Elle s'obtient ainsi par :

$$\epsilon = \frac{\Delta D}{D} \times \frac{P}{\Delta P} \tag{4.1}$$

où $\frac{\Delta D}{D}$ représente la variation relative de la demande et $\frac{\Delta P}{P}$ celle du prix.

A l'aide de cette formule, on peut facilement remarquer que le signe de l'élasticité au prix dépend de la monotonie de la fonction de demande. Pour la grande majorité des produits, la fonction de demande décroît avec le prix. Par suite, pour la grande majorité des produits, l'élasticité au prix est positive. Ce qui veut dire qu'une hausse des prix induirait une baisse de la demande. Le cas où l'élasticité est négative impliquerait une hausse de la demande avec le prix. Ce cas caractérise les produits de luxe, catégorie dans laquelle ne rentre pas, à priori, un produit d'assurance IARD.

Dans cette partie, nous allons nous intéresser à la sensibilité des assurés au prix. Il sera donc intéressant d'estimer l'élasticité du taux de résiliation à la prime payée

aussi appelée cotisation. Cette élasticité correspond donc à la variation relative du taux de résiliation lorsque la cotisation annuelle de l'assuré varie d'une quantité infinitésimale. On obtiendra alors un indicateur de l'impact sur le taux d'une légère augmentation ou diminution de la prime. Une élasticité égale à ϵ indiquera que pour une hausse de 1% du prix nous aurons une augmentation de ϵ % de la résiliation.

Avec des notations adaptées à notre étude, on notera :

$$\epsilon(P) = \frac{\partial \pi(P)}{\pi(P)} \times \frac{P}{\partial P} \tag{4.2}$$

où $\pi(P)$ représente le taux de résiliation du contrat d'assurance, qui est une fonction de la prime P.

4.2 Modélisation de l'élasticité

En s'appuyant sur les travaux de Fontaine [10] et de Markaoui [7], une modélisation de l'élasticité a pu être mise en oeuvre. En effet, l'objectif étant d'estimer la sensibilité d'un assuré au prix, l'introduction de variables qualitatives du modèle de résiliation croisées avec la cotisation permet de mesurer l'impact de la prime sur le taux de résiliation.

On rappelle ici (voir équation 2.6) que la probabilité de résiliation peut s'écrire :

$$\pi = \frac{\exp(X\beta)}{1 + \exp(X\beta)} \tag{4.3}$$

Ainsi, en scindant les variables explicatives du modèle de résiliation en deux et selon leurs effets, on peut écrire :

— L'effet relatif aux caractéristiques de l'assuré sans considération du prix :

$$\alpha(x_1, ..., x_k) = a_0 + \sum_{l=1}^{k} a_l(i_l)\mathbb{1}_{x_i=i_l} \tag{4.4}$$

Ce coefficient ne dépend pas de la prime payée P et correspond donc uniquement aux caractéristiques de l'assuré.

— L'effet relatif au prix :

$$\beta(x_1, ..., x_k) = b_0 + \sum_{l=1}^{k} b_l(i_l)\mathbb{1}_{x_i=i_l} \tag{4.5}$$

A l'inverse, ce coefficient est lié à la fois à la prime payée P mais aussi aux caractéristiques de l'assuré.

où :
- a_0 est le coefficient associé à la constante
- $a_1, ..., a_k$ sont les coefficients associés aux variables qualitatives x_l
- b_0 est le coefficient associé au niveau de prime P
- $b_1, ..., b_k$ sont les coefficients associés au prix P suivant les modalités des variables qualitatives x_l.

Or, la probabilité de résiliation d'un assuré peut alors s'écrire comme suit :

$$1 - \hat{\pi}_i = \frac{1}{1 + e^{\left(a_0 + a_1(i_1)\mathbb{1}_{x_i=i_1} + ... + a_k(i_k)\mathbb{1}_{x_i=i_k} + P(b_0 + b_1(i_1)\mathbb{1}_{x_i=i_1} + ... + b_k(i_k)\mathbb{1}_{x_i=i_k})\right)}} \tag{4.6}$$

Donc en utilisant les notations α et β, définies plus haut, on obtient :

$$1 - \hat{\pi}(P) = \frac{1}{1 + \exp(\alpha + \beta P)} \tag{4.7}$$

Ce qui nous permet d'écrire la probabilité de résiliation d'un assuré comme :

$$\hat{\pi}(P) = \frac{\exp(\alpha + \beta P)}{1 + \exp(\alpha + \beta P)} \tag{4.8}$$

Or, on rappelle qu'on peut estimer l'élasticité par :

$$\hat{\epsilon}(p) = \frac{\partial \hat{\pi}(P)}{\partial P} \times \frac{P}{\hat{\pi}(P)} \tag{4.9}$$

Donc à partir de cette écriture et de l'équation 4.8, on a :

$$\frac{\partial \hat{\pi}(P)}{\partial P} = \frac{\beta \exp(\alpha + \beta P)}{(1 + \exp(\alpha + \beta P))^2} = \frac{\beta\ \hat{\pi}(P)}{1 + \exp(\alpha + \beta P)} \tag{4.10}$$

Puis, en remarquant que :

$$\frac{1}{1 + \exp(\alpha + \beta P)} = 1 - \frac{\exp(\alpha + \beta P)}{1 + \exp(\alpha + \beta P)} = 1 - \hat{\pi}(P) \tag{4.11}$$

On obtient alors :

$$\hat{\epsilon}(P) = \frac{\partial \hat{\pi}(P)}{\partial P} \times \frac{P}{\hat{\pi}(P)} = \beta \hat{\pi}(P)(1 - \hat{\pi}(P)) \times \frac{P}{\hat{\pi}(P)} \tag{4.12}$$

Ce qui nous permet finalement de définir l'élasticité du taux de résiliation par rapport au prix par :

$$\hat{\epsilon}(P) = \beta P(1 - \hat{\pi}(P)) \tag{4.13}$$

En pratique, il est habituel de considérer que l'élasticité de la probabilité de résiliation au prix doit être positive : plus la cotisation augmente et plus l'assuré aura tendance à résilier son contrat d'assurance ; un contrat d'assurance automobile n'est pas, par nature, un bien de luxe. En revanche, des cas particuliers peuvent exister : certains profils peuvent être attachés à leur compagnie et ainsi privilégier leurs offres à celles de leurs concurrents sans trop de considération au prix.

L'équation 4.13 nous permet de remarquer plusieurs choses :

- Le signe de l'élasticité est déterminé par le coefficient β. En effet, on peut aisément considérer que la prime P ainsi que la probabilité de rétention en portefeuille $1 - \hat{\pi}(P)$ sont des grandeurs positives.
- En revanche, l'intensité de l'élasticité sera elle déterminée par l'ensemble des grandeurs présentes dans l'équation.

4.2.1 Modèle GLM

A nouveau, un modèle GLM a été implémenté. Ce dernier se base sur le premier modèle GLM de la partie 2.4 dont le but était de modéliser le taux de résiliation. En revanche, la variable croisée $prem_last \times polholder_age$ a été ajoutée. Les résultats sont disponibles en annexe D.

A partir de ce modèle, et de la formule 4.13 il nous est alors possible de calculer l'élasticité souhaitée. En effet, la régression logistique permet d'obtenir les coefficients α et β nécessaires au calcul de l'élasticité. L'élasticité pourra alors être étudiée pour différents profils, afin de mieux identifier les segments du portefeuille sensibles au prix.

4.2.2 Élasticité globale du portefeuille

A l'aide de notre modélisation, nous avons pu évaluer l'élasticité globale de notre portefeuille d'assurance. On obtient ainsi une élasticité globale égale à 0,128. Cela signifie ainsi que si le niveau de prime augmentait de 10% l'année suivante, le taux de résiliation des assurés augmenterait de 1,28% et passerait ainsi de 12,81% à 14,09%.

4.2.3 Élasticité au prix selon les variables explicatives

A partir de la régression mise en place, nous pouvons analyser l'élasticité de certains profils :

— Élasticité selon le genre de l'assuré :

Modalité	Prime moyenne	Élasticité
Male	378	0.131
Female	385	0.124

TABLE 4.1 – Élasticités moyennes selon le genre de l'assuré

Ce tableau (4.1) nous permet de remarquer que ce sont les hommes les plus sensibles à l'évolution des prix. Leur élasticité est en effet supérieure à celle des femmes pour des primes payées équivalentes.

— Élasticité selon l'ancienneté du contrat :

Modalité	Prime moyenne	Élasticité
0	424	0.120
1	370	0.110
2	398	0.126
3	403	0.138
4	364	0.136
5	299	0.124
6	343	0.168
7	318	0.158
8	316	0.165
9	294	0.161
10	277	0.161
11 et plus	281	0.164

TABLE 4.2 – Élasticités moyennes selon l'ancienneté du contrat

A l'aide du tableau 4.2, on remarque une tendance à la hausse : plus le contrat est ancien, plus l'assuré est sensible au prix.

— Élasticité selon l'âge de l'assuré :

Modalité	Prime moyenne	Élasticité
<30 ans	574	0.077
Entre 31 et 40 ans	370	0.044
Entre 41 et 50 ans	360	0.206
Entre 51 et 60 ans	333	0.200
Entre 61 et 70 ans	289	0.176
Entre 71 et 80 ans	291	0.180
Plus de 80 ans	321	0.200

TABLE 4.3 – Élasticités moyennes selon l'âge de l'assuré

Ce tableau (4.3) nous indique qu'il existe une tendance croissante de l'élasticité par rapport à l'âge de l'assuré : plus l'assuré vieillit et plus il est sensible au prix qu'on lui propose. En effet, un jeune conducteur n'a pas vraiment de pouvoir de négociation. A l'inverse, un conducteur plus expérimenté peut s'attendre à un tarif plus faible et est donc sensible à ce qu'il paye.

4.3 Étude des élasticités des segments déviants identifiés par l'algorithme PRIM

Pour finir, nous allons nous intéresser à l'intensité de l'élasticité au prix pour les profils identifiés grâce à l'algorithme PRIM. La différence est donc notable à travers les figures 4.1 et 4.2 : certains segments sont très élastiques (segments sous-tarifés 1 et 3 notamment ainsi que les segments sur-tarifés 2, 3, 4) tandis que d'autres sont moins sensibles à une augmentation de prix. C'est le cas des segments sous-tarifés 4 ainsi que des segments sur-tarifés 1 et 5.

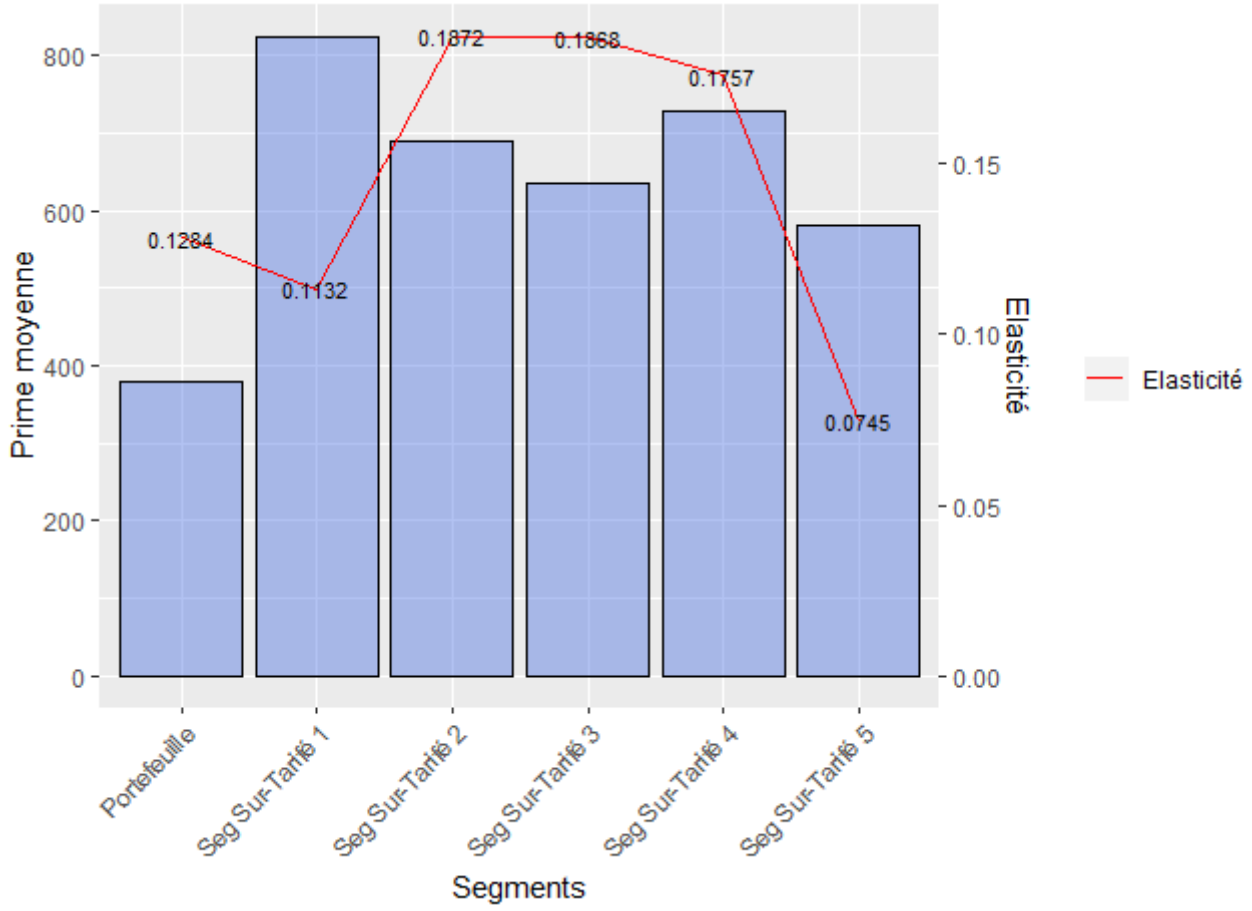

FIGURE 4.1 – Élasticité des segments sur-tarifés détectés par l'algorithme PRIM

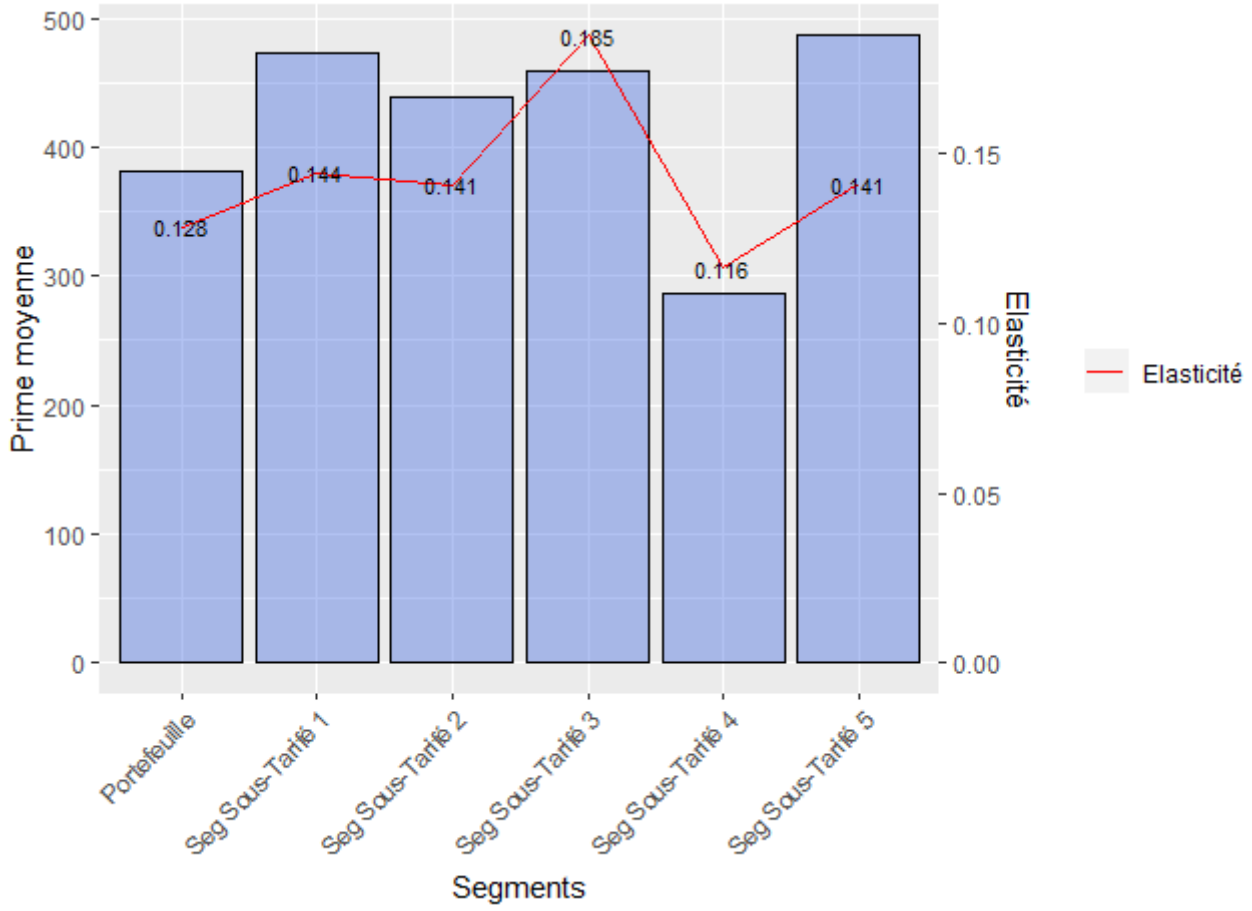

FIGURE 4.2 – Élasticité des segments sous-tarifés détectés par l'algorithme PRIM

L'étude de l'élasticité est donc indispensable pour mieux comprendre les flux d'un portefeuille d'assurance ou encore pour anticiper les conséquences de majorations de primes. Dans la partie suivante, nous nous replacerons à l'échelle du portefeuille d'assurance. En mettant à profit les modèles précedemment implémentés, nous chercherons à optimiser le profit total de l'assureur en respectant des contraintes de volume de portefeuille.

5 Optimisation tarifaire

L'optimisation est une discipline mathématique cherchant à résoudre des problèmes de minimisation ou de maximisation d'une fonction sur un ensemble. La résolution d'un tel problème peut être analytique ou numérique et pour cela, une étape préalable de modélisation du problème est nécessaire.

De nombreux domaines d'activité utilisent l'optimisation. On peut citer par exemple l'économie, l'industrie, l'ingénierie, les statistiques. En effet, si un problème peut être décrit par un modèle mathématique, il peut alors être optimisé. Le résultat obtenu dépendra de la pertinence du modèle choisi, du choix de la variable à optimiser ainsi que de l'efficacité de l'algorithme utilisé.

5.1 Objectifs de l'optimisation tarifaire

Après avoir détecté des sous-populations pour lesquelles le contrat d'assurance n'est pas à l'équilibre, l'élasticité nous a permis de comprendre les enjeux d'une hausse de tarif. En effet, cette hausse impacterait directement le taux de résiliation et ne peut donc pas être réalisée sans étude préalable.

Dans cette nouvelle partie, l'objectif est donc de mettre en oeuvre un programme d'optimisation tarifaire permettant de quantifier le gain de l'assureur selon plusieurs scénarios.

Le programme d'optimisation se basera sur un modèle précédemment implémenté : le modèle de résiliation.

Nous serons alors en mesure de fournir des réponses aux questions suivantes :

— De quel pourcentage serait-il judicieux d'augmenter ou de diminuer la prime d'assurance pour augmenter la marge de l'assureur ?
— Sur quelles populations est-il possible d'augmenter le tarif sans trop impacter le nombre de résiliations annuelles ?

Dans notre cadre, nous disposons d'un portefeuille de N assurés sur lequel :
— Une prime commerciale PC_i est appliquée.
— Un modèle de résiliation est implémenté : la probabilité de maintien en portefeuille π_i étant fonction des critères de risques et du niveau de prime commerciale.

5.2 Présentation du problème

5.2.1 Programme d'optimisation

En s'appuyant sur les travaux de Smets [4], on écrit communément un programme d'optimisation sous la forme :

$$\begin{aligned} \max_{x_1,x_2,...,x_n} \quad & f(x_1, x_2, ..., x_n) \\ \text{s.c.} \quad & \textit{contrainte}(j) \; j = 1, ..., m \end{aligned} \tag{5.1}$$

Dans ce programme, la fonction f est appelée fonction objectif et le but sera de trouver les valeurs $(x_1^*, x_2^*, ..., x_n^*)$ pour lesquelles la fonction est maximale sous les m contraintes. On appellera alors cette solution un optimum.

Concernant les contraintes, elles peuvent se présenter sous différentes formes :
— Contraintes d'inégalités : $g_j(x_1, x_2, ..., x_n) \leq c_j$ pour tout $j = 1, ..., m$.
— Contraintes de non-négativité : $x_i \geq 0$ pour tout $i = 1, ..., n$.

Dans le premier cas, si on note $\tilde{x}^*$ la solution du programme d'optimisation, deux possibilités se présentent :
— Si $g_j(\tilde{x}^*) = c_j$ alors, on dit que la contrainte est saturée à l'optimum ;
— Sinon, on a $g_j(\tilde{x}^*) < c_j$ et la contrainte est dite non saturée.

On notera également $\mathscr{DA}$ le domaine admissible qui représente l'ensemble des points x pour lesquels toutes les contraintes sont satisfaites.

5.2.2 Adaptation du problème d'optimisation à notre cas d'usage

Le modèle auquel nous allons nous intéresser est le suivant :

$$\begin{aligned} \max_{PC_1,PC_2,...,PC_n} \quad & f(PC_1, PC_2, ..., PC_n) \\ \text{s.c.} \quad & \textstyle\sum_{i=1}^{n} \pi_i \geq Volume \\ & inf_i \leq PC_i \leq sup_i \\ \forall\, i = 1, ..., n & \end{aligned} \tag{5.2}$$

Les bornes inf et sup encadrant l'amplitude possible de chaque prime commerciale sont définies par exemple à partir de la prime commerciale actuelle :

$$PC_i^0 - 2\% \leq PC_i \leq PC_i^0 + 3\% \qquad \forall\, i = 1, ..., n$$

Les amplitudes de variations peuvent dépendre évidemment de la cible marketing, des niveaux de probabilité de maintien ou de résiliation.

On rappelle que le modèle de résiliation est un modèle logit et nous pouvons ainsi écrire la probabilité de résiliation d'un assuré i comme :

$$\pi_i = \frac{e^{X_i \beta_i}}{1 + e^{X_i \beta_i}} \tag{5.3}$$

où X représente les variables explicatives et β le coefficient obtenu par modèle GLM.

Si on note β_P le coefficient associé à la prime commerciale du modèle de résiliation, on peut aussi écrire :

$$\pi_i = \frac{e^{\sum_{j=0}^{p} x_{ij}\beta_j}}{1 + e^{\sum_{j=0}^{p} x_{ij}\beta_j}} = \frac{e^{\sum_{j=0}^{p-1} x_{ij}\beta_j} e^{\beta_P PC_i}}{1 + e^{\sum_{j=0}^{p} x_{ij}\beta_j} e^{\beta_P PC_i}} = \frac{\alpha_i e^{\beta_P PC_i}}{1 + \alpha_i e^{\beta_P PC_i}} \tag{5.4}$$

Ainsi, la fonction objectif à maximiser sur les primes commerciales (en désignant PC_i par x_i) est la suivante :

$$\begin{aligned} \max_{x} \quad & f(x) = \textstyle\sum_{i=1}^{n} \left[(x_i - PR_i) * \frac{\alpha_i e^{\beta_P x_i}}{1+\alpha_i e^{\beta_P x_i}} \right] \\ \text{s.c.} \quad & \textstyle\sum_{i=1}^{n} \frac{\alpha_i e^{\beta_P x_i}}{1+\alpha_i e^{\beta_P x_i}} \geq Volume \\ & inf_i \leq x_i \leq sup_i \\ \forall\, i = 1, ..., n & \end{aligned} \tag{5.5}$$

5.3 Le module Price Simulation du logiciel addactis® Pricing

Le logiciel addactis® Pricing donne la possibilité, à travers son troisième module dénommé *Price Simulation*, de concilier des modèles de tarification et de comportement client afin de déterminer le meilleur ajustement du risque.
Pour cela, plusieurs modèles doivent être importés :

- Un modèle de tarification permettant de calculer la prime de chaque individu ou bien un fichier comportant d'ores et déjà une variable prime ;
- Un ou des modèle(s) de comportement client permettant de créer de premiers indicateurs sur lesquels les études seront réalisées.

A la suite de cela, des indicateurs peuvent être créés sur lesquels les méthodes de simulation seront appliquées.

- La première méthode du logiciel, intitulée *Simulation*, consiste à tester plusieurs scénarios dans lesquels la variable prime est modifiée. Les scénarios testés sont guidés directement par l'utilisateur ;
- La seconde méthode, connue sous le nom *Standard*, consiste à chercher le meilleur ajustement tarifaire selon des contraintes sur le portefeuille d'assurés.

5.3.1 Méthode *Simulation*

La méthode *Simulation* a pour vocation de tester l'impact de différentes combinaisons de paramètres sur la rentabilité du portefeuille.

Dans un premier temps, l'utilisateur doit choisir les variables explicatives (et modalités) qu'il souhaite étudier. Des variables aussi bien nominales qu'ordinales peuvent être choisies. Dans le cas où il s'agit d'une variable ordinale, l'utilisateur est en mesure de créer des intervalles. Les modalités ou intervalles choisis permettront de faire varier la prime payée des individus du portefeuille. Prenons l'exemple d'un individu possédant une maison avec piscine et payant une cotisation égale à 100€ pour son assurance habitation. Si l'utilisateur choisit d'intégrer la modalité : "Oui" de la variable "Présence d'une piscine" et qu'il associe une variation de 5% à cette modalité, le modèle considérera que l'assuré paye une prime de 105€.

En effet et dans un second temps, l'utilisateur du logiciel indique les variations qu'il souhaite appliquer à la variable *Premium*. Ces dernières sont multiplicatives, cela signifie que si la variable *Premium* doit à la fois subir une variation de 25% et une seconde de 10%, la variation finale sera de $1,25 \times 1,1 = 1,375$ ce qui signifie une variation finale de 37,5%. L'utilisateur peut cibler les modalités sur lesquelles il

souhaite appliquer des variations.

Il est également possible de borner la variable *Premium* à l'aide des options *Premium inf* et *Premium sup*. Cela permet de s'assurer que les primes restent dans un certain intervalle malgré les variations appliquées.

Seront donc implémentées un nombre de simulations qui dépend du nombre de modalités impactantes choisi ainsi que du nombre de variations. Par exemple, si deux variables impactent la prime : *Sexe* et *Situation Maritale*, si ces deux variables ont respectivement deux et trois modalités et si 5 variations sont choisies : -5%, $-2{,}5\%$, 0%, $2{,}5\%$, 5%, alors 3125 simulations ($5 \times 5 \times 5 \times 5 \times 5$) seront lancées.

Résultats de la méthode *Simulation*

Lorsqu'une méthode *Simulation* est lancée, l'utilisateur obtient différents résultats. Tout d'abord, l'ensemble des simulations exécutées est retranscrit et pour chacune d'elle nous obtenons les variations de la variable *Premium* et la valeur de l'indicateur choisi.

Ensuite, un graphique est proposé représentant l'indicateur en abscisse et la prime en ordonnée. Ce dernier permet une visualisation rapide du meilleur scénario ou encore de la meilleure combinaison de variations. En effet, l'utilisateur sera à la recherche d'un point d'inflexion de la courbe formée par les résultats des différents scénarios.

5.3.2 Méthode *Standard*

La seconde méthode disponible au sein du logiciel addactis® Pricing vise à optimiser l'indicateur principal en ajustant la variable *Premium* et tout en respectant une contrainte de volume ainsi qu'un intervalle de fluctuations de la variable *Premium*.
En revanche, la méthode *Standard* n'est applicable que dans le cas où l'indicateur principal est quadratique. Cela signifie qu'il se présente sous la forme : $(Premium - Variable1) * Model(Premium - Variable2)$.

Trois paramètres doivent alors être renseignés :
- Le volume cible, volume minimal souhaité pour le portefeuille ;
- Le nombre maximum d'itérations du programme d'optimisation ;
- L'écart entre deux itérations, noté Epsilon.

Résultats de la méthode *Standard*

Les résultats de la méthode *Standard* sont les mêmes que ceux issus de la méthode *Simulation*.

Ainsi, nous verrons dans la partie suivante les fondements théoriques de la méthode *Standard* avant de mettre en application les deux méthodes sur notre base de données.

5.4 Cadre théorique de l'approche *Standard* du logiciel - Méthodes de résolution de programmes d'optimisation

5.4.1 Lagrangien

La résolution de programmes d'optimisation tels que nous les avons décrits dans une partie antérieure (voir équation 5.1) peut être réalisée grâce à la méthode du Lagrangien. Cette dernière est notamment décrite dans l'article de Casanova [8].

Considérons pour cela le programme d'optimisation suivant :

$$\begin{array}{ll} \max\limits_{\tilde{x}} & f(\tilde{x}) \\ \text{s.c.} & g_j(\tilde{x}) \leq c_j \quad \forall j = 1, ..., m \end{array} \tag{5.6}$$

où $\tilde{x} = (x_1, x_2, ..., x_n)$.

Nous définissons le Lagrangien du problème d'optimisation par :

$$\mathscr{L}(\tilde{x}, \tilde{\lambda}) = f(\tilde{x}) - \sum_{j=1}^{m} \lambda_j (g_j(\tilde{x}) - c_j) \tag{5.7}$$

où $\tilde{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)$ sont les multiplicateurs de Lagrange associés aux m contraintes. Cette application va nous permettre de trouver et d'étudier les extremums de $f(\tilde{x})$ sous les contraintes définies précedemment.

5.4.2 Conditions de qualification des contraintes

L'utilisation du Lagrangien nécessite, pour le programme d'optimisation sous contraintes prenant la forme d'inéquations, de vérifier l'une des conditions suivantes :

— La matrice Jacobienne de ces m fonctions contraintes que nous pouvons noter :

$$\left[\frac{\partial g_j(\tilde{x})}{\partial x_i}\right]_{i=1,...,n\ j=1,...,m}$$

est de rang m lorsqu'elle est évaluée à l'optimum $\tilde{x}$.

— Les fonctions contraintes g_j, $j = 1, ..., m$ sont toutes linéaires.

5.4.3 Conditions nécessaires d'optimalité de Karush, Kuhn et Tucker

Dans un problème d'optimum avec contraintes d'inégalités, on suppose que :

— f, g_j $\forall\ j = 1, ..., m$ sont continues et différentiables,

— L'hypothèse de qualification des contraintes est vérifiée en un point $\tilde{x}^*$ du domaine admissible $\mathscr{DA}$.

Alors, si $\tilde{x}^*$ est un optimum local, il existe un vecteur $\tilde{\lambda}^*$ pour le Lagrangien qui vérifie :

— $\tilde{\lambda}^* \geq 0\ \forall\ j = 1, ..., m$

— $\tilde{\lambda}^*(g_j(\tilde{x}^*) - c_j) = 0 \ \Leftrightarrow\ g_j(\tilde{x}^*) = c_j$ ou $\tilde{\lambda}^* = 0\ \forall\ j = 1, ..., m$ cela signifie que le multiplicateur de Lagrange est nul pour les contraintes non saturées.

—
$$\frac{\partial \mathscr{L}(\tilde{x}^*, \tilde{\lambda}^*)}{\partial \lambda_j} \geq 0 \ \Leftrightarrow\ g_j(\tilde{x}^*) \leq c_j\ \forall\ j = 1, ..., m \tag{5.8}$$

—
$$\frac{\partial \mathscr{L}(\tilde{x}^*, \tilde{\lambda}^*)}{\partial x_i} = 0 \ \Leftrightarrow\ \frac{\partial f(\tilde{x}^*)}{\partial x_i} - \sum_{j=1}^{m} \lambda_j^* \frac{\partial g_j(\tilde{x}^*)}{\partial x_i} = 0\ \forall\ i = 1, ..., n \tag{5.9}$$

Ces résultats sont notamment utilisés par Grenet [9] dans son vadémécum sur la méthode de résolution des programmes d'optimisation statique.

5.4.4 Application au cadre quadratique

Considérations théoriques

Dans le cas quadratique, on écrit le programme d'optimisation comme :

$$\begin{array}{ll} \max\limits_{\tilde{x}^*} & f(x) = \frac{1}{2}x^T Q x + L^T x \text{ avec } \mathrm{x}^T = (x_1, ..., x_n) \\ \text{s.c.} & g_j(x) \leq c_j\ \forall\ j = 1, ..., m \end{array} \tag{5.10}$$

où :

— Q est une matrice semi-définie négative de taille n*n ;

— L et x sont deux vecteurs colonnes à n composantes ;
— les contraintes s'écrivent sous la forme : $g(x) = Ax + b$. De ce fait, $Ax + b \leq 0$ avec A une matrice m*n et b un vecteur colonne à m composantes.

Le Lagrangien du problème s'écrit :

$$\mathscr{L}(x, \lambda) = \frac{1}{2} x^T Q x + L^T x - \lambda^T (Ax + b) \tag{5.11}$$

Puis, si x^* est un optimum local, d'après les conditions de Karush, Kuhn et Tucker, il satisfait :

$$\frac{\partial \mathscr{L}(x^*, \lambda)}{\partial x_i} = 0 \tag{5.12}$$

avec $\lambda \geq 0$.

Par un calcul détaillé en annexe E, on remarque que chercher la solution au problème ci-dessous :

$$\begin{array}{ll} \max\limits_{\tilde{x}^*} & f(x) = \frac{1}{2} x^T Q x + L^T x \text{ avec } \mathrm{x}^T = (x_1, ..., x_n) \\ \text{s.c.} & g_j(x) \leq c_j \ \forall \ j = 1, ..., m \end{array} \tag{5.13}$$

revient de façon équivalente à résoudre le programme suivant que nous appelons programme dual :

$$\begin{array}{ll} \min\limits_{\lambda} & f(x) = \frac{1}{2} \lambda^T G \lambda + g^T \lambda \\ \text{s.c.} & \lambda_j \geq 0 \ \forall \ j = 1, ..., m \end{array} \tag{5.14}$$

avec $G = -AQ^{-1}A^T$ et $g^T = L^T Q^{-1} A^T - b^T$.

Le problème dual permettant de trouver λ^*, nous en déduisons alors x^* :

$$x^* = Q^{-1}(A^T \lambda^* - L) \tag{5.15}$$

Ensuite, pour résoudre le problème dual, l'algorithme de Gauss-Seidel peut être utilisé. Cet algorithme est notamment décrit par Buffat [11] :

— Initialisation : $\lambda^{[0]} = 0$
— Pour $i = 1, ..., m$, on calcule :

$$\lambda_i^{[k+1]} = max\left(0; u_i^{[k+1]}\right)$$

où :

$$u_i^{[k+1]} = \frac{-1}{G_{ii}} \left(\sum_{j=1}^{i-1} G_{ij} \lambda_j^{[k+1]} + \sum_{j=i+1}^{m} G_{ij} \lambda_j^{[k]} + g_i \right)$$

— Fin quand $|\lambda^{[k+1]} - \lambda^{[k]}| \leq \epsilon$

Exemple de programme quadratique adapté à notre cas d'usage :

Le programme que l'on souhaite optimiser est le suivant :

$$\begin{aligned} \max_x \quad & f(x) = \sum_{i=1}^{n} \left[(x_i - PR_i) * \frac{\alpha_i e^{\beta P x_i}}{1+\alpha_i e^{\beta P x_i}} \right] \\ \text{s.c.} \quad & \sum_{i=1}^{n} \frac{\alpha_i e^{\beta P x_i}}{1+\alpha_i e^{\beta P x_i}} \geq Volume \\ & PC_i^{inf} \leq x_i \leq PC_i^{sup} \\ \forall\, i = 1, ..., n & \end{aligned} \tag{5.16}$$

Pour exprimer le problème sous forme quadratique, chaque ratio $\frac{\alpha_i e^{\beta P x_i}}{1+\alpha_i e^{\beta P x_i}}$ est approximé linéairement sur l'intervalle $[PC_i^{inf}; PC_i^{sup}]$. Ainsi l'approximation linéaire s'effectue sur un intervalle restreint, imposé par la contrainte.

Deux valeurs doivent donc être calculées à l'initialisation du programme pour chaque assuré i : la pente a_i et l'ordonnée à l'origine b_i. Ces dernières sont liées à la contrainte de prime PC_i de la façon suivante :

$$a_i = \frac{\frac{\alpha_i e^{\beta PC_i^{sup}}}{1+\alpha_i e^{\beta PC_i^{sup}}} - \frac{\alpha_i e^{\beta PC_i^{inf}}}{1+\alpha_i e^{\beta PC_i^{inf}}}}{PC_i^{sup} - PC_i^{inf}}$$

$$b_i = \frac{\left(\frac{\alpha_i e^{\beta PC_i^{sup}}}{1+\alpha_i e^{\beta PC_i^{sup}}} + \frac{\alpha_i e^{\beta PC_i^{inf}}}{1+\alpha_i e^{\beta PC_i^{inf}}} \right) - a_i (PC_i^{sup} + PC_i^{inf})}{2}$$

Cela permet d'écrire le programme :

$$\begin{aligned} \max_x \quad & f(x) = \sum_{i=1}^{n} \left[(x_i - PR_i) * (a_i x_i + b_i) \right] \\ \text{s.c.} \quad & \sum_{i=1}^{n} (a_i x_i + b_i) \geq Volume \\ & PC_i^{inf} \leq x_i \leq PC_i^{sup} \\ \forall\, i = 1, ..., n & \end{aligned} \tag{5.17}$$

Puis :

$$\begin{aligned} \max_x \quad & f(x) = \sum_{i=1}^{n} \left[a_i x_i^2 + (b_i - a_i PR_i) x_i \right] \\ \text{s.c.} \quad & \sum_{i=1}^{n} (-a_i x_i - b_i) + Volume \leq 0 \\ & PC_i^{inf} - x_i \leq 0 \;\; \forall\, i = 1, ..., n \\ & x_i - PC_i^{sup} \leq 0 \;\; \forall\, i = 1, ..., n \end{aligned} \tag{5.18}$$

Enfin, si on veut que cela corresponde au programme théorique que nous avions écrit précédemment (5.10), les matrices Q, L, A et b sont les suivantes :

$$Q = Diag(2a_i)_{i=1,\dots,n}$$

$$L^T = (b_1 - a_1 PR_1, ..., b_n - a_n PR_n)$$

$$A = \begin{pmatrix} -a_1 & \cdots & -a_n \\ 1 & \cdots & 0 \\ -1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 1 \\ 0 & \cdots & -1 \end{pmatrix}$$

$$b = \begin{pmatrix} Volume - \sum_{i=1}^{n} b_i \\ -PC_1^{sup} \\ PC_1^{inf} \\ \vdots \\ -PC_n^{sup} \\ PC_n^{inf} \end{pmatrix}$$

Enfin, on peut appliquer l'algorithme de Gauss-Seidel en posant :

$$G = -AQ^{-1}A^T$$

et

$$g^T = (L^T Q^{-1} A^T - b^T)$$

5.5 Résultats sur la base de données

5.5.1 Import des données et modèles

Dans cette dernière partie, l'objectif est de mettre en pratique les considérations théoriques énoncées jusqu'à présent. Pour cela, nous nous aiderons du logiciel addactis® Pricing et utiliserons les résultats du modèle de résiliation précédemment

implémenté.

Dans un premier temps, les données de notre portefeuille ont été importées dans le logiciel. Ces données comportent bien entendu les variables descriptives explicatives qui ont été utilisées dans notre modèle de résiliation : l'âge, le genre de l'assuré, le type de garage de la voiture ou encore la prime payée pour le contrat d'assurance pour ne citer que ces exemples. Ensuite, le modèle de résiliation décrit dans l'annexe B est importé.

5.5.2 Définition des indicateurs

A la suite de cela, nous définissons la prime à l'étude ainsi que l'indicateur principal de notre étude par les encarts ci-dessous (figure 5.1) :

FIGURE 5.1 – Définition de la prime et de l'indicateur principal dans le logiciel addactis® Pricing

C'est cet indicateur qui sera calculé pour chacun des scénarios choisis dans la partie *Simulation* ou qui sera optimisé sous contraintes dans la partie *Standard*.

Un second indicateur a également été implémenté (figure 5.2) :

FIGURE 5.2 – Définition de l'indicateur secondaire dans le logiciel addactis® Pricing

Cet indicateur représente le volume du portefeuille, c'est-à-dire le nombre probable d'assurés n'ayant pas résilié leur contrat d'assurance.

5.5.3 Méthode *Simulation*

Nous allons dans cette partie tester plusieurs scénarios de majoration ou de réduction de primes à nos individus détectés dans les boîtes de l'algorithme PRIM.

Commençons par la première boîte sous-tarifée obtenue. Pour rappel, ce premier segment représente 1,0104% de la population totale et se compose des assurés ayant une ancienneté strictement inférieure à 5 ans, dont le véhicule est âgé de moins de 17 ans, vivant dans les régions 2, 5, 6, 9, 10, 12 ou 13 et dont le CRM est en hausse. Sur ce segment, l'ELR moyen est de 1,4562 et est donc 1,52 fois supérieur à l'ELR moyen du portefeuille. Ce paramétrage se traduit de la façon suivante dans le logiciel (figure 5.3) :

Factors	-15,0...	-12,0...	-9,00...	-6,00...	-3,00...	0,00 %	3,00 %	6,00 %	9,00 %	12,0...	15,0...	18,0...	21,0...	24,0...	26,0...
policy_age															
[0, 5[	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
[5, ∞[	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
...															
vehicl_age															
[0, 17[	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
[17, ∞[	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
...															
vehicl_region															
[Reg2, Reg5, Reg6, Reg9, Reg10, Reg12, Reg13]	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
...															
polholder_BMCevol															
[up]	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
...															

FIGURE 5.3 – Paramétrage des simulations sur la première boîte dans le logiciel addactis® Pricing

A gauche de la figure se trouvent les critères sur lesquels nous appliquons les majorations ou minorations de prime. Les critères ayant été déterminants dans la définition de la boîte ont été ciblés. Quant aux majorations ou minorations en question, elles sont affichées en haut de la figure. Dans la situation présentée, l'application de ce paramétrage induit la simulation de 50 625 scénarios.

Nous obtenons un graphique qui est le suivant (figure 5.4) :

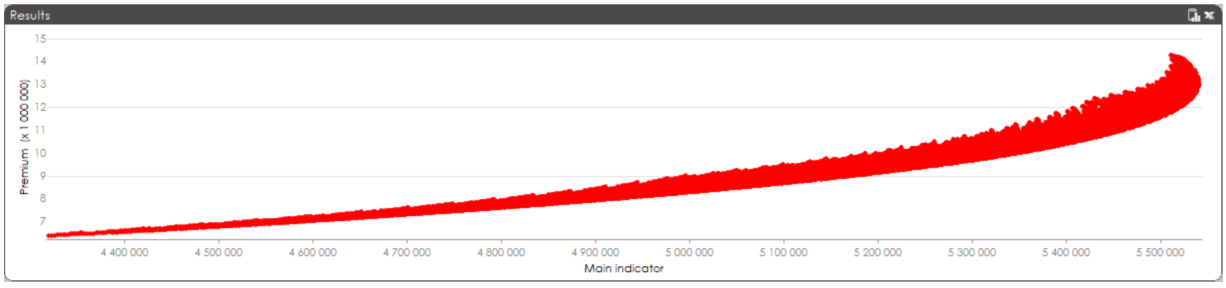

FIGURE 5.4 – Résultat graphique des simulations réalisées sur la première boîte sous-tarifée

Chaque point de ce graphique représente le résultat d'une simulation. Plusieurs conclusions peuvent être tirées :

— Dans un premier temps, le "Main indicator", qui représente dans notre cas le profit de l'assureur, croît avec la prime demandée aux assurés.
— On atteint ensuite un point d'inflexion et, suite à ce point, cette croissance n'est plus vraie : la prime demandée aux assurés étant trop élevée, ils vont faire le choix de résilier leur contrat. Le nombre de contrat en portefeuille baisse et le profit de l'assureur est directement impacté.

Il nous est alors possible de lier chaque point du graphique ci-dessus au scénario qui lui correspond. En triant notre tableau de résultats par ordre croissant du "Main indicator", nous obtenons les résultats ci-dessous (figure 5.5) :

policy_age	vehicl_age	vehicl_re...	polholder...			
[0, 5[	[0, 17[	[Reg2, Re...	[up]	Main indicator ▼	Premium	Vol
15,00 %	26,00 %	26,00 %	-9,00 %	5 542 001	12 975 011	12 904
15,00 %	26,00 %	26,00 %	-6,00 %	5 541 909	12 992 508	12 894
15,00 %	26,00 %	26,00 %	-12,00 %	5 541 890	12 957 514	12 914
15,00 %	26,00 %	26,00 %	-3,00 %	5 541 630	13 010 005	12 884
15,00 %	26,00 %	24,00 %	-9,00 %	5 541 623	12 890 976	12 952
15,00 %	26,00 %	24,00 %	-6,00 %	5 541 574	12 908 361	12 942
15,00 %	26,00 %	26,00 %	-15,00 %	5 541 556	12 940 016	12 925
18,00 %	26,00 %	24,00 %	-9,00 %	5 541 489	13 173 600	12 786
18,00 %	26,00 %	24,00 %	-12,00 %	5 541 473	13 155 816	12 797
15,00 %	26,00 %	24,00 %	-12,00 %	5 541 467	12 873 591	12 963

FIGURE 5.5 – Tableau des simulations réalisées sur la première boîte sous-tarifée

La première ligne indique qu'il faudrait augmenter le tarif des assurés répondant aux critères de la première boîte de 66% afin de maximiser le profit global. Il s'agit là de la solution optimale détectée par l'algorithme d'optimisation mais elle n'est pas envisageable en réalité. Bien que la majoration de 66% ne s'appliquerait qu'à 1% du portefeuille, cette augmentation drastique ferait fuir bon nombre d'entre eux et une réglementation existe la plupart du temps : les majorations de tarif sont souvent encadrées et précisées dans le contrat d'assurance. Cette majoration permettrait de conserver 17 190 assurés dans le portefeuille soit 75%. Ce scénario éliminerait ainsi l'ensemble des mauvais segments au sens de l'algorithme PRIM.

Si nous nous intéressons au second segment sous-tarifé, nous observons un comportement semblable. Pour rappel, ce segment représente 1,62% du portefeuille et est constitué des assurés de moins de 74 ans, dont le CRM est en hausse, qui vivent

en région 3, 4, 7, 8 ou 9 et pour lesquels le conducteur principal de la voiture n'est ni un apprenti, ni un jeune conducteur, ni un conducteur sans lien avec l'assuré. Sur ce segment, l'ELR moyen vaut 1,40.

L'algorithme nous indique qu'il serait optimal, au sens du profit de l'assureur, d'augmenter la prime d'assurance de ces assurés de 49% (figure 5.6). Là encore un tel chiffre n'est pas envisageable.

polholder...	polholder...	vehicl_re...	polholder...			
[0, 74[	[up]	[Reg3, Re...	[all drivers...	Main indicator ▼	Premium	Vol
20,00 %	-10,00 %	15,00 %	20,00 %	5 703 517	13 408 886	12 574
20,00 %	-11,00 %	15,00 %	20,00 %	5 703 516	13 402 906	12 577
20,00 %	-12,50 %	15,00 %	20,00 %	5 703 463	13 393 936	12 583
20,00 %	-7,50 %	15,00 %	20,00 %	5 703 411	13 423 836	12 565
20,00 %	-10,00 %	14,00 %	20,00 %	5 703 284	13 334 939	12 618
20,00 %	-11,00 %	14,00 %	20,00 %	5 703 270	13 328 993	12 622
20,00 %	-15,00 %	15,00 %	20,00 %	5 703 237	13 378 985	12 592
20,00 %	-7,50 %	14,00 %	20,00 %	5 703 207	13 349 803	12 610
20,00 %	-12,50 %	14,00 %	20,00 %	5 703 199	13 320 075	12 627
20,00 %	-5,00 %	15,00 %	20,00 %	5 703 153	13 438 786	12 556

FIGURE 5.6 – Tableau des simulations réalisées sur la seconde boîte sous-tarifée

En revanche, on remarque ici que le premier scénario permet d'obtenir un profit très élevé (proche de 5 704 000). Ceci s'obtient aux dépens du volume du portefeuille puisqu'il chute fortement passant à 12 600.

Il peut aussi être pertinent de s'intéresser aux segments sur-tarifés : le premier est composé des nouveaux contrats dont le CRM est en baisse, souscrits par des assurés de moins de 28 ans, ayant acheté un véhicule de plus de 10 ans et qui aujourd'hui a moins de 18 ans. Ces contrats doivent être souscrits par le conducteur lui-même et dans une région différente des régions 3, 9 et 12. Sur ce segment (1,11% de la population), l'ELR moyen vaut 0,798 et est donc plus de 1,2 fois inférieur à la moyenne globale. De nombreux critères définissent ce segment et cela a une influence sur le nombre de simulations. En effet, le nombre de simulations dépend à la fois des modalités sur lesquelles appliquer la majoration mais aussi des facteurs de majoration ou de minoration. Dans notre exemple, ce sont 7 modalités qui doivent intervenir dans la modification de la prime. Après plusieurs simulations intermédiaires, nous obtenons les résultats ci-dessous (figure 5.7 et 5.8) :

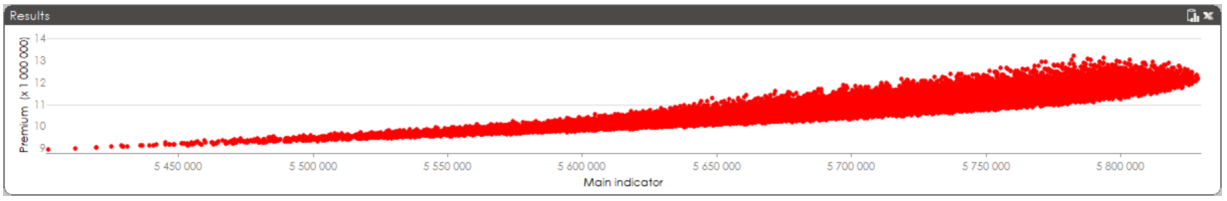

FIGURE 5.7 – Résultat graphique des simulations réalisées sur la première boîte sur-tarifée

policy_age	polholder...	polholder...	vehicl_ag...	vehicl_age	polholder...	vehicl_re...			
[0, 1[	[down]	[0, 28[	[10, ∞[	[0, 18[	[same]	[Reg1, Re...	Main indicator ▼	Premium	Vol
12,00 %	-25,00 %	-40,00 %	3,00 %	24,00 %	21,00 %	24,00 %	5 828 985	12 177 339	13 316
15,00 %	-25,00 %	-40,00 %	0,00 %	24,00 %	21,00 %	24,00 %	5 828 882	12 147 478	13 336
15,00 %	-25,00 %	-40,00 %	3,00 %	24,00 %	21,00 %	24,00 %	5 828 576	12 309 660	13 237
12,00 %	-25,00 %	-40,00 %	0,00 %	24,00 %	21,00 %	24,00 %	5 828 422	12 017 547	13 414
18,00 %	-25,00 %	-40,00 %	0,00 %	24,00 %	21,00 %	24,00 %	5 828 070	12 277 409	13 258
9,00 %	-25,00 %	-40,00 %	3,00 %	24,00 %	21,00 %	24,00 %	5 828 064	12 045 019	13 396
18,00 %	-25,00 %	-40,00 %	-3,00 %	24,00 %	21,00 %	24,00 %	5 827 010	12 112 838	13 359
15,00 %	-25,00 %	-40,00 %	-3,00 %	24,00 %	21,00 %	24,00 %	5 826 950	11 985 296	13 436
18,00 %	-25,00 %	-40,00 %	3,00 %	24,00 %	21,00 %	24,00 %	5 826 915	12 441 980	13 159
9,00 %	-20,00 %	-40,00 %	0,00 %	24,00 %	21,00 %	24,00 %	5 826 881	12 271 759	13 246

FIGURE 5.8 – Tableau des simulations réalisées sur la première boîte sur-tarifée

Dans ce cas, l'algorithme d'optimisation indiquerait que la meilleure solution consisterait à diminuer la prime de nos individus de 4,5%. Avec cela, nous obtiendrions un profit proche de 5 830 000 et conserverions 13 000 assurés en portefeuille.

Ces résultats, obtenus à l'aide de la méthode *Simulation* ne sont donc pas pleinement satisfaisants. En effet, nous ne pouvons pas déterminer la majoration optimale exacte pour chaque modalité : le temps de calcul est considérable et les scénarios ne peuvent donc pas être multipliés indéfiniment.

5.5.4 Méthode *Standard*

Comme nous l'avons vu à travers la partie précédente, la méthode *Simulation*, qui est une méthode manuelle, a présentée des limites certaines. Une optimisation "automatique" a ainsi été mise en place à l'aide de la méthode *Standard*.

Dans un premier temps, nous avons voulu connaître quels étaient les indicateurs de notre portefeuille dans le cas où l'on conserverait la prime d'assurance. Pour cela,

nous avons donc implémenté une méthode *Simulation* pour laquelle toutes les variations de primes sont égales à 0%. Nous obtenons une rentabilité valant 5 085 216. Le volume de ce portefeuille est égal à 15 576. Cela signifie que 7814 assurés décideraient de résilier leur contrat si la prime restait inchangée soit environ un tiers du portefeuille. Ce scénario sera notre référence pour la suite.

Suite à cela, nous pouvons implémenter une ou plusieurs méthode *Standard*. Celles-ci dépendent de plusieurs éléments :

— Une prime minimum et une prime maximum qui déterminent l'intervalle de fluctuation de la variable prime. Dans notre cas, la prime minimum est fixée à 0,875 fois la prime commerciale et la prime maximum est, quant à elle, fixée à 1,125 fois la prime commerciale. Cela signifie que, ligne à ligne, la prime commerciale ne peut être augmentée ou diminuée de plus de 12,5% par l'algorithme d'optimisation.

— Une contrainte de volume. Elle représente le volume cible du portefeuille déterminé par l'algorithme d'optimisation. C'est sur cette contrainte que des modifications ont été apportées afin de définir le meilleur scénario possible.

— Un nombre maximal d'itérations fixé à 1000.

— Une valeur Epsilon définissant la contrainte d'arrêt de l'algorithme fixée à 10^{-3}.

Le paramétrage dans le logiciel est le suivant (figure 5.9) :

Define constraints

	Individual premium constraint				Volume constraint						Effective values	
	Premium inf.	Premium sup.			Volume limits	Target volume	Max iterations	Epsilon			Nb. iterations	Epsilon
	PCinf	PCsup	▶	✔	[14 852,5 ; 16 300,6]	15 000,0 (65,0 %)	1000	1,00E-003	▶	✔	11	0,00E+000
•	PCinf	PCsup	▶	✔	[14 852,5 ; 16 300,6]	16 000,0 (69,4 %)	1000	1,00E-003	▶	✔	107	9,43E-004
	PCinf	PCsup	▶	✔	[14 852,5 ; 16 300,6]	15 500,0 (67,2 %)	1000	1,00E-003	▶	✔	90	9,00E-004

FIGURE 5.9 – Paramétrage des méthodes *Standard*

Les volumes cibles, choisis en fonction de l'intervalle proposé par le logiciel, sont donc 15 000, 15 500 ainsi que 16 000. Les bornes inférieures et supérieures de cet intervalle proposé sont calculées à partir des primes minimales et maximales renseignées par l'utilisateur. En effet, le logiciel se place dans le cas où toutes les primes valent la prime minimale (resp. maximale) pour définir la borne inférieure (resp. supérieure).

On obtient alors les résultats suivants (figure 5.10) :

	Tag	Name	Main indicator	Premium	Volume	Vol	Group
	☐	[PCinf, PCsup] Volume: 16000,00; Epsilon: 10^-3	5 160 781	8 254 399	16 000	16 001	
	☐	[PCinf, PCsup] Volume: 15500,00; Epsilon: 10^-3	5 443 888	8 934 944	15 500	15 501	
•	☐	[PCinf, PCsup] Volume: 15000,00; Epsilon: 10^-3	5 474 109	9 201 169	15 278	15 278	

FIGURE 5.10 – Résultats des méthodes *Standard*

On remarque tout d'abord que dans chacun des cas l'algorithme d'optimisation trouve une solution fructueuse puisque l'ensemble des profits (ou "Main Indicator") calculés sont supérieurs à 5 085 216, notre profit de référence. On peut également remarquer que c'est le troisième scénario qui maximise le profit. Il s'agit du scénario dont le volume cible vaut 15 000. Pour ce scénario, le volume probable vaut 15 278 et présente donc un léger déclin par rapport à notre référence. En revanche, le profit de ce scénario s'établit à 5 474 109 et représente une hausse de 7,7% par rapport toujours au scénario de référence. C'est ce scénario qui est ainsi retenu.

A l'inverse de la méthode *Simulation*, nous ne pouvons pas directement connaître les majorations ou minorations effectuées pour arriver à ce profit. Pour récupérer les effets par variable, un GLM a alors été implémenté afin d'expliquer la nouvelle prime. La même structure que la prime de référence a été conservée. La comparaison des deux modèles nous permet alors de comprendre quels sont les changements structurels de la prime d'assurance maximisant le profit.

Les tableaux suivants recensent les évolutions des effets par variable :

Variable	Modalité	Evolution de la prime moyenne
polholder_diffdriver	only partner	8,20%
polholder_diffdriver	young-learner17 drivers	−1,05%
polholder_diffdriver	same	4,20%
polholder_diffdriver	all drivers > 24	4,77%

TABLE 5.1 – Evolutions entre la prime du portefeuille et la prime obtenue par l'algorithme d'optimisation pour la variable polholder_diffdriver

Le tableau 5.1 indique que le modèle d'optimisation préconise une majoration de prime de 8% pour les assurés dont le conducteur principal est le conjoint. En revanche, une minoration de prime égale à 1% est optimale pour les jeunes conducteurs ainsi que les apprentis conducteurs.

Variable	Modalité	Evolution de la prime moyenne
policy_age	[0 ; 1[	1,22%
policy_age	[1 ; 2[	5,37%
policy_age	[2 ; 3[	5,33%
policy_age	[3 ; 4[	5,79%
policy_age	[4 ; 5[	7,34%
policy_age	[5 ; 6[	11,19%
policy_age	[6 ; 7[	10,22%
policy_age	[7 ; 8[	11,58%
policy_age	[8 ; 9[	11,50%
policy_age	[9 ; ∞[	11,95%

TABLE 5.2 – Evolutions entre la prime du portefeuille et la prime obtenue par l'algorithme d'optimisation pour la variable policy_age

On remarque à l'aide du tableau 5.2 que globalement, l'ensemble des polices d'assurance ayant une ancienneté supérieure à 1 an fait l'objet d'une majoration de prime dans la méthode Standard conséquente et que cette majoration croît avec l'ancienneté. L'effet est même pratiquement linéaire : plus la police d'assurance est ancienne et plus il est préconisé de majorer le tarif par rapport à son niveau actuel.

Variable	Modalité	Evolution de la prime moyenne
vehicl_age	[0 ; 9[	4,14%
vehicl_age	[9 ; 11[	4,27%
vehicl_age	[11 ; 12[	4,80%
vehicl_age	[12 ; 13[	4,32%
vehicl_age	[13 ; 14[	4,17%
vehicl_age	[14 ; 15[	3,99%
vehicl_age	[15 ; 16[	4,95%
vehicl_age	[16 ; 17[	5,51%
vehicl_age	[17 ; ∞[	6,48%

TABLE 5.3 – Evolutions entre la prime du portefeuille et la prime obtenue par l'algorithme d'optimisation pour la variable vehicl_age

De même, on observe (tableau 5.3) que, selon notre modèle d'optimisation, les assurés seraient sous-tarifés en fonction de l'âge de leur véhicule. Une majoration de prime comprise entre 4% et 6% permettrait d'augmenter le profit de l'assureur.

Enfin, en ciblant nos analyses sur les segments de l'algorithme PRIM nous obtenons les résultats suivants (5.4) :

Segment étudié	Moyenne des évolutions ligne à ligne
Boîte sous-tarifée 1	5,97%
Boîte sous-tarifée 2	7,15%
Boîte sur-tarifée 1	−4,99%

TABLE 5.4 – Résultats de la méthode Standard sur les segments détectés

Il apparaît donc clairement que les segments détéctés par l'algorithme PRIM ne sont pas optimalement tarifés. Concernant les contrats appartenant au segment sur-tarifé, la méthode Standard propose une minoration, en moyenne, de près de 5% des primes. Quant aux segments sous-tarifés, il serait optimal d'augmenter la prime d'assurance de 6% à 7% selon le segment.

La méthode Standard permet ainsi d'expliciter la stratégie tarifaire à mettre en place afin d'optimiser la rentabilité de l'assureur. Pour aller plus loin, cette méthode pourrait être challengée par des méthodes d'optimisation plus poussées qui pourraient s'avérer plus performantes. En effet, c'est une méthode du Lagrangien Standard qui a été utilisée, méthode s'appliquant aux fonctions quadratiques uniquement. Une linéarisation a ainsi été nécessaire. La mise en place, par exemple, d'une méthode du Lagrangien augmenté, permettrait de retirer cette contrainte.

6 Limites opérationnelles des données et modèles implémentés

Les résultats des modèles que nous avons mis en oeuvre dépendent fortement de la base de données et donc de la stratégie commerciale de l'assureur qui porte le risque en question. De tels résultats peuvent évidemment être remis en question et challengés. Prenons l'exemple de la variable représentant l'ancienneté du contrat. Nous avons vu que l'ELR était quasiment croissant en fonction de l'ancienneté au sein de notre base : cela signifie que plus le contrat est récent plus il est rentable. Or, il est usuel de voir l'inverse chez les assureurs : la prime d'assurance croît habituellement avec les années (majorations annuelles) bien que la sinistralité de ces contrats diminue (conducteurs expérimentés, meilleure connaissance de leur risque par leurs distributeurs d'assurance). Souvent les assureurs acceptent donc de perdre en rentabilité au début du contrat pour attirer les clients en affaire nouvelle mais cette perte de rentabilité est rattrapée avec le temps. Ce n'est pas le cas de notre portefeuille pour lequel l'ELR croît avec l'ancienneté.

Plusieurs interprétations peuvent alors être données à cette hausse de l'ELR en fonction de l'ancienneté du contrat :

— Pour que la stratégie commerciale évoquée précedemment soit adoptée et fonctionne, cela suppose que le taux de résiliation du portefeuille ne soit pas trop élevé : au sein de notre portefeuille, en moyenne, la sensibilité des assurés au prix payé augmente avec leur ancienneté (4.2). L'ancienneté moyenne s'établit donc à 2,44 ans ce qui ne semble pas être assez élevé pour que la compensation sur plusieurs années s'effectue. Le choix est donc fait de ne pas sous-tarifer les affaires nouvelles quitte à ne pas les voir transformer leur devis.

— Ensuite, il est fortement probable que le modèle de tarification implémenté ne prédise pas précisément la prime pure d'un assuré en affaire nouvelle : de nombreuses caractéristiques du conducteur échappent au modèle qui ne

peut calibrer parfaitement le risque. C'est donc, au fil des années, par un suivi précis des contrats que l'assureur ou son distributeur peuvent réajuster le tarif ou même résilier un contrat dont le risque s'avère trop élevé. Plus l'ancienneté d'un contrat augmente et plus la prime commerciale doit donc être calibrée pour être à l'équilibre ou même excédentaire selon la stratégie de l'assureur. Si ce suivi de portefeuille n'est pas réalisé, nous pouvons nous retrouver dans la situation que l'on connaît : malgré une ancienneté qui augmente, le contrat ne gagne pas en rentabilité. Il peut donc s'agir d'un manque de suivi et de réajustement de la prime commerciale selon les sinistralités réelles observées. Connaître les informations passées concernant la sinistralité des contrats pourrait nous permettre de vérifier ces informations.

— Aussi, comme abordé précedemment, l'indicateur ELR présente des limites (voir partie 2.2) qui ne permettent pas d'assurer qu'un contrat dont l'ELR est plus élevé qu'un autre est forcément plus déficitaire. La prime pure est prédite, à partir du modèle de tarification de l'assureur, à une date T. La prime commerciale est alors évaluée à cette même date mais elle peut être réévaluée plus souvent. Par exemple, si le coefficient bonus-malus d'un assuré diminue, sa prime commerciale diminuera également selon le système appliqué dans certains pays européens. Au fil des années, et donc lorsque l'ancienneté du contrat augmente, la différence entre prime pure et prime commerciale peut décroître. L'ELR augmente mais cela ne signifie tout de même pas que la rentabilité des contrats est impactée et diminue.

Ensuite et concernant les modélisations effectuées, de tels modèles sont conçus au sein d'une direction technique. Ils doivent donc être confrontés à la réalité du marché et, en premier lieu, ce sont les équipes commerciales et distributeurs qui doivent en être informés pour les mettre en œuvre. Les distributeurs, qu'ils soient agents généraux, courtiers ou directement assureurs, sont au contact direct avec les clients et présentent ainsi un avantage certain : ils connaissent leurs assurés. Des potentielles majorations à appliquer seront à leur sens des freins pour exercer leur activité. Plusieurs solutions peuvent exister afin d'appliquer de façon plus réaliste un programme d'optimisation théorique :

— La première est d'évaluer les taux de déviation seuils :
  — Tout d'abord, on peut définir le taux au-delà duquel la résiliation doit être directement imposée à l'assuré. Cela permettra notamment de simplifier la gestion de ces contrats. En revanche, cette solution présente un risque concernant le volume du portefeuille.
  — Ensuite, on peut également définir le taux de déviation en dessous duquel il n'est pas opportun pour l'assureur d'engendrer des démarches pour

majorer la prime d'assurance de contrats, majoration qui s'avérerait dangereuse puisque l'acte de majoration peut être un élément déclencheur invitant l'assuré à faire jouer la concurrence.

— La seconde est de définir une stratégie de majoration. En effet, une majoration brutale de cotisation peut entraîner une forte résiliation des assurés. Lisser cette majoration sur plusieurs années peut ainsi s'avérer être un bon compromis pour l'assureur.

addactis® Software étant un éditeur de logiciels, nous n'avons pas une vision assez proche du marché sur de tels sujets et n'avons ainsi pas de préconisations quant aux seuils ou stratégies à appliquer.
En revanche, des données complémentaires concernant les évolutions tarifaires mises en place dans le passé pourraient être des ressources importantes dans le but de définir les taux seuils précédemment évoqués ou encore pour définir des stratégies de majoration lorsque celles-ci s'avèrent indispensables. Apprendre du comportement des assurés en cas de majoration permettrait très probablement de gagner en rentabilité.
Il aurait également été pertinent de s'intéresser à la valeur client de chaque police. Cet indicateur est en effet devenu important dans de nombreuses industries, et l'attention croissante portée à cet indicateur a vu émerger différentes définitions. Elles présentent toutes un point commun : celui d'évaluer deux quantités :

— Les revenus générés par le client : ils peuvent être directement financiers (les cotisations en assurance par exemple). Mais les bénéfices apportés par le client peuvent également être indirects : bouche à oreille permettant la transformation d'une affaire nouvelle par exemple ;

— Les coûts également générés par ce même client : ce peut être classiquement les remboursements d'un sinistre. Mais des coûts antérieurs à la prise d'effet du contrat d'assurance peuvent également être inclus dans la définition de la valeur client : on peut citer les coûts d'acquisition.

La modélisation de la valeur client nécessiterait la connaissance de données supplémentaires mais permettrait une meilleure maîtrise du portefeuille d'assurance et sans doute une prise de décision optimale quant aux évolutions tarifaires à mettre en oeuvre.

Conclusion

L'analyse d'un portefeuille d'assurance peut aisément s'avérer complexe tant par le volume de données dont disposent les assureurs que par les nombreux indicateurs pouvant être étudiés. Que ce soit la rentabilité, le comportement client ou la sinistralité, tous peuvent faire l'objet d'une étude approfondie. Dans ce rapport, le choix s'est porté sur l'indicateur de résiliation mais la pertinence de travailler avec la transformation des devis en contrat est évidente. Cet aspect n'est en effet pas négligeable : l'assureur est tenu de proposer des offres attractives lui permettant de gagner des parts de marché et d'assurer sa pérennité.

Les modèles classiquement utilisés en assurance, bien que très performants, ne permettent pas de suivre, tout au long de leur vie, l'ensemble des contrats d'assurance. Une méthode automatique a ainsi été développée, permettant de déterminer, à partir de contraintes liées au contexte d'utilisation, des groupes de contrats pour lesquels l'assureur est fortement déficitaire ou bénéficiaire. Les résultats d'une telle méthode peuvent être utiles à des fins commerciales comme à des fins techniques au sein d'une équipe d'actuaires. Automatiser l'analyse d'un portefeuille d'assurance répond ainsi à des besoins grandissants. De plus et à l'heure des modèles de Machine Learning qualifiés de "Black Box", l'algorithme PRIM utilisé propose un processus entièrement clair et traçable constituant une plus-value certaine.

Bien que cette méthode ait été utilisée, au sein de cette étude, dans le cadre de l'assurance automobile et de l'Expected Loss Ratio, elle ouvre la voie à de nombreux autres domaines que ce soit en assurance (IARD, santé, prévoyance) ou en finance. L'application de l'algorithme PRIM sur un Loss Ratio, dans le but de détecter des contrats pour lesquels la sinistralité réelle dévie fortement de la cotisation payée, formerait une étude enrichissante à mener pour un assureur.

Nous avons ensuite mis en place un modèle permettant de déterminer l'élasticité de la résiliation des assurés au prix. De tels résultats nous prouvent les limites de la majoration annuelle classique des assureurs : pour certains assurés, très sensibles au prix, la résiliation en sera directement impactée.

En définitive, l'approche présentée dans ce rapport propose en premier lieu deux études disjointes. La première, basée sur l'indicateur de rentabilité Expected Loss Ratio nous apporte de l'information quant aux populations de notre portefeuille dont la rentabilité s'éloigne de la moyenne. La seconde, basée sur l'indicateur de résiliation nous permet de cibler les populations les moins fidèles et de connaître leur élasticité au prix. Ces analyses, indépendantes, sont rapprochées dans notre troisième partie, apportant des solutions concrètes aux assureurs afin de concilier les résultats des deux premières modélisations. Cette dernière partie propose une optimisation tarifaire du portefeuille et apporte des résultats chiffrés aux assureurs quant aux majorations ou minorations de prime à mettre en place pour optimiser leur rentabilité.

Les modèles ainsi construits ont été entraînés sur un jeu de données réelles mais présentant des limites majeures. Tout d'abord, le faible nombre d'observations présentes dans la base nuit à l'optimisation des modèles implémentés. La séparation de la base de données en une base d'apprentissage et une base de validation, bien que commune et efficace, n'était pas réalisable dans cette étude. Le manque d'information et d'expertise sur les données constitue une seconde limite.

Les parties de ce mémoire nous ont ainsi permis de nous positionner à différentes échelles d'analyse d'un portefeuille d'assurance. Ces analyses ne sont pas monnaie courante chez les assureurs puisqu'aucun aspect réglementaire ne leur est attribué. Les méthodes exposées au sein de ce mémoire ont ainsi permis d'apprécier les enjeux ainsi que la complexité inhérente à la gestion d'un portefeuille d'assurance. L'application de telles méthodes au sein d'une compagnie permettrait de challenger l'approche proposée en évaluant notamment les bénéfices financiers que celle-ci pourrait susciter. Ces travaux s'inscriront dans la durée et nécessiteront d'être appréciés avec un recul nécessaire.

Bibliographie

[1] Site internet d'addactis® : La solution de tarification addactis® Pricing. 2021.

[2] OTT Armin. Package PRIM. 2017.

[3] ELIE Romuald CHARPENTIER Arthur, DENUIT Michel. Segmentation et mutualisation, les deux faces d'une même pièce ? 2015.

[4] SMETS Didier. Programmation linéaire et Optimisation.

[5] CHARPENTIER Arthur DUTANG Christophe. Package CASdatasets. 2020.

[6] FISHER Nicholas I. FRIEDMAN Jerome H. Bump-hunting in high-dimensional data. 1999.

[7] MARKAOUI Hajar. Analyse de la probabilité de résiliation en assurance automobile : comparatif de deux méthodologies d'estimation et conséquences sur la tarification. 2016.

[8] CASANOVA Jean-Jérôme. La méthode du Lagrangien.

[9] GRENET Julien. Vademecum : optimisation statique, Lagangien et conditions de Kuhn et Tucker.

[10] FONTAINE Léonard. La modélisation de la valeur contrat (pnpv) par la refonte du modèle de résiliation. 2011.

[11] BUFFAT Marc. Méthode de Gauss-Seidel.

[12] HAPFELMEIER Alexander OTT Armin. Nonparametric subgroup identification by prim and cart : A simulation and application study. 2017.

A Annexe : Le logiciel addactis® Pricing

addactis® Pricing [1] est la solution de la gamme de logiciels d'addactis® conçue pour améliorer l'ensemble du processus tarifaire des organismes assureurs. Ce logiciel a été développé pour répondre aux besoins de ces derniers dans un contexte concurrentiel et réglementaire en constante évolution. En effet, le positionnement concurrentiel ainsi que la conquête de parts de marché sont des enjeux actuels majeurs qui nécessitent le recours à des outils performants. La solution addactis® Pricing propose ainsi l'optimisation conjointe de la croissance et de l'équilibre technique.

addactis® Pricing est un outil tout-en-un disposant de trois modules distincts :

— Le module « Tariff » comprend toutes les étapes du processus de tarification, de l'analyse de données à la production du tarif, en utilisant les techniques actuarielles les plus abouties, au sein d'une application intuitive et transparente. Ce module met à la disposition de ses utilisateurs un puissant moteur GLM permettant de modéliser des indicateurs tarifaires clés : fréquence, coût moyen, prime de risque. Des méthodes de "reverse engineering" sont également incluses, toujours dans le but d'améliorer la structure tarifaire finale. Enfin, grâce au module Data Science, l'utilisateur peut recourir aux méthodes de Machine Learning les plus performantes : CART, Random Forest et Gradient Boosting. Une fois les modèles implémentés, le logiciel permet d'intégrer des rapports sur mesure. Les graphiques et tableaux ainsi créés sont des outils certains d'aide à l'analyse des modèles statistiques implémentés.

— Le module « Customer Behavior » est un module permettant d'étudier le comportement client à des instants clés de la vie de son contrat d'assurance : souscription, résiliation ou encore avenant. Il est alors possible de modéliser la réaction d'un assuré face aux changements tarifaires, de modéliser la transformation d'un devis en contrat et bien d'autres indicateurs.

— Enfin, le module « Price Simulation » est une composante destinée à la re-

cherche du meilleur ajustement du risque. En effet, l'ajustement du risque se base sur les modèles tarifaires et les analyses du comportement réalisés dans les composantes précédentes. Le meilleur ajustement sera alors défini en appliquant des modifications sur les paramètres des modèles afin d'ajuster au mieux le risque par rapport aux attentes.

Ce logiciel s'inscrit aussi dans une logique de transparence puisqu'un outil, l'« Audit Trail », permet une traçabilité complète du projet tarifaire réalisé. Les modifications et décisions prises par les actuaires sont ainsi partageables et interprétables permettant à l'ensemble du processus de tarification d'être documenté.

B

Annexe : Résultats du modèle de résiliation

Factor	Modality	Value	Standard error	Lower conf. limit	Upper conf. limit	Wald Chi-2	Pr > Chi-2	Probability (%)	Obs. count
(constant)	(constant)	-1,6435	0,1224	-1,8834	-1,4036	180,31	0,0000	16,1992	
polholder_age	[19 ; 35[	0,0000	0,0000	0,0000	0,0000	0,00	0,0000	16,1992	5 644
polholder_age	[35 ; 41[	-0,2100	0,0567	-0,3212	-0,0988	13,71	0,0002	13,5464	5 528
polholder_age	[41 ; 49[	-0,2313	0,0585	-0,3458	-0,1167	15,65	0,0001	13,2994	5 873
polholder_age	[49 ; ∞[	-0,3649	0,0639	-0,4901	-0,2396	32,60	0,0000	11,8328	6 015
polholder_diffdriver	only partner	0,0994	0,0455	0,0102	0,1886	4,77	0,0290	17,5940	8 128
polholder_diffdriver	young-leaner17 drivers	0,3404	0,0708	0,2016	0,4791	23,11	0,0000	21,3641	1 997
polholder_diffdriver	same	0,0000	0,0000	0,0000	0,0000	0,00	0,0000	16,1992	11 167
polholder_diffdriver	all drivers > 24	-0,0145	0,0803	-0,1718	0,1428	0,03	0,8563	16,0028	1 768
polholder_gender	Female	-0,1667	0,0427	-0,2505	-0,0830	15,22	0,0001	14,0614	8 339
polholder_gender	Male	0,0000	0,0000	0,0000	0,0000	0,00	0,0000	16,1992	14 721
polholder_job	medical	-0,0901	0,0417	-0,1720	-0,0083	4,66	0,0308	15,0125	9 482
polholder_job	normal	0,0000	0,0000	0,0000	0,0000	0,00	0,0000	16,1992	13 578
vehicl_garage	parking deck	-0,0809	0,0701	-0,2182	0,0565	1,33	0,2486	15,1312	2 243
vehicl_garage	private	0,0000	0,0000	0,0000	0,0000	0,00	0,0000	16,1992	12 718
vehicl_garage	street	-0,0305	0,0507	-0,1299	0,0689	0,36	0,5481	15,7899	5 468
vehicl_garage	underground garage	0,2304	0,0626	0,1077	0,3530	13,55	0,0002	19,5743	2 631
vehicl_region	Reg10	0,2577	0,1102	0,0418	0,4737	5,47	0,0193	20,0087	1 057
vehicl_region	Reg12	0,5987	0,0946	0,4133	0,7842	40,03	0,0000	26,0237	1 325
vehicl_region	Reg2	-0,3784	0,1743	-0,7200	-0,0368	4,71	0,0299	11,6926	556
vehicl_region	Reg8	0,2100	0,0817	0,0499	0,3701	6,61	0,0101	19,2561	3 074
vehicl_region	Reg1-5	0,0000	0,0000	0,0000	0,0000	0,00	0,0000	16,1992	3 009
vehicl_region	Reg11-14	0,4379	0,0888	0,2638	0,6120	24,31	0,0000	23,0488	1 847
vehicl_region	Reg3-4-6-7	0,1625	0,0679	0,0293	0,2957	5,72	0,0168	18,5280	10 427
vehicl_region	Reg9-13	0,4102	0,0914	0,2311	0,5893	20,15	0,0000	22,5606	1 765
policy_age	Order 1	-0,0535	0,0086	-0,0703	-0,0366	38,83	0,0000		
prem_last	Order 1	0,0003	0,0001	0,0002	0,0005	12,97	0,0003		
vehicl_age	Order 1	-0,0238	0,0077	-0,0389	-0,0088	9,65	0,0019		
vehicl_agepurchase	Order 1	0,0430	0,0143	0,0151	0,0710	9,09	0,0026		
vehicl_agepurchase	Order 2	-0,0039	0,0008	-0,0055	-0,0023	22,02	0,0000		

FIGURE B.1 – Résultats du modèle de résiliation obtenus à partir du logiciel addactis® Pricing

Traitement de la variable "vehicl_region" :

La variable indiquant la région de vie de la personne assurée a fait l'objet d'une attention particulière. En effet, cette dernière présentait un grand nombre de modalités s'avérant pour la plupart non-significatives dans les premiers modèles implémentés. Il a donc été nécessaire de réaliser des regroupements. La figure B.2 présente la moyenne des taux de résiliation pour chacune des régions :

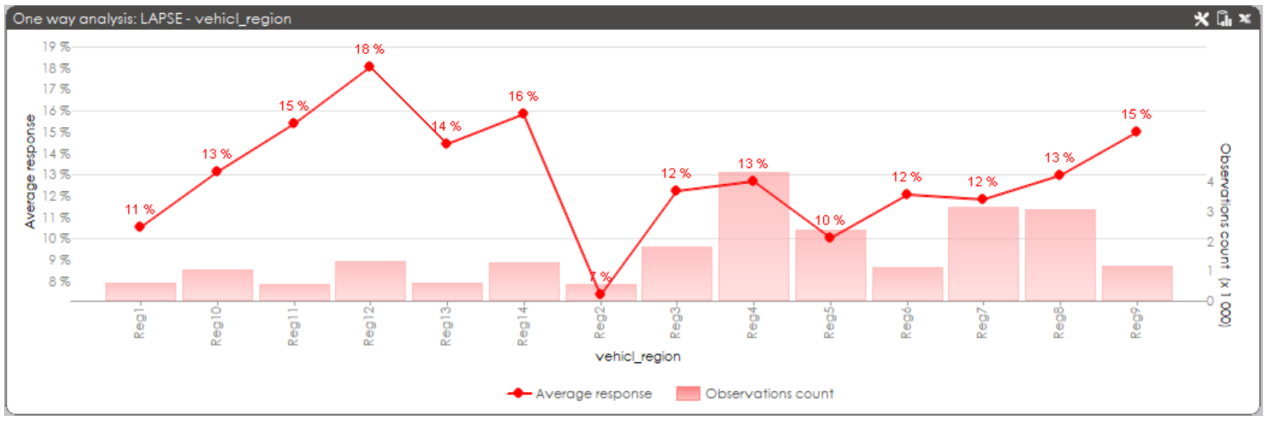

FIGURE B.2 – Analyse graphique univariée de la variable "vehicl_region" avant regroupements

Après regroupements effectués avec le logiciel addactis® Pricing, on obtient (figure B.3) :

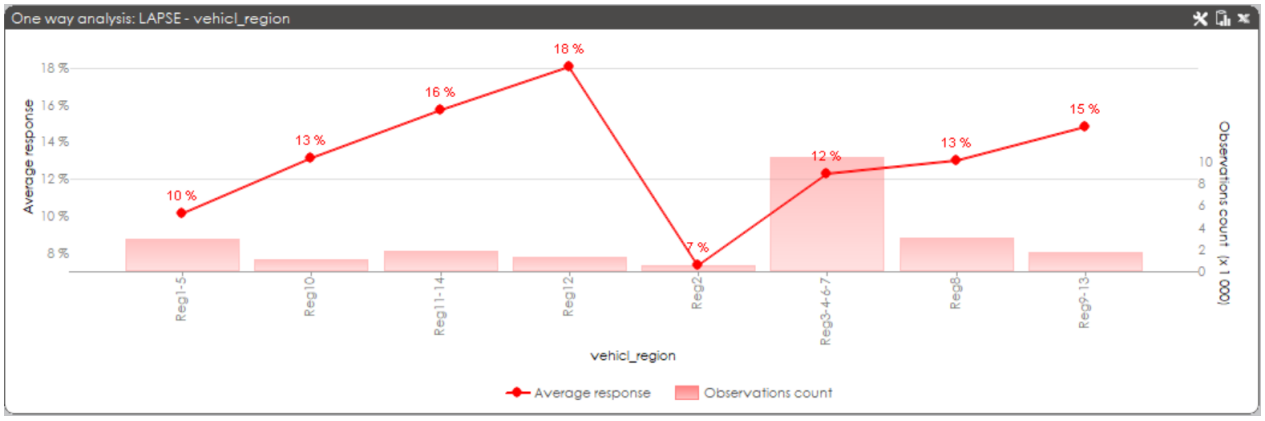

FIGURE B.3 – Analyse graphique univariée de la variable "vehicl_region" après regroupements

C Annexe : Exemple d'utilisation de l'algorithme PRIM

Le but ici est d'illustrer le fonctionnement de l'algorithme PRIM de manière plus concrète et étape par étape. Pour cet exemple, nous considérons le jeu de données suivant :

Identifiant	Sexe	Situation maritale	Nombre d'enfant(s)	Milieu de vie	Y
1	H	Célibataire	1	Rural	0
2	F	Célibataire	2	Citadin	0
3	H	Marié(e)	3	Citadin	0
4	H	Célibataire	0	Rural	0
5	F	Marié(e)	4	Citadin	0
6	F	Marié(e)	2	Citadin	1
7	H	Marié(e)	2	Citadin	1
8	H	Marié(e)	2	Rural	1
9	F	Célibataire	4	Citadin	1
10	F	Marié(e)	1	Citadin	1
11	F	Célibataire	3	Citadin	0
12	H	Marié(e)	1	Rural	0
13	H	Célibataire	0	Rural	0
14	F	Marié(e)	5	Citadin	0
15	F	Marié(e)	2	Rural	0
16	F	Celibataire	2	Rural	1
17	H	Célibataire	4	Citadin	1
18	F	Marié(e)	0	Citadin	1
19	H	Marié(e)	2	Citadin	1
20	H	Célibataire	1	Rural	1

Table C.1 – Jeu de données exemple

L'algorithme PRIM nécessite également deux méta-paramètres. Dans cet exemple, nous fixons $\beta = 0,2$ et $\alpha = \frac{1}{3}$. Cela signifie qu'au maximum, en sortie de l'algorithme PRIM, les boîtes contiendront 4 individus $(0,2 \times 20)$.

La variable cible sera donc la variable Y qui est une variable binaire. Sur notre jeu de données, on a $\overline{Y_B} = 0,5$, la variable cible est parfaitement répartie dans notre population. Nous allons ainsi chercher à maximiser cette moyenne.

Boîte numéro 1 :

1) Top-down peeling :

a) Etape 1 : Lors de cette étape, l'objectif est de recenser l'ensemble des boîtes éligibles pour chaque variable explicative. A la suite de cela, la boîte éligible retenue sera celle qui permettra de maximiser la moyenne sur $B \setminus b$.

Etape 1 Peeling				
Variable	b	Nombre d'observations retirées	Moyenne sur $B \setminus b$	$\geq \overline{Y_B}$ (éligible) ?
Nombre d'enfants	$\leq \frac{1}{3}$	7	0,54	Oui
	≥ 2	13	0,43	Non
Situation maritale	Marié(e)	11	0,44	Non
	Célibataire	9	0,55	Yes
Milieu de vie	Citadin	12	0,375	Non
	Rural	8	0,58	Oui
Sexe	F	10	0,5	Non
	H	10	0,5	Non

TABLE C.2 – Etape 1 de la procédure de peeling

Le tableau ci-dessus (C.2) présente les calculs implicites réalisés. Concernant les variables nominales : situation maritale, milieu de vie et sexe, la procédure est simple. Pour chaque variable, on retire tour à tour chacune des modalités. Par exemple, pour la variable situation maritale, si l'on retire l'ensemble des individus qui présentent la modalité "Marié(e)", on retire 11 individus et notre nouvelle base $B \setminus b$ sera la suivante (C.3) :

Identifiant	Sexe	Situation maritale	Nombre d'enfant(s)	Milieu de vie	Y
1	H	Célibataire	1	Rural	0
2	F	Célibataire	2	Citadin	0
4	H	Célibataire	0	Rural	0
9	F	Célibataire	4	Citadin	1
11	F	Célibataire	3	Citadin	0
13	H	Célibataire	0	Rural	0
16	F	Celibataire	2	Rural	1
17	H	Célibataire	4	Citadin	1
20	H	Célibataire	1	Rural	1

TABLE C.3 – Base de données à laquelle les marié(é)s ont été retirés

Sur ce nouvel ensemble, on peut calculer la moyenne qui vaut $\frac{4}{9}$ soit environ 0,44. Or, 0,44 étant inférieur à notre moyenne initiale de 0,5, cette boîte n'est pas éligible pour la suite.

Pour la variable numérique nombre d'enfants, le processus est différent. En effet, l'algorithme va utiliser les quantiles de niveau α et $1 - \alpha$ qui valent dans notre cas $\frac{4}{3}$ et 3. On retirera donc dans un premier temps l'ensemble des individus ayant au plus un enfant puis ceux ayant 2 enfants et plus.

Une fois que l'on connaît l'ensemble des boîtes éligibles, la boîte retenue (et donc la modalité ou le sous-ensemble retiré) sera celle telle que la moyenne sur $B \setminus b$ est maximisée. Dans notre cas, la plus grande moyenne est 0,58. L'ensemble des individus vivant en milieu rural est donc retiré et on obtient B : "Individus citadins". On doit alors vérifier que le critère d'arrêt n'est pas atteint. Or, ici, 12 observations appartiennent à notre nouvelle boîte et $\frac{12}{20} > \beta$, le critère d'arrêt n'est pas respecté, l'algortihme de peeling continue.

b) Etape 2 : A la deuxième itération du top-down peeling, l'algorithme progresse avec les données suivantes (C.4) :

Identifiant	Sexe	Situation maritale	Nombre d'enfant(s)	Milieu de vie	Y
2	F	Célibataire	2	Citadin	0
3	H	Marié(e)	3	Citadin	0
5	F	Marié(e)	4	Citadin	0
6	F	Marié(e)	2	Citadin	1
7	H	Marié(e)	2	Citadin	1
9	F	Célibataire	4	Citadin	1
10	F	Marié(e)	1	Citadin	1
11	F	Célibataire	3	Citadin	0
14	F	Marié(e)	5	Citadin	0
17	H	Célibataire	4	Citadin	1
18	F	Marié(e)	0	Citadin	1
19	H	Marié(e)	2	Citadin	1

TABLE C.4 – Individus citadins

Etape 2 Peeling				
Variable	b	Nombre d'observations retirées	Moyenne sur $B \setminus b$	$\geq 0,58$ (éligible) ?
Nombre d'enfants	≤ 2	6	0,33	Non
	$\geq \frac{10}{3}$	4	0,625	Oui
Situation maritale	Marié(e)	8	0,5	Non
	Célibataire	4	0,625	Oui
Sexe	F	8	0,75	Oui
	H	4	0,5	Non

TABLE C.5 – Etape 2 de la procédure de peeling

Le tableau nous détaillant l'étape 2 de la procédure de peeling (C.5) nous indique qu'à cette étape, ce sont les femmes qui vont être retirées puisque privées de celles-ci, les données présentent une moyenne cible élévée (égale à 0,75).

Une fois les citadins ainsi que les femmes retirées des données initiales, quatre individus demeurent. On atteint donc le critère d'arrêt puisque $\frac{4}{12} = 0,2 = \beta$, la procédure de peeling s'arrête à cette étape.

La boîte provisoire contient alors les observations suivantes (C.6) :

Identifiant	Sexe	Situation maritale	Nombre d'enfant(s)	Milieu de vie	Y
3	H	Marié(e)	3	Citadin	0
7	H	Marié(e)	2	Citadin	1
17	H	Célibataire	4	Citadin	1
19	H	Marié(e)	2	Citadin	1

TABLE C.6 – Individus citadins masculins

2) Bottom-up pasting :

Dans cette seconde procédure, on s'intéresse aux données ayant été précédemment retirées par le top-down peeling qui sont les suivantes (C.7) :

Identifiant	Sexe	Situation maritale	Nombre d'enfant(s)	Milieu de vie	Y
1	H	Célibataire	1	Rural	0
2	F	Célibataire	2	Citadin	0
4	H	Célibataire	0	Rural	0
5	F	Marié(e)	4	Citadin	0
6	F	Marié(e)	2	Citadin	1
8	H	Marié(e)	2	Rural	1
9	F	Célibataire	4	Citadin	1
10	F	Marié(e)	1	Citadin	1
11	F	Célibataire	3	Citadin	0
12	H	Marié(e)	1	Rural	0
13	H	Célibataire	0	Rural	0
14	F	Marié(e)	5	Citadin	0
15	F	Marié(e)	2	Rural	0
16	F	Celibataire	2	Rural	1
18	F	Marié(e)	0	Citadin	1
20	H	Célibataire	1	Rural	1

TABLE C.7 – Données exclues de la boîte issue de la procédure de top-down peeling

A partir de ces données, la procédure de bottom-up pasting va chercher à cette fois-ci élargir la boîte détectée par le top-down peeling. Pour cela, une procédure semblable est répétée puisque pour chacune des variables, on regarde si l'ajout d'une modalité (variable nominale) ou d'un sous-ensemble (variable numérique) permettrait d'améliorer la moyenne.

Etape 1 Pasting				
Variable	b	Nombre d'observations ajoutées	Moyenne sur $B \cup b$	$\geq 0,75$ (éligible) ?
Nombre d'enfants	≤ 1	7	0,55	Non
	≥ 2	9	0,54	Non
Situation maritale	Marié(e)	8	0,58	Non
	Célibataire	8	0,5	Non
Milieu de vie	Citadin	8	0,58	Non
	Rural	8	0,5	Non
Sexe	F	10	0,57	Non
	H	6	0,5	Non

TABLE C.8 – Etape 1 de la procédure de pasting

Dans notre cas, le tableau C.8 nous indique qu'aucun ajout de données ne permet d'améliorer la moyenne de la variable cible Y. Le bottom-up pasting prend ainsi fin et nous avons obtenu la boîte B_1 : $B_1 = B\backslash$ (Milieu de vie = Citadin $\cup$ Sexe = F) ou autrement dit, les hommes citadins présentent un fort écart à la moyenne sur la variable cible Y.

Mais l'algorithme PRIM ne s'arrête pas là, il peut en effet chercher à détecter de nouveaux sous-ensembles (ou boîtes). La même procédure sera répétée mais, cette fois-ci, le jeu de donnée sur lequel le top-down peeling débutera sera la base initiale à laquelle on retirera la boîte B_1.

D Annexe : Résultats du modèle d'élasticité

Significativité globale des variables du modèle :

Variable	Degré de liberté	Test du Chi2
prem_last	1	4.106e-09
policy_age	1	< 2.2e-16
polholder_age_class	3	1.859e-06
vehicl_garage	3	0.0002514
polholder_job	1	0.0228109
polholder_diffdriver	3	6.532e-07
polholder_gender	1	0.0002750
vehicl_age	1	2.030e-12
vehicl_agepurchase	2	3.839e-08
vehicl_Region	7	6.339e-14
prem_last*polholder_age_class	3	0.0305717

Table D.1 – Significativité globale des variables du modèle

A nouveau, le tableau D.1 indique que les onze variables inclues dans le modèle apportent de l'information à ce dernier puisque la probabilité indiquée dans la seconde colonne est proche de 0 (au seuil de 5% notamment).

Significativité des paramètres :

Ici, l'interprétation est la même que pour le modèle de l'annexe B. La colonne "Pr > Chi-2" du tableau D.2 présente des valeurs proches de 0 qui nous permettent d'affirmer que les paramètres du modèle sont significatifs (différents de 0) dans le modèle à l'étude.

Variable	Coefficient	Pr > Chi-2
Intercept	-1.549e+00	< 2e-16
prem_last	1.608e-04	0.227315
policy_age	-5.274e-02	8.71e-10
polholder_age_class [35,41)	-2.257e-01	0.043637
polholder_age_class [41,49)	-4.393e-01	5.75e-05
polholder_age_class [49,86)	-5.736e-01	3.39e-07
vehicl_garage parking deck	-7.982e-02	0.254760
vehicl_garage street	-3.348e-02	0.509512
vehicl_garage underground garage	2.243e-01	0.000346
polholder_job medical	-9.219e-02	0.027290
polholder_diffdriver all drivers >24	-1.785e-02	0.823977
polholder_diffdriver young-learner 17 drivers	3.103e-01	1.55e-05
polholder_diffdriver only partner	9.398e-02	0.039144
polholder_gender Female	-1.696e-01	7.31e-05
vehicl_age	-2.315e-02	0.002611
vehicl_agepurchase	-1.416e+01	0.004515
vehicl_agepurchase^2	-1.484e+01	4.43e-06
vehicl_Region10	2.624e-01	0.017269
vehicl_Region11-14	4.372e-01	8.61e-07
vehicl_Region12	5.974e-01	2.78e-10
vehicl_Region9-13	4.084e-01	7.90e-06
vehicl_Region2	-3.806e-01	0.029096
vehicl_Region3-4-6-7	1.621e-01	0.017093
vehicl_Region8	2.076e-01	0.011084
prem_last*polholder_age_class [35,41)	-3.435e-05	0.888400
prem_last*polholder_age_class [41,49)	4.978e-04	0.024838
prem_last*polholder_age_class [49,86)	5.194e-04	0.026289

Table D.2 – Résultats du modèle d'élasticité

E Annexe : Passage du problème primal au problème dual

On rappelle que dans le cadre quadratique, un programme d'optimisation sous contraintes peut s'écrire :

$$\begin{array}{ll} \max\limits_{\tilde{x}^*} & f(x) = \frac{1}{2}x^T Q x + L^T x \text{ avec } \mathrm{x}^T = (x_1, ..., x_n) \\ \text{s.c.} & g_j(x) \leq c_j \ \forall \ j = 1, ..., m \end{array} \tag{E.1}$$

où :

- Q est une matrice semi-définie négative de taille n*n ;
- L et x sont deux vecteurs colonnes à n composantes ;
- les contraintes s'écrivent sous la forme : $g(x) = Ax + b$. De ce fait, $Ax + b \leq 0$ avec A une matrice m*n et b un vecteur colonne à m composantes.

Ce programme est le problème primal.

Le Lagrangien de ce dernier s'écrit :

$$\mathscr{L}(x, \lambda) = \frac{1}{2}x^T Q x + L^T x - \lambda^T (Ax + b) \tag{E.2}$$

Puis, si x^* est un optimum local, d'après les conditions de Karush, Kuhn et Tucker, il satisfait :

$$\frac{\partial \mathscr{L}(x^*, \lambda)}{\partial x_i} = 0 \tag{E.3}$$

avec $\lambda \geq 0$.

Soit $Qx^* + L - A^T\lambda = 0$ et ainsi $x^* = Q^{-1}(A^T\lambda - L)$.

Ce qui nous permet d'écrire :

$$\mathscr{L}(x^*, \lambda) = \frac{1}{2}x^{*T} Q x^* + L^T x^* - \lambda^T (Ax^* + b) \tag{E.4}$$

D'où :

$$\begin{aligned}\mathscr{L}(x^*,\lambda) = \frac{1}{2}(A^T\lambda - L)^T(Q^{-1})^T QQ^{-1}(A^T\lambda - L) + \\ L^TQ^{-1}(A^T\lambda - L) - \lambda^T\left[AQ^{-1}(A^T\lambda - L) + b\right]\end{aligned} \tag{E.5}$$

Qui se simplifie en :

$$\begin{aligned}\mathscr{L}(x^*,\lambda) = \frac{1}{2}\left[(\lambda^TAQ^{-1}A^T\lambda) - \lambda^TAQ^{-1}L - L^TQ^{-1}(A^T\lambda - L)\right] + \\ L^TQ^{-1}(A^T\lambda - L) - \lambda^T[AQ^{-1}(A^T\lambda - L) + b]\end{aligned} \tag{E.6}$$

Puis,

$$\begin{aligned}\mathscr{L}(x^*,\lambda) = -\frac{1}{2}(\lambda^TAQ^{-1}A^T\lambda) + \frac{1}{2}L^TQ^{-1}(A^T\lambda - L) + \\ \frac{1}{2}\lambda^TAQ^{-1}L - \lambda^Tb\end{aligned} \tag{E.7}$$

Or, on a :

$$\begin{aligned}\frac{1}{2}\lambda^TAQ^{-1}L &= \frac{1}{2}(L^TQ^{-1}A^T\lambda)^T \\ &= \frac{1}{2}(L^TQ^{-1}A^T\lambda)\end{aligned} \tag{E.8}$$

puisque la matrice est scalaire.

De même, $\lambda^Tb = b^T\lambda$.

D'où :

$$\begin{aligned}\mathscr{L}(x^*,\lambda) = -\frac{1}{2}(\lambda^TAQ^{-1}A^T\lambda) + (L^TQ^{-1}A^T - b^T)\lambda - \\ \frac{1}{2}L^TQ^{-1}L\end{aligned} \tag{E.9}$$

D'après cela et si on pose : $G = -AQ^{-1}A^T$ et $g^T = L^TQ^{-1}A^T - b^T$, on obtient :

$$\mathscr{L}(x^*,\lambda) = \frac{1}{2}\lambda^TG\lambda + g^T\lambda - \frac{1}{2}L^TQ^{-1}L \tag{E.10}$$

Ainsi, nous venons de montrer que chercher la solution au problème primal ci-dessous :

$$\begin{array}{ll} \max\limits_{\tilde{x}^*} & f(x) = \frac{1}{2}x^T Q x + L^T x \text{ avec } \mathrm{x}^T = (x_1, ..., x_n) \\ \text{s.c.} & g_j(x) \leq c_j \ \forall \ j = 1, ..., m \end{array} \tag{E.11}$$

revient de façon équivalente à résoudre le programme suivant que nous appelons programme dual :

$$\begin{array}{ll} \min\limits_{\lambda} & f(x) = \frac{1}{2}\lambda^T G \lambda + g^T \lambda \\ \text{s.c.} & \lambda_j \geq 0 \ \forall \ j = 1, ..., m \end{array} \tag{E.12}$$

avec $G = -AQ^{-1}A^T$ et $g^T = L^T Q^{-1} A^T - b^T$

Liste des abréviations

— GLM - Modèles Linéaires Généralisés
— PRIM - Patient Rule Induction Method
— LR - Loss Ratio
— ELR - Expected Loss Ratio
— PP - Prime Pure
— PC - Prime Commerciale
— BI - Business Intelligence
— KPI - Key Performance Indicator (ou Indicateur Clé de Performance)
— CRM - Clause de Réduction-Majoration (ou Bonus-Malus)

Table des figures

2.1 Profil de référence du portefeuille étudié 19
2.2 Distribution de l'âge des assurés du portefeuille 20
2.3 Différence de statut entre l'assuré et le conducteur principal du véhicule 21
2.4 Analyse de l'ELR en fonction de l'évolution du CRM 22
2.5 Analyse de l'ELR en fonction de l'âge de la police d'assurance 23
2.6 Analyse de l'ELR en fonction de l'âge de l'assuré 24
2.7 Analyse de la résiliation en fonction de l'évolution du CRM de l'assuré 25
2.8 Analyse de la résiliation en fonction de l'âge de l'assuré 26
2.9 Test de type III appliqué aux variables du modèle de résiliation . . . 31
2.10 Graphique Lift 32
2.11 Matrice de confusion 33
2.12 Courbe ROC du modèle de résliation implémenté 34

3.1 Stratégie de couverture de l'algorithme PRIM 41
3.2 Illustration de l'implémentation de la procédure de peeling en deux dimensions 44
3.3 Exemple de trajectoire d'un processus de peeling 47
3.4 Segments déviants obtenus par l'algorithme PRIM 52
3.5 Segments déviants obtenus par l'algorithme PRIM sans la variable CRM 54
3.6 Diagramme de fréquences relatives associé à l'âge du souscripteur . . 55
3.7 Diagramme de fréquences relatives associé au statut de l'assuré par rapport au souscripteur 55
3.8 Diagramme de fréquences relatives associé au genre du souscripteur . 56
3.9 Diagramme de fréquences relatives associé au secteur d'activité du souscripteur 56
3.10 Diagramme de fréquences relatives associé à l'ELR de l'assuré 57
3.11 Segments déviants obtenus par l'algorithme PRIM 58
3.12 Diagramme de fréquences relatives associé à l'ancienneté de la police d'assurance 59
3.13 Diagramme de fréquences relatives associé au moyen de garage de la voiture assurée 60
3.14 Diagramme de fréquences relatives associé au genre de l'assuré 60

3.15 Diagramme de fréquences relatives associé à l'ELR de l'assuré 61

4.1 Élasticité des segments sur-tarifés détectés par l'algorithme PRIM . . 68
4.2 Élasticité des segments sous-tarifés détectés par l'algorithme PRIM . 69

5.1 Définition de la prime et de l'indicateur principal dans le logiciel addactis® Pricing 80
5.2 Définition de l'indicateur secondaire dans le logiciel addactis® Pricing 80
5.3 Paramétrage des simulations sur la première boîte dans le logiciel addactis® Pricing 81
5.4 Résultat graphique des simulations réalisées sur la première boîte sous-tarifée 81
5.5 Tableau des simulations réalisées sur la première boîte sous-tarifée . . 82
5.6 Tableau des simulations réalisées sur la seconde boîte sous-tarifée . . 83
5.7 Résultat graphique des simulations réalisées sur la première boîte sur-tarifée 84
5.8 Tableau des simulations réalisées sur la première boîte sur-tarifée . . 84
5.9 Paramétrage des méthodes *Standard* 85
5.10 Résultats des méthodes *Standard* 86

B.1 Résultats du modèle de résiliation obtenus à partir du logiciel addactis® Pricing 97
B.2 Analyse graphique univariée de la variable "vehicl_region" avant regroupements 98
B.3 Analyse graphique univariée de la variable "vehicl_region" après regroupements 98

Liste des tableaux

2.1 Tableau croisé dynamique des résiliations estimées par profil 35
2.2 Tableau croisé dynamique des résiliations estimées par profil 36
2.3 Odds Ratio pour la variable polholder_age 37

3.1 Etude des dissimilarités entre les différentes boîtes 57
3.2 Étude des dissimilarités entre les différentes boîtes 61

4.1 Élasticités moyennes selon le genre de l'assuré 66
4.2 Élasticités moyennes selon l'ancienneté du contrat 66
4.3 Élasticités moyennes selon l'âge de l'assuré 67

5.1 Evolutions entre la prime du portefeuille et la prime obtenue par l'algorithme d'optimisation pour la variable polholder_diffdriver . . . 86
5.2 Evolutions entre la prime du portefeuille et la prime obtenue par l'algorithme d'optimisation pour la variable policy_age 87
5.3 Evolutions entre la prime du portefeuille et la prime obtenue par l'algorithme d'optimisation pour la variable vehicl_age 87
5.4 Résultats de la méthode Standard sur les segments détectés 88

C.1 Jeu de données exemple 99
C.2 Etape 1 de la procédure de peeling 100
C.3 Base de données à laquelle les marié(é)s ont été retirés 101
C.4 Individus citadins 102
C.5 Etape 2 de la procédure de peeling 102
C.6 Individus citadins masculins 103
C.7 Données exclues de la boîte issue de la procédure de top-down peeling 103
C.8 Etape 1 de la procédure de pasting 104

D.1 Significativité globale des variables du modèle 105
D.2 Résultats du modèle d'élasticité 106