INSTITUT DES ACTUAIRES

le cnam
Conservatoire national des arts et métiers

Mémoire présenté devant le CNAM pour l'obtention du diplôme de la filière Actuariat et l'admission à l'Institut des Actuaires

le 10/02/2021

Par : Sètondji Fiacre KINMAGBAHOHOUE

Titre : Segmentation des agents généraux pour une optimisation des mesures d'accompagnement

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

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

Membres présents du jury de l'Institut des Actuaires	*Signatures*	*Entreprise :*
Mme Edith BOCQUAIRE		AXA France
Mme Florence PICARD		
M. François-Xavier NEGRI		

Membres présents du jury du CNAM

Mme Sandrine LEMERY *(Présidente)*

M. Olivier DESMETTRE

M. David FAURE

M. François WEISS

Directeur de mémoire en entreprise :

Nom : M. Guillaume GORGE

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

Secrétariat

Bibliothèque :

40 rue des Jeûneurs,
75002 Paris

Remerciements

Je tiens à remercier tout particulièrement mon encadrant, Guillaume GORGE, pour m'avoir guidé tout au long de ce travail, pour sa confiance et sa disponibilité.

Je remercie également toute l'équipe de la Direction Techniques Distributeurs AXA PIE pour leur convivialité, leur professionnalisme et leur coopération. Ensemble, vous m'avez tous permis d'avoir l'environnement idéal pour réaliser ce mémoire.

Enfin, toute ma gratitude à ma femme, pour sa disponibilité et son soutien indéfectible. Merci infiniment d'avoir supporté mes absences et d'avoir été là pour nos filles.

Résumé — L'apparition de nouveaux acteurs tels que les bancassureurs et la vente en ligne ont transformé et rendu plus compétitive la distribution des produits d'assurances. Conscient de l'importance de la distribution pour le développement et le multi-équipement, AXA a toujours accompagné ses agents généraux afin de leur permettre de conserver leur portefeuille et d'attirer de nouveaux clients. Ces mesures visant à améliorer la performance ne sont pas toujours adaptées au contexte de chaque agent.

Ce mémoire a pour objectif de rationaliser les mesures d'accompagnement et de les rendre plus efficaces. Il propose une segmentation des agents en groupes homogènes de comportements et utilise les caractéristiques de chaque groupe pour identifier les mesures adéquates à ce dernier.

La première partie est consacrée à la sélection d'indicateurs pertinents et susceptibles de caractériser les agents suivant trois dimensions : la production, la maîtrise des coûts et l'utilisation des budgets commerciaux. Ensuite l'information suivant ces trois dimensions sera utilisée pour classer les agents en cinq groupes homogènes à l'aide d'un k-means. Enfin, des méthodes de classifications supervisées (Régression logistique multinomiale et CART) seront comparées pour l'expliquer les groupes obtenus à partir des indicateurs sélectionnés. L'apprentissage supervisé permettra de mieux caractériser chaque groupe et d'opérationnaliser la segmentation.

Mots clés : ACP, K-means, Régression logistique multinomiale, Arbre de décision CART, A/B Testing.

Abstract — The emergence of new players such as "bancassurance" and online sales has made the distribution of insurance products more competitive. Aware of the importance of distribution for development and multi-equipment, AXA has always accompanied its distributors to enable them to maintain their portfolio and attract new customers. These measures to improve performance are not always adapted to the context of each agent.

The purpose of this thesis is to streamline support measures and make them more effective. It proposes a segmentation of the distributors into homogeneous groups of behaviors and uses the characteristics of each group to identify the appropriate measures.

The first part is devoted to the selection of relevant indicators that can characterize agents according to three dimensions : production, cost control and the use of commercial budgets. Then the information following these three dimensions will be used to classify the agents into five homogeneous groups using a k-means method. Finally, supervised classification methods (multinomial logistic regression and CART) will be compared for learning on the groups obtained from the selected indicators. This supervised learning will better characterize each group and operationalize the segmentation.

Keywords : PCA, K-means, multinomial logistic regression, Decision Tree CART, A/B Testing.

Table des matières

Introduction **1**

1 Contexte de l'étude et problématique **3**

1.1 AXA et la distribution des produits d'assurance 3

1.2 Les agents généraux AXA 7

1.3 Problématique de l'étude 14

2 Construction de la base de données **17**

2.1 Périmètre de l'étude 17

2.2 Choix des indicateurs 18

2.3 Analyse exploratoire des données 21

2.4 Construction de la base de segmentation 32

3 Segmentation des distributeurs **35**

3.1 Classification non supervisée 35

3.2 Interprétation des groupes 40

4 Opérationnalisation de la segmentation **49**

4.1 Classification supervisée 49

4.2 Généralisation de la segmentation 65

4.3 A/B Testing avec une dotation ST aux groupes 68

4.4 Mesures opérationnelles 70

Conclusion **73**

A Définition des indicateurs de l'étude **75**

B Réduction de dimensions avec l'ACP **79**

B.1 ACP sur les indicateurs de maîtrise de coût 79

B.2 ACP sur les indicateurs d'utilisation de budgets commerciaux 82

C Modèles Linéaires Généralisés **87**

C.1 Hypothèses et principes des GLM 87

C.2 Estimation des paramètres 89

Bibliographie **91**

Introduction

Dans un marché de plus en plus concurrentiel, la distribution des produits d'assurance constitue un pilier pour favoriser le développement de l'activité et multi-équiper les clients. Historiquement gérés par les intermédiaires tels que les courtiers, les agents généraux..., les canaux de distribution n'ont cessé d'évoluer ces dernières décennies.

L'arrivée de nouveaux acteurs tels que les bancassureurs, la vente en ligne, les comparateurs d'assurances a fragilisé ces réseaux de distribution traditionnels. Ces derniers doivent faire face à une concurrence féroce pour conserver ou attirer des clients de plus en plus sensibles au tarif. Ces changements ne sont pas sans conséquences sur la qualité de la souscription des intermédiaires d'assurances et impactent les résultats techniques. Le réseau des agents généraux, canal le plus utilisé par AXA France pour la distribution des produits IARD sur le marché des particuliers et des professionnels, est fortement touché par ces mutations.

Pour les accompagner, AXA France utilise des mesures d'accompagnement tels que les latitudes commerciales, les leads, les analyses techniques de portefeuille pour aider ses agents à booster leur production tout en veillant à la rentabilité et au développement du chiffre d'affaires. Cependant, ces mesures d'accompagnement ne tiennent pas compte du comportement de chaque agent.

Pour optimiser ces mesures et améliorer la rentabilité des agents, ce mémoire propose une segmentation des agents généraux en groupes homogènes de comportements. Cette segmentation permettra de créer des profils d'agents et d'adapter les mesures d'accompagnement sur chaque segment afin de les rendre plus efficaces.

Après avoir présenté l'organisation actuelle des agents généraux d'AXA France et les accompagnements dont ils bénéficient, nous évoquerons les insuffisances de la répartition de ces mesures ainsi que les risques liés aux comportements des agents. Ensuite nous identifierons des indicateurs pertinents pour caractériser le comportement des agents. Ces indicateurs serviront pour la création d'une base que nous utiliserons pour faire une classification non supervisée.

Enfin, nous allons utiliser le résultat de cette classification pour faire de l'apprentissage supervisé afin de mieux caractériser les groupes et rendre opérationnelle la segmentation. Nous finirons notre étude par un A/B test sur les groupes et la proposition de mesures d'accompagnement adaptées à chaque groupe.

Chapitre 1

Contexte de l'étude et problématique

Sommaire

1.1 AXA et la distribution des produits d'assurance 3
1.1.1 Les canaux de distribution 3
1.1.2 AXA France et la direction de l'offre IARD 5
1.1.3 Les intermédiaires de distribution d'AXA France 6
1.2 Les agents généraux AXA 7
1.2.1 Les budgets commerciaux des AGA 8
1.2.2 Le ratio combiné protocolaire 10
1.2.3 Coefficient Technique Personnalisé (CTP) : 11
1.2.4 Les risques liés aux comportements des agents 12
1.3 Problématique de l'étude 14

1.1 AXA et la distribution des produits d'assurance

1.1.1 Les canaux de distribution

Le marché de l'assurance est de plus en plus concurrentiel et chaque acteur utilise tous les leviers à sa disposition pour assurer la croissance de l'activité. Les assureurs peuvent compter sur différents canaux de distribution pour attirer de nouveaux clients tout en améliorant les résultats techniques. Ces canaux de distribution peuvent être regroupés en quatre catégories :

— **Les intermédiaires d'assurance** : ils sont mandatés par une ou plusieurs compagnies d'assurance pour la commercialisation des produits d'assurance. Depuis 2007, ils ont l'obligation de s'enregistrer auprès de l'ORIAS (Organisme pour le Registre unique des Intermédiaires en Assurance, Banque et Finance). Ils doivent informer et conseiller les clients avant la souscription d'un contrat.

— **La vente directe** : développée par les sociétés mutuelles d'assurances, il s'agit d'une distribution sans intermédiaire qui est apparue dans les années 1970.

— **Les bancassureurs** : ce sont des réseaux bancaires qui font la distribution de produits d'assurance. En effet, les banques utilisent leur proximité avec les clients pour leur proposer des produits d'assurance. Fortement présente sur le marché de l'assurance vie, cette forme de distribution se développe également en assurance dommages des particuliers et professionnels.
— **La distribution par internet** : utilisée par toutes les sociétés d'assurance, ce mode de distribution est privilégié par les compagnies qui commercialisent en direct telles que Direct Assurance.

Ces canaux ne se développent pas de la même façon et les études comparatives montrent qu'ils se basent sur des stratégies différentes. Les intermédiaires, les bancassureurs et la vente directe utilisent la proximité et le devoir de conseils pour attirer les clients alors que la distribution sur internet s'appuie sur la facilité d'accès au tarif pour avoir une clientèle qui peut être très volatile.

La concurrence et la nécessité de performance obligent les assureurs à privilégier la profitabilité. La priorité est donc la conservation du portefeuille et sa rentabilité. En effet les transformations digitales ont facilité l'accès au tarif en ligne. Les clients sont alors plus sensibles au prix et changent plus facilement d'assureurs. Il est donc capital de fidéliser le client à travers le multi-équipement (qui garantit une meilleure duration du contrat). Cette stratégie est adaptée à l'approche des intermédiaires qui misent sur le développement d'une relation avec le client pour le fidéliser. C'est ce qui justifie les meilleurs résultats des réseaux d'intermédiaires par rapport aux autres modes de distribution.

De plus, les coûts de distribution (voir l'étude comparative [Eur14] à la figure 1.1) montrent que contrairement aux idées reçues, les intermédiaires d'assurance compensent leurs frais d'acquisition élevés par la qualité de leur souscription. Ils ont donc des coûts moins élevés par rapport à d'autres modes de distribution.

Bien que le marché de la distribution soit de plus en plus concurrentiel, le réseau des intermédiaires reste encore un mode de distribution privilégié qui favorise le multi-équipement. Il s'agit du canal de distribution le plus utilisé par AXA France pour la distribution des produits IARD.

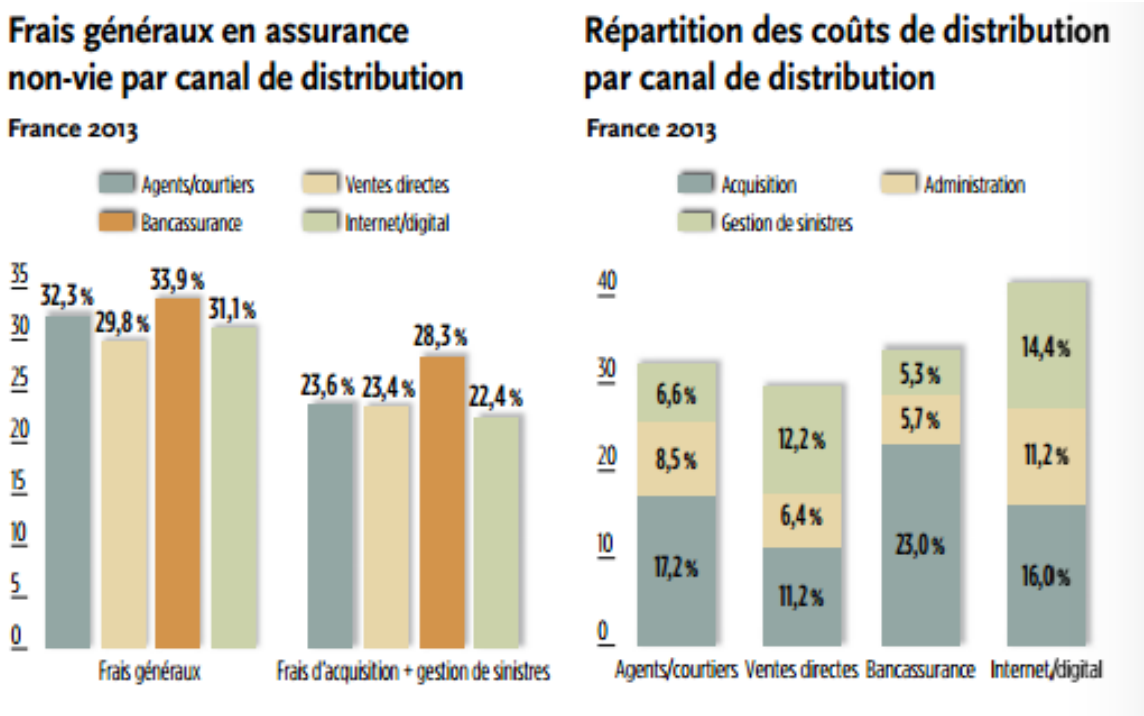

FIGURE 1.1 – Étude comparative des canaux de distribution

1.1.2 AXA France et la direction de l'offre IARD

AXA France est la filiale la plus importante du groupe AXA avec près de 25% du chiffre d'affaires. C'est le premier assureur en France avec près de 33 000 collaborateurs au service de 7,3 millions de clients. AXA France dispose de trois entités opérationnelles qui lui permettent de couvrir l'ensemble des branches d'assurances :

— **AXA Particuliers et IARD Entreprises (AXA PIE)** qui regroupe les activités IARD, Épargne et Prévoyance Individuelles ainsi que la Distribution et la vente au travers des différents réseaux concernés.

— **AXA santé et collectives** qui regroupe l'ensemble des activités Collectives (Santé, Prévoyance, Epargne et Retraite en France et à l'international) et la santé Individuelle.

— **Directions Marques et Client Multi-accès** qui a vocation à intervenir, de façon transverse, sur l'expérience client, le Digital, la Data, la Marque, les Partenariats.

Pour une meilleure organisation des activités, chaque entité est divisée par sous-entité d'activités. Ainsi l'entité AXA PIE possède en son sein une sous-entité IARD Particuliers et Professionnels qui gère la conception des produits, leur distribution ainsi que les engagements. L'une des directions de cette sous-entité IARD Particuliers et Professionnels est la direction de l'offre.

Cette direction a pour mission :
— la définition de la politique tarifaire des produits IARD Particuliers et Professionnels ;
— la tarification des produits auto, dommages particuliers et multirisque professionnelle ;
— la définition de l'offre, de la stratégie et des opérations commerciales pour la commercialisation des produits auto, dommages particuliers et professionnels et MRP ;
— le suivi des performances et veille à la rentabilité des portefeuilles ainsi que l'accompagnement nécessaire aux distributeurs.

Il s'agit donc d'une direction qui dispose des leviers nécessaires pour développer l'IARD particuliers et professionnels d'AXA France.

1.1.3 Les intermédiaires de distribution d'AXA France

AXA France s'appuie sur des réseaux de distribution assez diversifiés pour la commercialisation de ses produits. Ils sont organisés sur la base de la collaboration avec AXA ou suivant le type de produits distribués :

— **Les agents indépendants et salariés commerciaux** : ils sont présents sur tout le territoire. Ils distribuent exclusivement les produits AXA et sont regroupés suivant les types de produits qu'ils commercialisent :

  Agents Généraux AXA (AGA) : ils sont constitués uniquement d'indépendants et font la distribution de tous les produits AXA. C'est le canal le plus utilisé pour la distribution des produits IARD.

  Agents généraux Prévoyance Patrimoine (A2P) : ce réseau est spécialisé dans le patrimoine et la prévoyance.

  Conseillers AXA Épargne et Protection (AEP) : contrairement aux AGA et A2P, ce sont des salariés AXA spécialisés dans le conseil et la distribution des produits d'épargne et de protection.

— **Courtiers :** ce sont des indépendants qui ne commercialisent pas que des produits AXA.

Les produits IARD particuliers et professionnels sont surtout commercialisés par les agents généraux comme le montre la répartition du chiffre d'affaires IARD Particuliers Professionnels (figure 1.2).

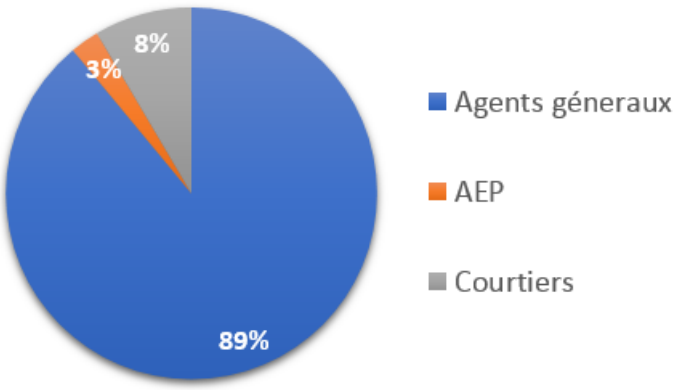

FIGURE 1.2 – Répartition CA IARD PP par canal de distribution

1.2 Les agents généraux AXA

Ce sont des indépendants qui achètent auprès d'AXA un ou plusieurs portefeuilles de contrats qu'ils se chargent de développer tout au long de leur carrière. Leur objectif premier est de faire fructifier leur investissement de façon à s'assurer un revenu et à garder un portefeuille qui sera valorisé à leur départ en retraite. Ils sont donc commissionnés sur leur production et bénéficient d'un accompagnement de la part d'AXA.

Cet accompagnement les aide à améliorer leur production mais également à comprendre et à mettre en oeuvre la politique de souscription d'AXA. L'accompagnement est assuré par par les directeurs régionaux de vente (DRV) et les inspecteurs commerciaux qui quadrillent le territoire.

Ils sont également appuyés dans cette tâche par des équipes AXA qui analysent la production de chaque agent en vue d'un suivi personnalisé. Ces analyses visent à identifier des points d'amélioration qui vont garantir la rentabilité du portefeuille.

Les agents peuvent donc compter sur leur pyramide commerciale (DRV et inspecteurs) pour vite s'approprier les nouveaux produits mais aussi pour remonter leurs difficultés. Quant à la direction, cet accompagnement permet d'avoir un suivi de proximité et de s'assurer que le développement des agents soit cohérent avec la stratégie globale du groupe. En effet, le système de commissionnement et la volonté de rentabiliser leur investissement poussent chaque agent à avoir ses propres stratégies de développement. Ils peuvent donc se mettre en association avec d'autres agents pour construire un portefeuille plus solide et mutualiser les coûts, comme ils peuvent s'entourer de plusieurs collaborateurs pour accélérer le développement.

En plus de ces accompagnements, les agents généraux disposent de divers leviers dont

ils peuvent se servir pour baisser la prime commerciale et attirer de nouveaux clients.

1.2.1 Les budgets commerciaux des AGA

Les budgets commerciaux constituent le principal moyen utilisé par la direction pour accompagner les agents afin de favoriser leur développement. Ces budgets constituent des leviers permettant aux agents de faire des gestes commerciaux pour :

— fidéliser les clients représentant un contexte commercial qu'il est important de préserver ;
— favoriser le multi-équipement de nouveaux clients.

Il existe deux types de budgets commerciaux à la disposition des agents sur le marché IARD particuliers et professionnels :

— La Souplesse Tarifaire (ST) : elle est utilisée pour la branche auto sur les produits auto 4 roues particuliers et moto sans les formules commerciales. Il s'agit d'une dotation que la direction fait à chaque agent en se basant sur son résultat technique et son chiffre d'affaires dans la branche auto. Cette dotation est utilisée par l'agent pour faire des rabais sur le produit auto. La souplesse tarifaire représente environ 4% du Chiffre d'affaires auto.
— L'Allocation Régionale de Crédit Commercial (ARCC) : elle est utilisée pour les autres branches IARD non auto à l'exception de la prévoyance et la santé. Elle est octroyée suivant le chiffre d'affaires et les résultats techniques dommages particuliers et MRP. Elle représente 3% du chiffre d'affaires dommages particuliers.

Bien que les budgets commerciaux représentent un levier à la disposition des agents, la direction a défini des règles afin de garantir une utilisation judicieuse. Ces règles permettent de limiter une mauvaise utilisation de ces budgets.

Types d'utilisation :

Les budgets commerciaux peuvent donc être utilisés de diverses manières et l'impact n'est pas le même suivant le type d'utilisation que le distributeur choisit. Il peut choisir d'utiliser le budget sur une cotisation ou sur un sinistre :

- **Utilisation du budget commercial sur une cotisation** :
  Elle consiste à accorder une réduction à un client ou à un prospect afin de baisser sa cotisation. Cette utilisation s'applique à la cotisation d'une affaire nouvelle, d'un remplacement ou d'un terme. Elle peut être une utilisation ponctuelle ou une utilisation reconductible.
  L'utilisation ponctuelle vise à offrir un tarif d'appel à un prospect ou à un client sans antécédent. Dans ce cas, la direction prend en charge une partie de la réduction que l'agent accorde. Les impacts du crédit ponctuel sont annuels et ne vont pas au-delà.

Mais l'agent peut également utiliser son crédit commercial pour réduire durablement la cotisation d'un client. Dans ce cas, il s'agit d'une utilisation reconductible du budget commercial. Cette utilisation est supportée entièrement par l'agent. Ce qui garantit une meilleure maîtrise de son impact sur les ressources.
Le budget commercial représente donc un levier qui permet à l'agent de baisser à court terme ou à long terme la cotisation d'un client. Mais il peut être également utilisé dans le cadre d'un sinistre :

- **Utilisation du budget commercial sur un sinistre** :
  Il permet au distributeur de faire un geste commercial à un client qui fait face à un sinistre. Cette utilisation a pour but de soutenir les bons clients.

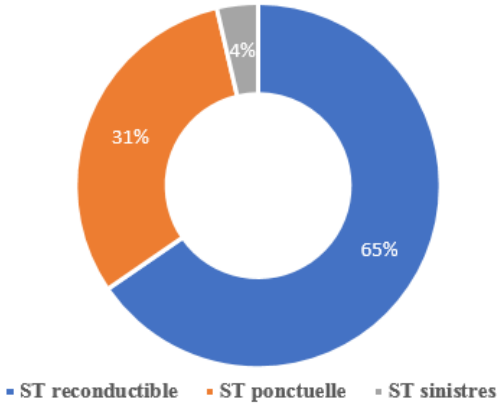

FIGURE 1.3 – Utilisation de la ST

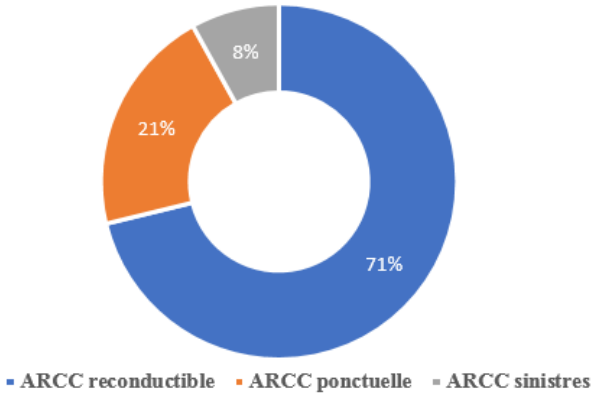

FIGURE 1.4 – Utilisation de l'ARCC

L'utilisation est globalement orientée vers la cotisation plutôt que les sinistres et varie selon les agents. Ils peuvent l'utiliser sur les affaires nouvelles ou plutôt sur le terme. Cette répartition dépend de l'agent et peut influer sur la rentabilité des contrats qu'il souscrit. En effet, si l'agent accorde plus de souplesse tarifaire pour attirer un profil jugé risqué à travers le tarif, cela va impacter la rentabilité de son portefeuille. Il est donc important de contrôler l'allocation de budget et l'usage que les agents en font. Le principe actuel de dotation est basé sur le chiffre d'affaires et le ratio combiné. Ce qui permet de réguler la dotation en fonction des résultats et également du poids de l'agent.

Cependant cette approche ne permet pas de contrôler l'utilisation qu'en fait l'agent et ne tient pas compte de l'adéquation de la dotation avec les besoins de l'agent. En effet, les agents n'ont aucune obligation d'utilisation du budget qu'ils reçoivent et peuvent constituer une réserve assez conséquente de budget. Cette réserve échappe au contrôle d'AXA (car AXA ne peut plus contrôler le montant total de budgets que l'agent peut utiliser sur une année) et l'agent peut alors l'utiliser pour faire des réductions. De plus, malgré cette réserve, la direction peut continuer de les doter. Ce qui pose un problème d'adéquation entre le besoin de l'agent et la dotation qu'on lui accorde.

Le budget commercial peut donc constituer un réel moyen permettant aux agents d'impacter positivement ou négativement les résultats techniques. Il faut donc tenir compte du besoin en budget de l'agent et de l'usage qu'il en fait. Le ratio combiné est le moyen qu'utilise la direction pour contrôler ce levier.

1.2.2 Le ratio combiné protocolaire

Le ratio combiné protocolaire est un indicateur de performance qui permet à une compagnie d'assurance d'évaluer la rentabilité technique d'un portefeuille. C'est le rapport entre ce que la compagnie dépense et ce qu'elle encaisse au titre de ce portefeuille. Son calcul n'intègre pas tous les postes de charges d'un compte de résultat compagnie (pas de coût de la réassurance). Le ratio combiné protocolaire est égal à la somme des trois éléments suivants :

— le rapport sinistres à cotisations écrêtées et mutualisées : c'est le coût des sinistres par rapport au montant des primes (S/P).
  Les sinistres prennent en compte les indemnités versées aux assurés mais également l'estimation des charges des sinistres en cours (sinistres connus mais non réglés, sinistres survenus mais pas encore connu et le montant du réajustement des sinistres passés).
  Son écrêtement-mutualisation permet de le stabiliser dans le temps en lissant la charge des sinistres importants au niveau national.
— le taux moyen de rémunération comprenant les commissions versées aux intermédiaires (agents, courtiers...) et allocations.
— le taux de frais généraux : c'est le rapport entre les frais de gestion et les primes perçues.

Il est calculé tous les mois, sur 36 mois de déclaration de sinistres (quelle que soit la survenance), vieillis de 3 mois.

FIGURE 1.5 – Calcul du ratio combiné

Il permet de piloter la performance de chaque agent et est utilisé pour :

— l'attribution et la révision des délégations techniques ;
— le calcul de l'intéressement technique de l'agence.

Le ratio combiné est calculé pour chaque branche et traduit surtout la situation passée d'une agence. Il ne permet pas d'identifier le potentiel d'un portefeuille. Il joue tout de même un rôle capital pour l'agent car il sert de base au calcul de son budget commercial mais également au calcul d'un autre indicateur qui détermine la tarification auto appliquée par chaque agent à ses clients. Il s'agit du coefficient technique personnalisé.

1.2.3 Coefficient Technique Personnalisé (CTP) :

A travers les budgets commerciaux, nous avons montré que les agents n'accordent pas les rabais de la même façon. Cette situation montre qu'il y a un comportement agent qui agit sur le tarif qu'il propose à chaque client. Des études antérieures ont permis de définir un indicateur qui caractérise chaque agent. Cet indicateur est fonction de la rentabilité technique du portefeuille de l'agent et permet de valoriser un agent par rapport à ses pairs.

Le coefficient technique personnalisé (CTP) est donc un indicateur de qualité technique d'un agent. Il permet de fournir un avantage tarifaire à une agence ayant des résultats plus favorables que la moyenne des agents. Il est déterminé en classant les agents selon leur ratio combiné, en leur attribuant une note de 1 à 100. Une fois le CTP fixé par AXA, on parle alors de CTP de base. Il détermine la part de la prime commerciale qu'appliquera l'agent pour ses clients. Ainsi, pour le même profil, le tarif appliqué varie en fonction du CTP de l'agent. Le profil aura un tarif plus bas si l'agent a un CTP faible.

Le CTP constitue un double levier à la main de la direction et des agents :

— Utilisation du CTP pour favoriser le développement des agents ayant de bons résultats.
  En effet, il est utilisé par la direction pour permettre aux agents ayant de bons résultats techniques de bénéficier d'un tarif préférentiel par rapport aux agents ayant de mauvais résultats. Il s'agit donc d'un indicateur qui oblige les agents à avoir de bons résultats afin de bénéficier de tarifs préférentiels.
— Utilisation du CTP pour augmenter le budget commercial de l'agent. L'agent peut acheter des points supplémentaires pour augmenter son CTP (on parle alors de CTP effectif) et en contrepartie avoir plus de budget de souplesse tarifaire. Dans ce cas, l'agent décide de réduire l'avantage qu'offre un bon CTP pour disposer de plus de budget qu'il pourra utiliser à sa guise.
  L'achat des points CTP, bien que limité à trois points, permet à chaque agent peut revoir le tarif à la hausse afin d'avoir plus de budget. La majorité des agents achète des points et transforme ces points en souplesse tarifaire (la figure 1.6 montre le CTP de base par rapport au CTP effectif). Cela peut entraîner une baisse de la

production car les tarifs seront plus élevés. Ce choix laissé aux agents de piloter dans une certaine mesure le tarif à la hausse peut être contre-productif.

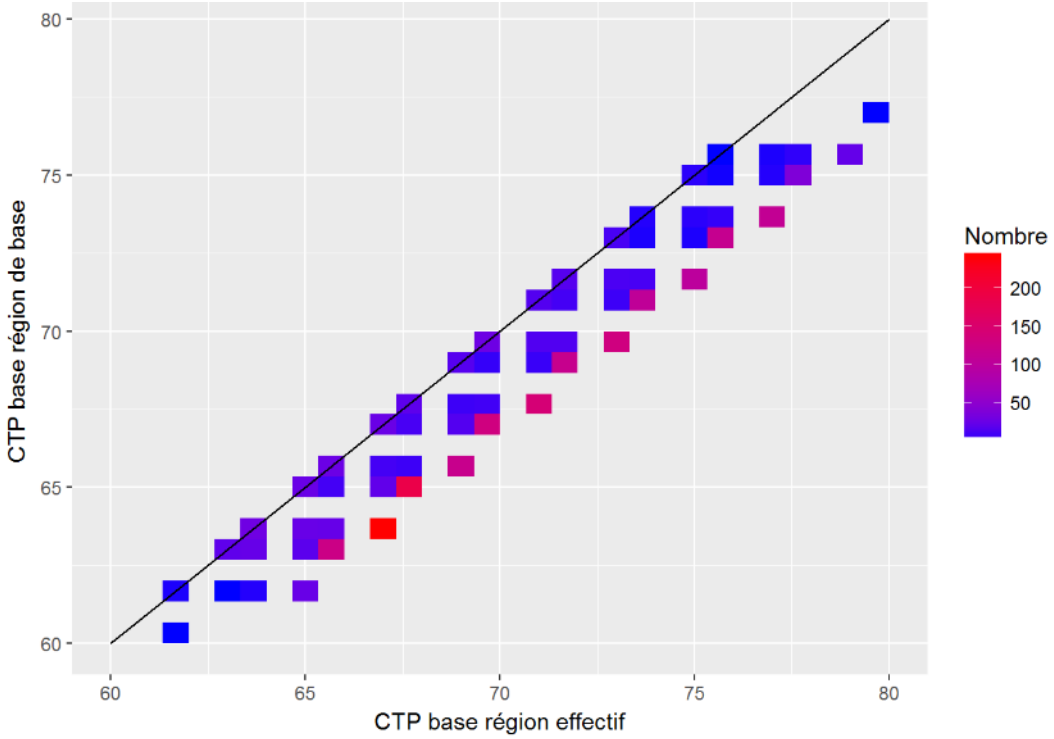

FIGURE 1.6 – CTP de base par rapport au CTP effectif

Le CTP bien que limité au produit auto permet donc à la direction de moduler la prime commerciale suivant le comportement de chaque agent.

Les budgets commerciaux et le CTP montrent que les agents ne sont donc pas de simples distributeurs de produits AXA. Ils disposent de quelques leviers qu'ils utilisent pour améliorer leur production. Bien que la direction essaye de contrôler l'utilisation de ces leviers à travers le pilotage par le ratio combiné de l'agent, ces leviers ne sont pas toujours bien utilisés et ne garantissent pas forcément une amélioration des performances commerciales et techniques. Ces mauvais usages peuvent entraîner un risque de souscription ou d'antisélection.

1.2.4 Les risques liés aux comportements des agents

Les agents utilisent les leviers selon leur stratégie et leur contexte. Ainsi, le comportement de chaque agent détermine son utilisation et peut donc influencer son résultat. Cette différence comportementale n'est pas prise en compte dans l'accompagnement des agents mais elle impacte leur production et les résultats techniques. Ces comportements peuvent

engendrer différents risques.

1.2.4.1 Le risque d'antisélection :

L'antisélection est un phénomène très répandu en assurance qui caractérise le dysfonctionnement résultant de l'asymétrie d'information.

La mauvaise utilisation des budgets commerciaux par un agent peut engendrer un phénomène d'antisélection. Un agent peut, dans l'optique de produire davantage, mettre beaucoup de rabais pour attirer un profil A risqué. Ce qui va baisser la prime du profil A. Cette prime peut alors être plus faible que la prime d'un profil B moins risqué pour lequel l'agent ne fait aucun rabais. Cette utilisation du budget pour attirer des profils risqués va engendrer de mauvais résultats techniques. Ce qui aura pour conséquence une augmentation générale du tarif même pour le profil B qui est moins risqué. Cette situation va pousser le profil B à résilier pour souscrire ailleurs où il bénéficiera d'un tarif plus juste. Notre agent va progressivement perdre ses bons profils et pour maintenir son développement attirer uniquement des profils risqués (en leur accordant du rabais). Ainsi le budget commercial aura servi à rendre encore plus risqué le portefeuille de l'agent. Il s'agit donc d'antisélection due à la mauvaise utilisation du budget.

1.2.4.2 Le risque d'une mauvaise utilisation de la délégation

La délégation en souscription est le fait de déléguer au distributeur le pouvoir d'effectuer la souscription de certains risques sans demander l'accord des services d'engagements. Les agents ayant une délégation tarifaire peuvent aller au-delà du seuil fixé de rabais sans avoir l'accord des services d'engagements. Ces délégations de souscription ont été mises en place pour faciliter le travail de l'agent. Mais elles enlèvent à la direction le pouvoir de contrôler certains profils avant la souscription. Ce pouvoir est donc délégué à l'agent qui devient responsable de la sélection du risque.

Les agents en délégations tarifaires font donc une souscription directe même lorsqu'ils dépassent le seuil de rabais fixé. Ces agents peuvent donc dans le but de produire et de générer plus de commissions mettre plus de budgets sur des profils risqués. La délégation peut donc entraîner une augmentation du risque lorsque les agents ne font pas bon usage des leviers de rabais que la direction met à leur disposition.

Ces risques constituent des limites au système actuel et ne garantissent pas une amélioration de la rentabilité des portefeuilles.

1.3 Problématique de l'étude

Objectif de l'étude

Nous avons vu précédemment que les agents généraux constituent le socle de la production IARD. Ils sont aidés par la direction qui met à leur disposition différents accompagnements et leviers pour booster leur production. La direction utilise le ratio combiné pour surveiller la rentabilité de chaque agent. Le CTP permet aussi de favoriser les agents ayant de bons résultats. Cependant, les agents à travers leur comportement peuvent toujours influencer les résultats.

Cette influence du distributeur sur les résultats techniques n'est pas prise en compte en amont dans la tarification et elle dépend de chaque agent et de sa situation. Le pilotage actuel des agents bien qu'il s'appuie sur les résultats techniques, n'intègre pas la stratégie de chaque agent. De plus, les mesures d'accompagnement ne tiennent pas compte du profil de chaque agent et ne garantissent pas une profitabilité de leur portefeuille.

Face à cette situation, comment pourrait-on améliorer l'accompagnement des agents ? Comment peut-on rendre cet accompagnement plus efficace en l'adaptant à la stratégie et au besoin de chaque agent ? Comment pourrait-on s'appuyer sur le comportement et les risques techniques de chaque agent pour différencier les mesures d'accompagnement ? Enfin comment pourrait-on utiliser les données agents pour rendre plus efficace ces mesures tout en permettant à chaque agent d'améliorer la rentabilité de son portefeuille ?

Une réponse est de créer des profils d'agents basés sur les comportements et l'utilisation des budgets commerciaux puis de proposer une stratégie leur permettant d'améliorer leur rentabilité.

L'objectif de notre étude est donc de classer les distributeurs en populations homogènes de comportements et de risques techniques dans le but de différencier les mesures d'accompagnement.

Démarche de l'étude

Notre démarche pour effectuer une classification des agents consistera à identifier et créer dans un premier temps des indicateurs permettant de catégoriser les agents. Ces indicateurs nous serviront à constituer la base de données de notre étude.

Nous utiliserons ensuite une méthode de réduction de dimensions (ACP), tout en veillant à ce que cette méthode préserve une information suffisante dans chacune des dimensions suivantes :

— la production ;
— la maîtrise des coûts ;
— l'utilisation des budgets commerciaux.

Ensuite, nous utiliserons la nouvelle base constituée pour faire une classification non supervisée. Cette classification nous permettra d'attribuer un groupe à chaque agent et de décrire chaque groupe.

Enfin, nous allons utiliser le résultat de cette classification non supervisée pour définir une nouvelle variable à expliquer. Une classification supervisée ayant comme variable de réponse cette nouvelle variable permettra de faire une meilleure caractérisation des groupes obtenus et une généralisation de la segmentation. Nous proposerons des mesures d'accompagnement basées sur les caractéristiques de ces groupes.

Chapitre 2

Construction de la base de données

Sommaire

2.1 Périmètre de l'étude **17**
2.2 Choix des indicateurs **18**
2.2.1 Indicateurs de production 18
2.2.2 Indicateurs de maîtrise des coûts 20
2.2.3 Indicateurs d'utilisation des budgets commerciaux 20
2.3 Analyse exploratoire des données **21**
2.3.1 Principe général de l'ACP : 21
2.3.2 Réduction de dimensions avec l'ACP 25
2.4 Construction de la base de segmentation **32**

La segmentation de nos distributeurs en groupes homogènes exige qu'on identifie des indicateurs capables de faire ressortir leur comportement. Nous allons dans ce chapitre définir le périmètre de l'étude, puis présenter les indicateurs ou variables qui vont permettre de caractériser les agents. Nous finirons par la réduction de dimensions à l'aide de l'ACP qui permettra de construire une nouvelle base de données pour la classification.

2.1 Périmètre de l'étude

L'étude va se porter sur la branche IARD particuliers professionnels et plus particulièrement sur les produits Auto et Dommages particuliers. Ce choix s'explique par le fait que ces deux produits représentent la majorité du chiffre d'affaires des distributeurs en IARD particuliers professionnels. De plus, les leviers utilisés par AXA pour accompagner les distributeurs se basent sur les résultats de ces deux produits.

Nous allons considérer tous les distributeurs en activité pour notre étude. Nous avons donc 3 132 agents généraux qui vont être segmentés. Nous avons choisi de considérer leur situation au 31 décembre 2017.

Les données utilisées proviennent des bases suivantes :

— Base de production : c'est la base qui recense les données de production telles que la taille de production, le chiffre d'affaires, les affaires nouvelles, les résiliations ... par produit et par portefeuille.
— Base de suivi de ratio combiné : c'est la base qui permet le pilotage des résultats techniques par distributeur.
— Base de suivi de budget : elle permet d'analyser et de suivre l'utilisation du budget commercial ainsi que les rabais qu'accorde le distributeur.

Nous utiliserons les données 2016 et 2017 pour construire les indicateurs de comportements et les données 2015 à 2017 pour le ratio combiné. Le choix de n'utiliser que les données 2016 et 2017 pour les indicateurs s'explique par le fait que nous ne disposons pas d'assez d'historiques sur les données. De plus la situation des distributeurs évolue très vite car ils peuvent céder ou acquérir de nouveaux portefeuilles. Ce qui implique une modification de la situation du distributeur. Il est donc important, dans l'optique d'avoir une segmentation qui se base sur la situation présente du distributeur, d'utiliser des données actualisées.

Une fois que la base de données est créée, nous allons construire une autre base de données avec les années 2015-2016 qui servira de base de test. Cette base de test sera utilisée pour identifier le groupe dans lequel l'agent aurait été si on se basait sur les données 2015-2016. Ce test permettra de conclure sur la stabilité du modèle de segmentation que nous construirons.

2.2 Choix des indicateurs

A partir du contexte opérationnel actuel, que nous avons présenté au chapitre précédent, la production apparaît comme un facteur clé dans la stratégie de chaque agent. En effet, elle conditionne le développement de l'activité. De plus, la rentabilité et plus globalement la maîtrise des coûts par le distributeur permettent également de se projeter sur son comportement et son impact sur les résultats techniques. Enfin, l'utilisation du budget permet de mieux comprendre sa stratégie. Nous avons donc utilisé ces trois aspects comme dimensions qui permettront de classifier les agents. Nous avons alors construit des indicateurs qui permettent de bien cerner chacune de ces dimensions.

2.2.1 Indicateurs de production

Nous entendons par production, la dynamique commerciale de l'agent. Il s'agit de la capacité de l'agent à faire plus d'affaires nouvelles qu'il n'en perd. Diverses variables per-

mettent d'expliquer la production d'un distributeur. Nous avons donc construit à partir de ces variables des indicateurs qui nous permettent de caractériser la capacité d'un agent à :

— se procurer des opportunités (à travers le nombre de devis qu'il réalise par rapport à sa taille) ;
— transformer les devis ;
— multi-équiper ses clients (à travers le nombre de devis clients qu'il transforme par rapport au nombre de devis clients réalisés) ;
— faire de la croissance à travers son apport net ;
— faire plus d'affaires nouvelles et perdre moins de contrats ;
— augmenter le nombre d'affaires nouvelles qu'il réalise d'une année à l'autre ;
— utiliser les nouveaux produits (Clic and Go [1] et Ma Maison [2]).

Ces indicateurs de production étant fortement liés à la taille du portefeuille, nous allons également considérer la taille du portefeuille ainsi que le chiffre d'affaires global IARD. Le tableau ci-dessous récapitule les indicateurs de production que nous avons jugés pertinents pour caractériser la production d'un distributeur. L'annexe A présente plus en détails chaque indicateur ainsi que les raisons qui sous-tendent leur choix.

Variables	Descriptions
CA_IARD	Chiffre d'affaires IARD
PTF_IARD	Portefeuille IARD
TxCROIS_AutoP	Taux de croissance Auto
TxPROD_AutoP	Taux de production Auto
TxRESIL_AutoP	Taux de résiliation Auto
VarAN_AutoP	Variation des affaires nouvelles Auto
VarRes_AutoP	Variation des résiliations Auto
PartClicAndGO	Part de contrat Clic and GO
TxCROIS_DomP	Taux de croissance dommages particuliers
TxPROD_DomP	Taux de production dommages particuliers
TxRESIL_DomP	Taux de résiliation dommages particuliers
VarAN_DomP	Variation des affaires nouvelles dommages particuliers
VarRes_DomP	Variation des résiliations dommages particuliers
PartMaMaison	Part de contrat Ma Maison
TxTransf_Pros	Taux de transformation pour des devis prospects
TxTransf_Cli	Taux de transformation pour des devis clients
TxDevis	Taux de devis réalisés par le distributeur

Table 2.1 – Indicateurs de production

1. Clic and GO est le nouveau produit auto pour lequel le tarif du socle reste très concurrentiel et ne peut bénéficier d'aucun rabais. L'agent pourra mettre du budget commercial uniquement pour les options.
2. C'est l'équivalent de Clic and Go pour la MRH.

2.2.2 Indicateurs de maîtrise des coûts

Afin d'inclure dans notre analyse la rentabilité du portefeuille de l'agent, nous allons nous appuyer sur des indicateurs techniques qui montrent la maîtrise des coûts par le distributeur. Ces indicateurs permettent d'inclure dans notre étude les résultats techniques de l'agent ainsi que la sous-tarification qu'il fait à ses clients. Les indicateurs ci-dessous permettent de caractériser la maîtrise des coûts de l'agent.

La sous-tarification moyenne permet de mesurer quel est le taux de réduction que l'agent applique en moyenne à ses contrats. Cet indicateur permet de comprendre si l'agent arrive à attirer des clients avec moins de rabais ou non. Le lecteur est invité à se référer à l'annexe A pour plus de détails sur chaque indicateur.

Variables	Descriptions
RC_AUTO	Ratio combiné Auto
RC_MRH	Ratio combiné dommages particuliers
RC_TOT	Ratio combiné total IARD
Sous_Tarif_Auto	Sous-tarification moyenne Auto
Sous_Tarif_DomP	Sous-tarification moyenne dommages particuliers
CTP_BASE	Coefficient Technique Personnalisé de base régionale
PointsCTP	Nombre de points CTP achetés
PartCntRedAuto	Part de contrat auto ayant bénéficiée de réduction
PartCntRedDomP	Part de contrat dommages particuliers ayant bénéficiée de réduction

Table 2.2 – Indicateurs de maîtrise des coûts

2.2.3 Indicateurs d'utilisation des budgets commerciaux

Comme nous l'avons vu, le budget commercial est un levier dont dispose les agents pour booster leur production. Son utilisation peut être caractérisée par :

— le profil de client sur lequel l'agent utilise son budget. Nous pouvons donc distinguer l'utilisation sur les mono-détenteurs mais aussi l'utilisation sur des clients sinistrés. Ces deux cas qui peuvent constituer des usages inadéquats permettent de mieux catégoriser l'agent.
— la répartition que l'agent fait du budget. Quelle est la part utilisée pour défendre son portefeuille ? Cette utilisation du budget s'oppose à l'utilisation pour faire des affaires nouvelles.
— la réserve de budget dont il dispose à travers son solde et sa consommation
— l'impact du rabais sur son chiffre d'affaires à travers son taux de rabais.

Ces caractéristiques sont exprimées par les indicateurs récapitulés dans le tableau suivant :

Variables	Descriptions
TermeST	Part du budget commercial en ST utilisée pour le Terme
TermeARCC	Part du budget commercial en ARCC utilisée pour le Terme
MonoST	Part du Budget ST utilisée pour des mono-détenteurs
MonoARCC	Part du Budget ARCC utilisée pour des mono-détenteurs
CliSinST	Part de Budget ST utilisée pour des clients sinistrés
CliSinARCC	Part de Budget ARCC utilisée pour des clients sinistrés
TxRabais_AutoP	Taux de rabais en souplesse tarifaire (ST)
TxRabais_DomP	Taux de rabais en ARCC
UseST	Consommation rapportée à la dotation ST
UseARCC	Consommation rapportée à la dotation ARCC
NbreMois_ST	Nombre de mois de consommation de réserve en ST
NbreMois_ARCC	Nombre de mois de consommation de réserve en ARCC

Table 2.3 – Indicateurs d'utilisation des budgets commerciaux

Tous ces indicateurs et la pertinence de leur utilisation sont décrits plus en détails à l'annexe A. Cette sélection a permis également de noter le nombre important de ces indicateurs. Le choix de ces indicateurs s'explique par le fait qu'ils pourraient apporter plus d'informations pour mieux segmenter les distributeurs. Cependant, nous savons que cette logique suivant laquelle plus de variables permettraient de gagner en information ne garantit pas forcement la pertinence des informations. C'est ce qui nous amène à faire une analyse exploratoire plus détaillée de ces indicateurs.

2.3 Analyse exploratoire des données

L'analyse consistera à réduire le nombre de d'indicateurs de l'étude tout en maîtrisant la perte d'information. Pour faire une réduction de dimension avec une perte d'information maîtrisée, nous allons utiliser l'Analyse en Composantes Principales (ACP) :

2.3.1 Principe général de l'ACP :

L'ACP [Bac10] est une méthode d'analyse factorielle dont les débuts remontent en 1901 avec un article de Karl Pearson. Elle est généralement utilisée pour synthétiser l'information lorsqu'on a une quantité importante de données quantitatives à traiter. Elle permet la transformation de variables initialement corrélées en de nouvelles variables non corrélées.

Chaque nouvelle variable appelée composante principale est une combinaison linéaire des variables initiales. L'ACP est donc une méthode qui permet de condenser l'information en gardant les relations caractéristiques (proximités entre variables et individus).

Approche de l'ACP et construction des composantes principales [TRI13] :

Considérons p variables quantitatives observées sur n individus tel que le tableau de données X suivant :

$$X = \begin{bmatrix} x_{11} & ... & x_{1j} & ... & x_{1p} \\ ... & & ... & & ... \\ x_{i1} & ... & x_{ij} & ... & x_{ip} \\ ... & & ... & & ... \\ x_{n1} & ... & x_{nj} & ... & x_{np} \end{bmatrix}^t \tag{2.1}$$

Le nuage des points-individus est représenté par les coordonnées des n individus et chaque individu a pour coordonnées :

$$X_i = \begin{bmatrix} x_{i1} & ... & x_{ij} & ... & x_{ip} \end{bmatrix} \tag{2.2}$$

Le nuage des points-variables est représenté par les coordonnées des p variables et chaque variable a pour coordonnées :

$$^tX_j = \begin{bmatrix} x_{1j} & ... & x_{ij} & ... & x_{nj} \end{bmatrix} \tag{2.3}$$

La représentation du nuage d'individus et du nuage de variables permettent d'analyser les liens entre les individus ou les variables. Malheureusement, il est impossible de faire une telle représentation lorsque la dimension est grande.

Le but de l'ACP et des méthodes factorielles en général est de permettre une telle représentation tout en contrôlant la perte d'informations. L'ACP détermine donc un sous-espace E_q de faible dimension (q<=p), sur lequel est projeté le nuage de points du tableau de données de sorte à conserver les liens significatifs entre ces données.
E_q est alors le sous-espace permettant la dispersion (l'inertie) maximale du nuage projeté. Ce sous-espace est appelé espace factoriel. Les nouveaux axes sur lesquels le nuage est projeté sont appelés les composantes principales.

Soit I l'inertie totale du nuage, G le centre de gravité du nuage et m_i le poids de chaque individu. L'inertie étant la somme des distances des points au centre de gravité pondéré par le poids de chaque point, on a :

$$I = \sum_{i=1}^{n} m_i d^2(X_i, G) \tag{2.4}$$

Pour faciliter le calcul d'inertie, il faut transformer les données afin de ramener le centre de gravité à l'origine. De plus, l'inertie étant basée sur la distance, il faut normaliser les données afin d'annihiler les différences des unités ou des échelles de mesures de chaque variable sur la variance. On parle alors dans ce cas d'une ACP normée. Pour la suite, nous allons supposer que les individus ont le même poids et que notre tableau de données est normalisé, donc le centre de gravité est confondu à l'origine.

L'espace factoriel est constitué en déterminant itérativement ces composantes principales. On détermine la première composante principale qui représente la composante pour laquelle l'inertie projetée est la plus grande. Les autres composantes sont obtenues par la même approche. Considérons $\overrightarrow{Gu_1}$ le vecteur directeur de la première composante principale et I_1 l'inertie projetée sur cette composante, il faut :

$$I_1 \text{ maximale}$$

$$\text{Avec } \|\overrightarrow{u_1}\| = 1$$

Soit c_{i1} la projection orthogonale de l'individu X_i sur la première composante principale :

$$c_{i1} = {}^t\overrightarrow{X_1}.\ \overrightarrow{u_1}.$$

L'inertie projetée sur la première composante est alors :

$$I_1 = \sum_{i=1}^{n} c_{i1}^2 \tag{2.5}$$

Soit C_1 le vecteur de projection des n individus sur cette composante :

$${}^tC_1 = \begin{bmatrix} c_{11} & ... & c_{i1} & ... & c_{n1} \end{bmatrix} \tag{2.6}$$

On a donc :

$$C_1 = X.u_1 \tag{2.7}$$

On en déduit que :

$$I_1 =^t C_1.C_1 =^t u_1^t XXu_1 \tag{2.8}$$

Soit la matrice d'inertie S$={}^tXX$, l'inertie projetée peut se mettre sous la forme :

$$I_1 =^t u_1 S u_1 \tag{2.9}$$

Il faut donc choisir u_1 de sorte à maximiser ${}^t u_1 S u_1$. On utilise la méthode des multiplicateurs de Lagrange pour déterminer u_1. On pose donc :

$$\text{L}(u_1)={}^t u_1 S u_1\text{-}\lambda_1({}^t u_1 u_1 - 1)$$

La dérivée matricielle donne :

$$\frac{\partial L(u_1)}{\partial u_1} = 2Su_1 - 2\lambda_1 u_1$$

On a alors :

$$Su_1 - \lambda_1 u_1 = 0$$

$${}^t u_1 S u_1 -^t u_1 \lambda_1 u_1 = 0$$

Sachant que $\|\overrightarrow{u_1}\| = 1$, on en déduit :

$${}^t u_1 S u_1 =^t u_1 u_1 \lambda_1 = \lambda_1$$

D'où

$$I_1 = \lambda_1 \tag{2.10}$$

λ_1 est la valeur propre la plus grande de S.

On détermine ensuite la deuxième composante sur laquelle le nuage se déforme le moins en projection, après la première composante, tout en étant orthogonal au premier. On réitère jusqu'à l'obtention de p composantes. L'inertie de la kieme composante est obtenue par la même approche et l'inertie projetée sur la kieme composante est :

$$I_k = \lambda_k$$

Les composantes ainsi déterminées étant toutes orthogonales deux à deux, on a donc :

$$I = \sum_{k=1}^{p} I_k = \sum_{i=k}^{p} \lambda_k \tag{2.11}$$

Dans l'espace factoriel ainsi déterminé, les axes sont non corrélés et ne véhiculent pas la même information selon leur rang : leur capacité à « résumer » le nuage se détériore au fur et à mesure que l'on observe des axes de rang élevé.

2.3.2 Réduction de dimensions avec l'ACP

Notre objectif étant de garder de l'information suivant chacune de nos trois dimensions (production, maîtrise de coût et utilisation de budget), une ACP sur tous les indicateurs ne nous permet pas de garder ces informations. L'ACP va consolider toutes les informations et nous risquons de perdre l'information d'une dimension car moins influente par rapport à une autre dimension.

Nous allons donc effectuer une ACP par dimension (trois ACP au total). Ce qui nous permettra de conserver l'information de chaque dimension. Nous présenterons ci-dessous l'ACP sur les indicateurs de production. Les ACP sur les indicateurs de maîtrise de coût et celle des indicateurs d'utilisation de budget sont détaillées dans l'annexe B.

ACP sur les indicateurs de production

Corrélation des variables liées à la production :

Avant de faire une ACP sur la base de variables liées à la production, nous avons fait une matrice de corrélation qui nous a permis de porter un regard sur les variables. En effet, si tous les coefficients de la matrice de corrélation sont proches de 0, il n'y a aucun intérêt à faire une analyse en composante principale car l'ACP n'a de sens que s'il y a suffisamment de corrélations entre les variables. La matrice de corrélation des variables de la base est représentée sur le schéma 2.1 :

Cette matrice (figure 2.1) montre qu'il y a un intérêt à faire une analyse en composantes principales car les variables ne sont pas toutes corrélées entre elles et à l'opposé, les coefficients ne sont pas tous proches de 0. Nous pouvons donc faire une ACP afin de réduire le nombre de dimension. Nous constatons aussi que le chiffre d'affaires et le portefeuille sont fortement corrélés. Nous les gardons tout de même car le chiffre d'affaires permet d'avoir le poids des cotisations de l'agent alors que le portefeuille donne le nombre de contrats.

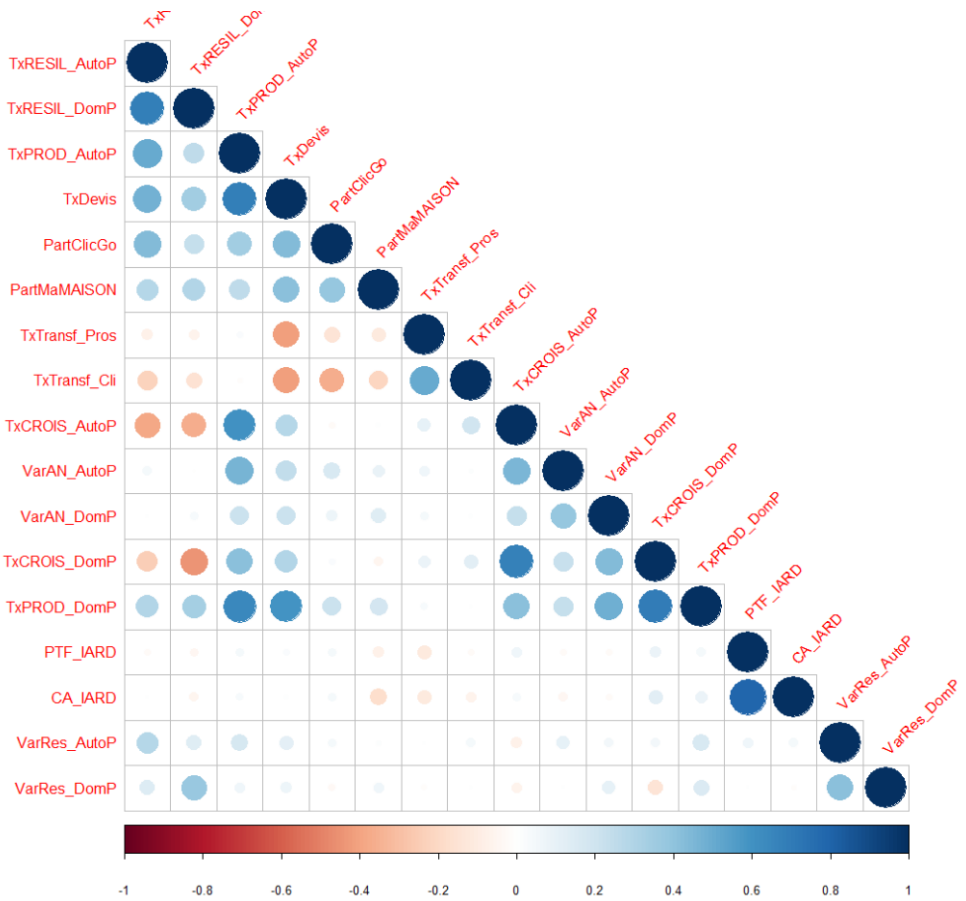

FIGURE 2.1 – Matrice de corrélation

Analyse en composantes principales :

Comme nous l'avons expliqué dans la section précédente, nous allons faire une ACP normée afin d'éviter qu'une variable prédomine sur les autres. Nous nous appuierons sur quelques sorties de R afin d'interpréter ces résultats et de choisir les composantes principales qui serviront à notre étude.

Choix du nombre de composantes à retenir :

Diverses techniques sont utilisées pour définir le nombre de composantes principales qui seront retenues. Dans le tableau ci-dessous, la valeur propre représente la quantité de variance expliquée par chaque composante principale. En examinant donc la proportion expliquée par chaque composante, Nous constatons que le premier plan (les deux premières composantes) ne porte que 46,13% soit respectivement 26,43% pour le premier axe et 19,69% pour le second. En nous appuyant sur le critère de Kaiser [3], nous pouvons

3. Critère de Kaiser : il propose d'utiliser les valeurs propres pour décider du nombre d'axes à retenir.

retenir les quatre premières composantes. En effet, chaque composante qui a une valeur propre supérieure à 1, représente plus de variance par rapport à une seule variable initiale. Ces quatre composantes portent 68.5% d'inertie expliquée, ce qui représente un bon seuil d'explication. Nous allons analyser davantage ces quatre axes et les interpréter afin d'en garder uniquement les composantes qui portent un sens business.

	Valeurs propres	Variances (%)	Variances cumulées (%)
Composante 1	3.96	26.43	26.43
Composante 2	2.96	19.69	46.13
Composante 3	1.91	12.78	58.92
Composante 4	1.44	9.57	68.5
Composante 5	0.95	6.38	74.89
...	...	...	...
Composante 17	0	0	100

Table 2.4 – Valeurs propres et variances

Interprétation des composantes :

Nous allons nous focaliser sur les quatre premières composantes et analyser quelles sont les indicateurs qui contribuent le plus à chaque composante ainsi que les individus qui participent également à la formation de cette composante. Pour apprécier cette participation, nous allons donc nous appuyer sur la contribution et la qualité de représentation de chaque variable. Le cercle de corrélation sera également utilisé pour comprendre les corrélations de chaque variable avec la composante à interpréter.

Première composante :

La figure 2.2 montre les dix variables qui contribuent le plus à la composante avec la ligne rouge qui représente la contribution moyenne au-dessus de laquelle la contribution est significative. On constate donc que les variables TxDevis (taux de devis), TxPROD_Auto, TxPROD_DomP (taux de production) sont celles qui participent le plus à la réalisation de cette composante.

En nous appuyant sur le cercle de corrélation du premier plan (composante 1 et composante 2), on s'aperçoit que certaines de ces variables sont bien représentées car elles sont proches du cercle. De plus, la composante isole le taux de devis et les taux de production. Ces variables sont isolées sur sa partie positive car elles sont positivement corrélées à la composante. Il s'agit d'un effet taille. Enfin, quand on regarde les individus qui contribuent le plus comme l'agent PHIL, on constate qu'il a un taux de production auto et dommages

Nous allons retenir uniquement les axes dont l'inertie est supérieure à l'inertie moyenne I/p. Kaiser en ACP normée : I/p= 1 et donc les composantes retenues sont celles dont la valeur propre supérieure à 1

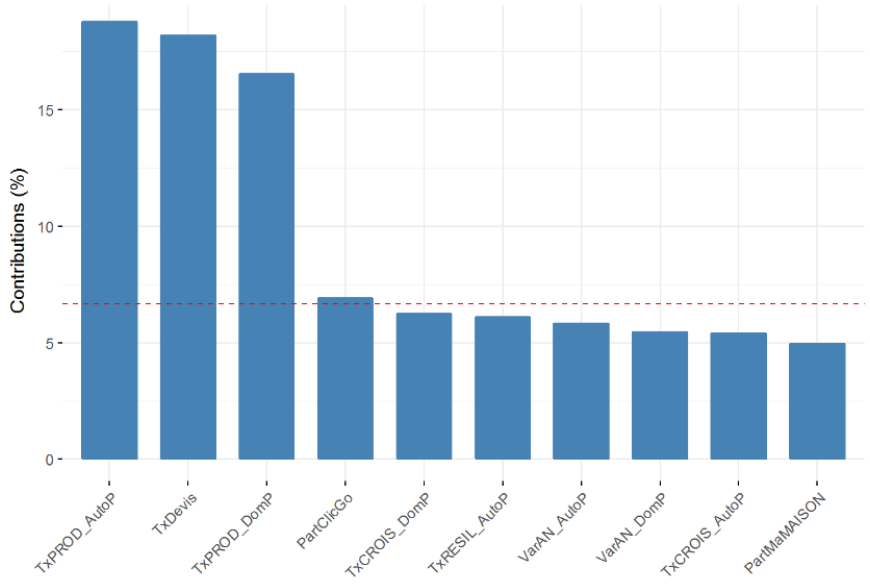

FIGURE 2.2 – Contribution composante 1

au-dessus de 30% avec un taux de devis à plus de 50%. Nous pouvons dire que cette composante exprime la **production**. Plus la valeur d'un individu est élevée sur cette composante plus il a une production forte.

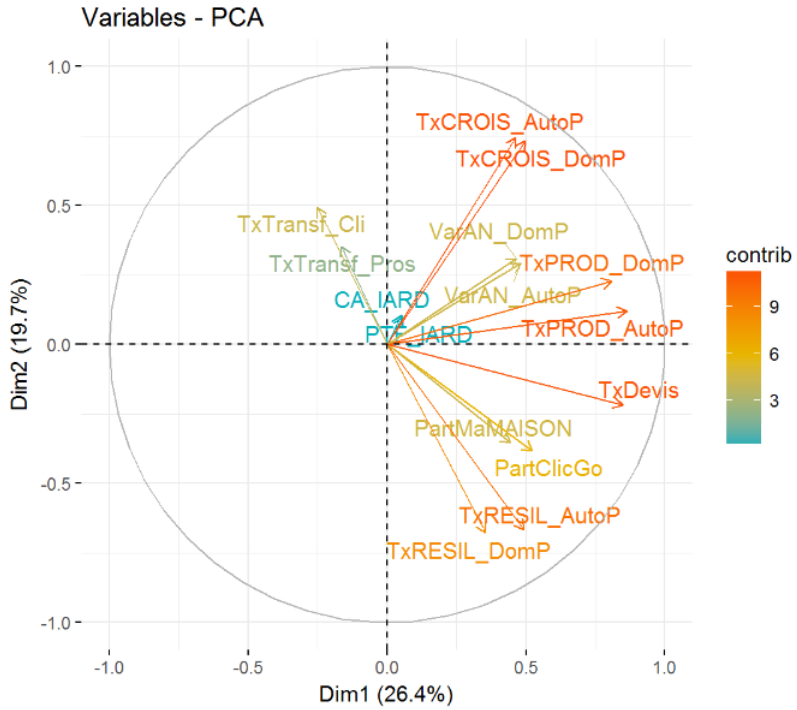

FIGURE 2.3 – Cercle de corrélation du premier plan

Deuxième composante :

La figure 2.4 montre la contribution des variables à la deuxième composante. On constate que TxCrois_AutoP, TxCrois_DomP (taux de croissance), TxRESIL_AutoP et TxRESIL_DomP (taux de résiliation) sont les variables qui participent le plus à la composante.

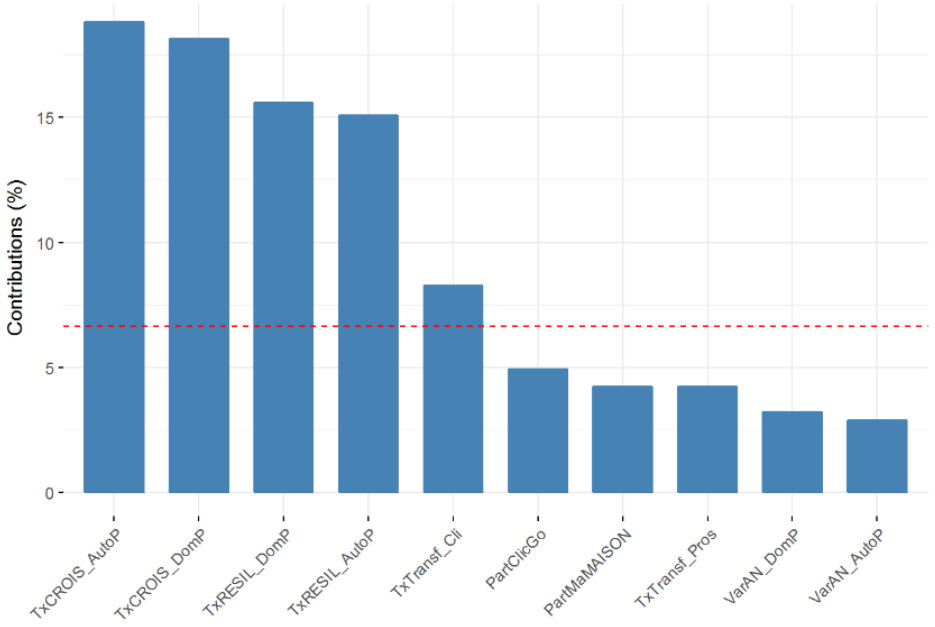

FIGURE 2.4 – Contribution composante 2

D'après le cercle de corrélation du premier plan, ces variables sont bien représentées sur le cercle de corrélation car elles sont toutes proches du cercle. De plus, les taux de résiliation sont négativement corrélés avec la composante alors que les taux de croissance sont corrélés positivement avec la composante. On en déduit que la composante oppose les taux de résiliation Auto et dommages particuliers aux taux de croissance Auto et dommages particuliers. Il s'agit d'un effet de forme. Enfin, en regardant les individus qui contribuent le plus à la composante, on constate que l'agent PIERRE a des taux de croissance auto et dommages à -23% et l'agent STEPH quant à lui a respectivement 17% et 23% pour ces deux taux. On a donc une composante qui oppose la croissance et le taux de résiliation. En d'autres termes, plus le taux de résiliation est faible, plus on a de la croissance. Nous pouvons donc conclure que la deuxième composante est un **axe de croissance et de résiliation**.

Troisième composante :

Comme l'indique la figure 2.5, les variables CA_IARD et PTF_IARD sont les deux

variables qui contribuent significativement à la construction de la composante.

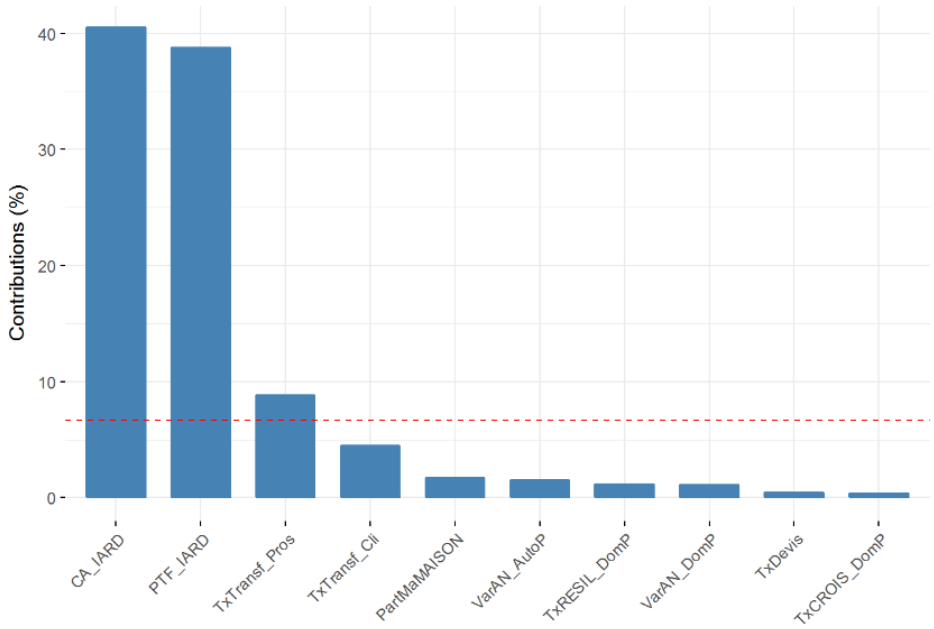

FIGURE 2.5 – Contribution composante 3

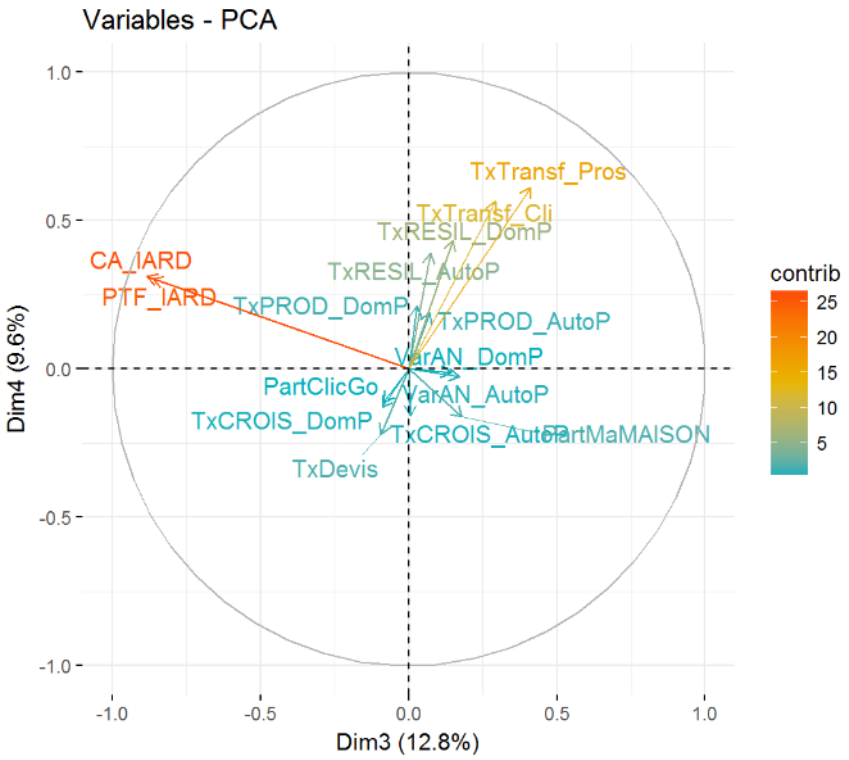

FIGURE 2.6 – Cercle de corrélation du second plan (composante 3 et composante 4)

De plus, ces deux variables sont très bien représentées sur le cercle de corrélation (voir figure 2.6) et sont toutes les deux corrélées positivement avec cette composante. Enfin les individus qui participent le plus à cette composante sont les distributeurs qui possèdent un grand portefeuille. Nous pouvons conclure que cette composante exprime la taille du portefeuille de chaque agent. Il s'agit d'un composant qui définit le poids de chaque distributeur. Nous pouvons le qualifier d'**axe de poids du distributeur**.

Quatrième composante :

En se basant sur la contribution de la quatrième composante (2.8 et le cercle de corrélation, on déduit que les variables TxTransf_PRO (taux de transformation prospect), TxTransf_CLI (taux de transformation client), TxRESIL_AutoP et TxRESIL_DomP sont les plus significatives à la construction de la composante. De plus, elles sont toutes corrélées positivement avec la composante. Ce qui signifie que les agents qui ont un taux de transformation élevé sont les agents qui subissent le plus grand nombre de résiliations. Nous pouvons dire qu'il s'agit d'un **axe de transformation de devis** et que la transformation de devis est également liée à la résiliation des contrats sur cette composante.

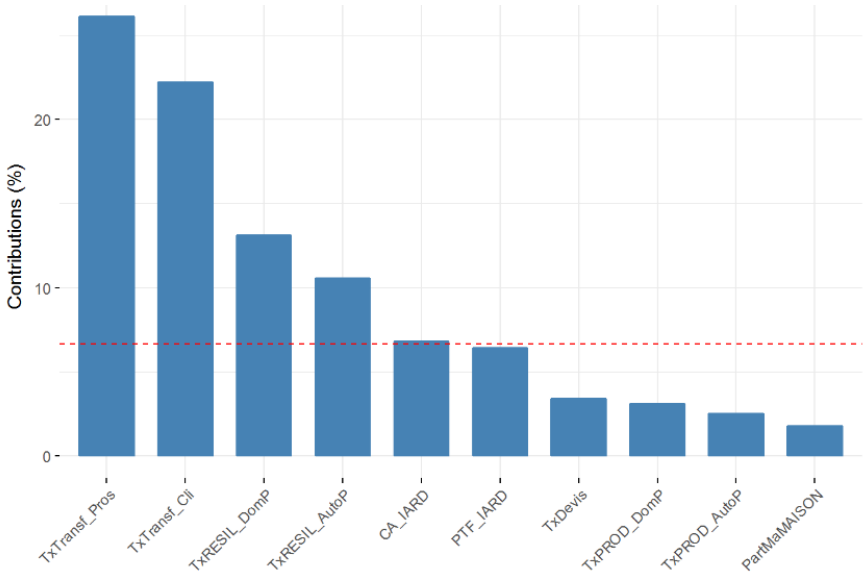

FIGURE 2.7 – Contribution composante 4

En somme, nous allons retenir pour notre étude les quatre premières composantes de l'ACP sur les indicateurs de production. Ces quatre composantes expriment :

- Composante 1 : la production ;
- Composante 2 : la croissance et la résiliation ;
- Composante 3 : le poids du distributeur ;
- Composante 4 : la transformation des devis et la résiliation.

ACP sur les indicateurs de maîtrise de coûts

L'analyse en composante principale est développée en annexe B . Nous retenons de cette analyse deux composantes qui portent plus de 63% de l'information. Ces composantes expriment :

- Composante 1 : la rentabilité et la réduction (sous-tarification moyenne et part d'affaires nouvelles avec réduction) en auto ;
- Composante 2 : la réduction en dommages particuliers.

ACP sur les indicateurs d'utilisation de budgets

L'analyse en composante principale est également développée en annexe B . Nous retenons de cette analyse quatre composantes qui portent 60% de l'information de cette dimension et expriment :

- Composante 1 : la réserve en nombre de mois de consommation ;
- Composante 2 : la répartition du budget ;
- Composante 3 : la part de budget utilisée pour le terme ;
- Composante 4 : le taux de rabais.

Cette réduction de dimension nous a permis de préserver au moins 60% de l'information suivant chacune des trois dimensions tout en réduisant le nombre de variables à 10.

2.4 Construction de la base de segmentation

Après avoir réduit les dimensions tout en conservant l'information de chaque dimension, nous avons regroupé les composantes des trois ACP pour constituer une nouvelle base que nous allons utiliser pour effectuer notre segmentation. La figure 2.8 présente l'approche que nous avons adoptée pour construire cette base. Nous garderons la base initiale (base AGA) que nous utiliserons par la suite pour caractériser les groupes issues de la segmentation.

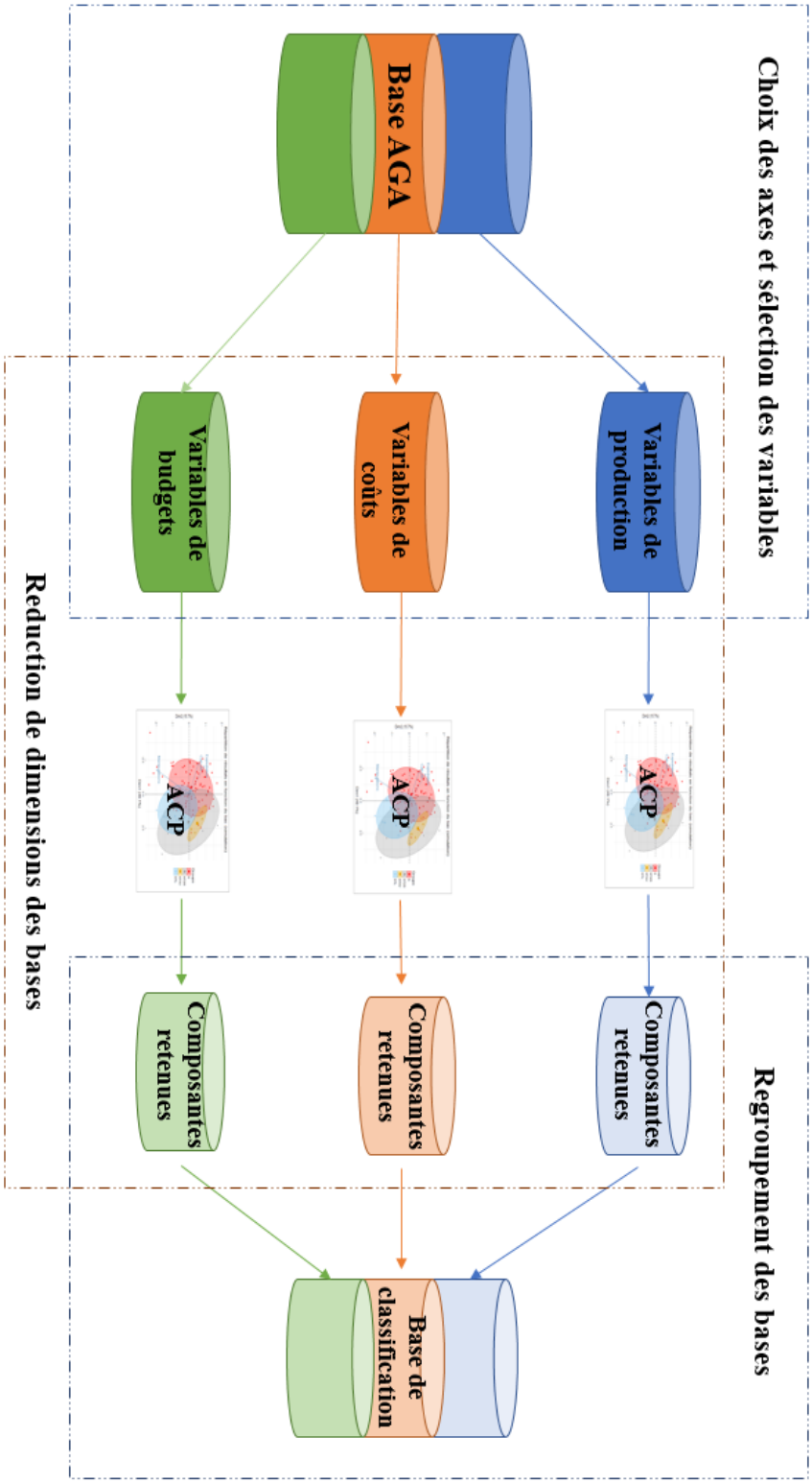

FIGURE 2.8 – Construction de la base de classification

Chapitre 3

Segmentation des distributeurs

Sommaire

3.1 Classification non supervisée **35**

3.1.1 Méthode de partitionnement : k-means 37

3.1.2 Classification des distributeurs avec k-means 38

3.2 Interprétation des groupes **40**

3.1 Classification non supervisée

La classification non supervisée est le fait de regrouper en classes des individus ou objets en se basant sur les similarités existantes entre eux. Il s'agit d'une classification sans à priori, basée sur la ressemblance de n objets suivant p critères. Elle permet de classer les objets de sorte que deux objets seront dans le même groupe s'ils se ressemblent le plus possible et dans des groupes distincts s'ils diffèrent le plus possible. Pour garantir un tel critère dans le regroupement en classes, la classification non supervisée utilise la notion d'inertie :

Soit X_1, ...,X_i,...X_n les objets classés dans K classes C_1...C_k (avec $K < n$).

En considérant X_G le centre de gravité de ce nuage de point et X_{iG} le centre de gravité de classe C_i, l'inertie totale est :

$$I = \sum_{i=1}^{n} d^2(X_i - X_G) = \sum_{i=1}^{n} \|X_i - X_G\|^2$$

Les classes étant distinctes, l'inertie totale est la somme des inerties de chaque classe :

$$I = \sum_{k=1}^{K} \sum_{i \in C_k} \|X_i - X_{kG}\|^2 \tag{3.1}$$

En s'appuyant sur le théorème de Huygens, on obtient :

$$\begin{aligned}
I &= \sum_{k=1}^{K} \sum_{i \in C_k} \|X_i - X_{kG} + X_{kG} - X_G\|^2 \\
&= \sum_{k=1}^{K} \sum_{i \in C_k} (\|X_i - X_{kG}\|^2 + \|X_{kG} - X_G\|^2) \\
&= \sum_{k=1}^{K} \sum_{i \in C_k} d^2(X_i - X_{kG}) + \sum_{k=1}^{K} \sum_{i \in C_k} d^2(X_{kG} - X_G) \\
&= I_w + I_b
\end{aligned}$$

Avec :

$$\begin{aligned}
I_w &= \sum_{k=1}^{K} \sum_{i \in C_k} d^2(X_i - X_{kG}) \\
I_b &= \sum_{k=1}^{K} \sum_{i \in C_k} d^2(X_{kG} - X_G)
\end{aligned}$$

I_w mesure l'inertie intra-classe, elle est la somme des distances entre les objets d'une classe et le barycentre de cette classe. Une classe homogène est une classe où l'inertie intra-classe est faible.

I_b mesure l'inertie inter-classe, elle est la somme des distances entre le barycentre des objets et le barycentre de chaque classe. Plus l'inertie inter-classe est grande, plus les deux classes sont hétérogènes.

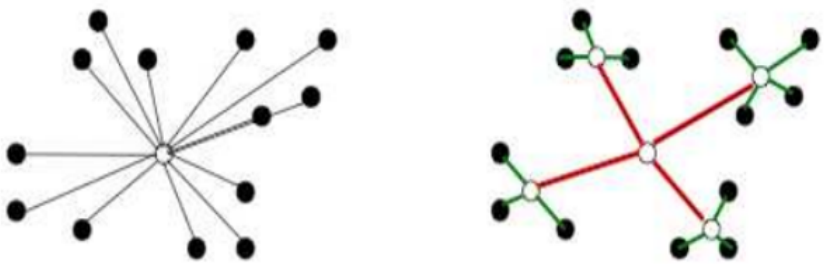

FIGURE 3.1 – Inertie intra-classe vs inertie inter-classe

La segmentation en K classes revient à trouver les classes de telles sorte que l'inertie

intra-classe soit minimale et l'inertie inter-classe soit maximale. En d'autres termes, la classification consiste à regrouper les individus de sorte à maximiser la similarité intra-groupe et à minimiser la similarité inter-groupe.

Il existe plusieurs méthodes de classification [St8] dont les deux grandes familles les plus utilisées sont : les méthodes hiérarchiques et les méthodes de partitionnement. Dans cette partie théorique, nous nous contenterons d'expliquer uniquement la méthode de partitionnement k-means. Notre choix d'utiliser cette méthode au-delà des résultats plus probants qu'elle nous donne, s'appuie sur les travaux présentés dans l'article [SAJO05] qui expose une étude comparative de plusieurs méthodes.

3.1.1 Méthode de partitionnement : k-means

Il existe plusieurs variantes de méthodes de partitions. La plus connue et la plus utilisée est celle des k-means [JHS12] :

— **Principe :**
Cette méthode se base sur un nombre k défini de partitions pour regrouper les objets en k classes. Le principe consiste à diviser arbitrairement les objets en k classes que l'on améliore itérativement jusqu'à la convergence du critère choisi (voir figure 3.2).

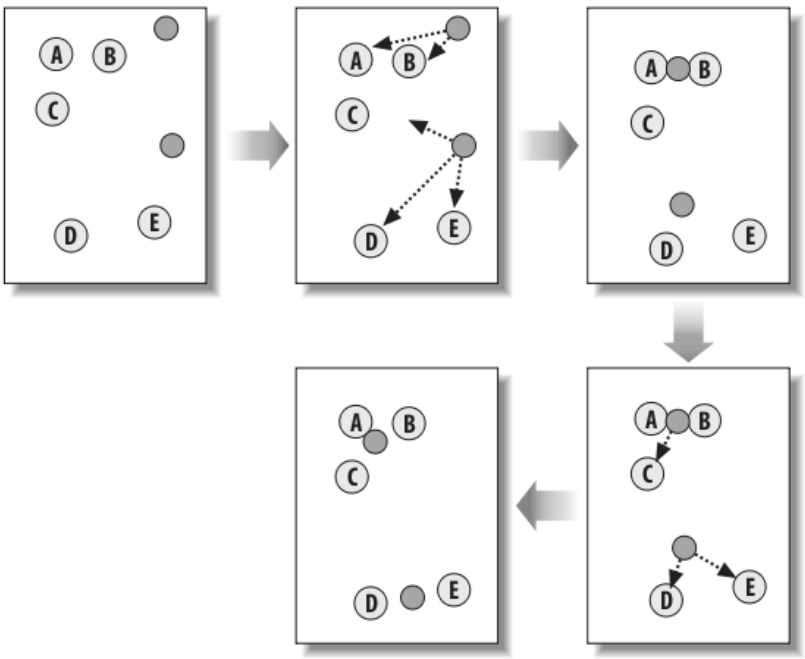

FIGURE 3.2 – Classification k-means

— **Algorithme :**
Initialisation :
On choisit aléatoirement k points comme les centres des k classes.
Itération :
Tous les objets/individus sont affectés à la classe dont le centre est le plus proche au sens de la distance choisie. On obtient ainsi les k classes.
On calcule les barycentres des classes créées qui deviennent les k nouveaux centres.
On itère jusqu'à la converge. On peut aussi fixer le nombre d'itération.

Les itérations de k-means permettent aux objets de changer de classes, ce qui n'est pas possible avec la classification ascendante hiérarchique pour laquelle une affectation est irréversible. Le k-means permet également d'explorer plusieurs solutions en changent les points de départ et les répétitions.

L'inconvénient de cette méthode est qu'elle ne permet pas de savoir quel est le nombre de classes convenables, ni de visualiser la proximité entre les classes ou les objets. Mais il existe différentes techniques qui permettent de choisir le nombre de groupe optimal par rapport à ces données.

3.1.2 Classification des distributeurs avec k-means

Après avoir expliqué le principe du k-means, nous allons l'utiliser pour regrouper nos distributeurs en classes homogènes. Il faut donc dans un premier temps, identifier le nombre de classes optimales avant d'appliquer la méthode.

3.1.2.1 Choix du nombre de classes optimales : Average silhouette

Il existe diverses approches qui permettent de déterminer le nombre de clusters optimal. Nous avons utilisé la méthode "Average silhouette" pour choisir le nombre de clusters.

Il s'agit d'une approche qui permet de mesurer la qualité d'une classification en déterminant le niveau d'appartenance de chaque objet à sa classe. La méthode calcule pour chaque nombre de clusters donné la silhouette moyenne des objets. Une silhouette moyenne élevée montre une bonne qualité de classification pour le nombre de clusters choisi. Le nombre de clusters optimal est alors la valeur pour laquelle la silhouette moyenne est maximale.

La figure 3.3 présente la silhouette lorsqu'on applique la méthode à notre base d'étude. Les résultats nous montrent que le nombre optimal de clusters est 2 puis 3 ou 5. Mais une classification en 2 ou 3 groupes n'est pas très pertinente. Nous allons donc choisir une

division en cinq groupes car au-delà de 3 groupes, le nombre qui maximise la qualité de la classification est 5.

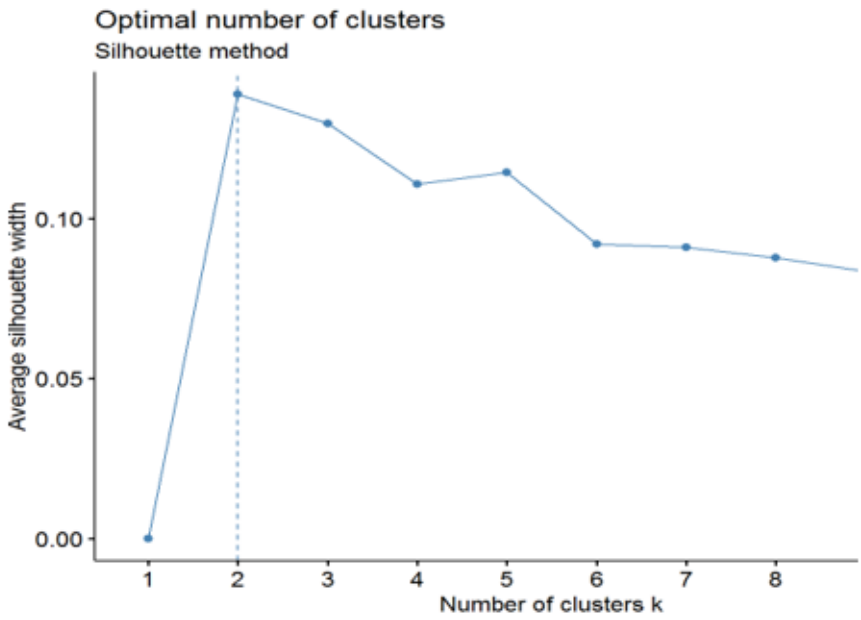

FIGURE 3.3 – Nombre de clusters optimal par la méthode de silhouette

3.1.2.2 Implémentation

Nous appliquons la methode k-means sous R en utilisant la fonction *kmeans*. Nous choisissons de découper en 5 groupes avec un nombre d'itération qui garantit la convergence. On obtient le découpage présenté sur le tableau suivant.

Groupe	Nombre d'individus	Pourcentage
1	232	7.4%
2	601	19.2%
3	1 115	35.6%
4	603	19.2%
5	581	18.6%

TABLE 3.1 – Groupes k-means

On constate également qu'avec 5 classes, nous avons 60% d'inertie expliquée (voir 3.4. Ce qui représente un pourcentage acceptable.

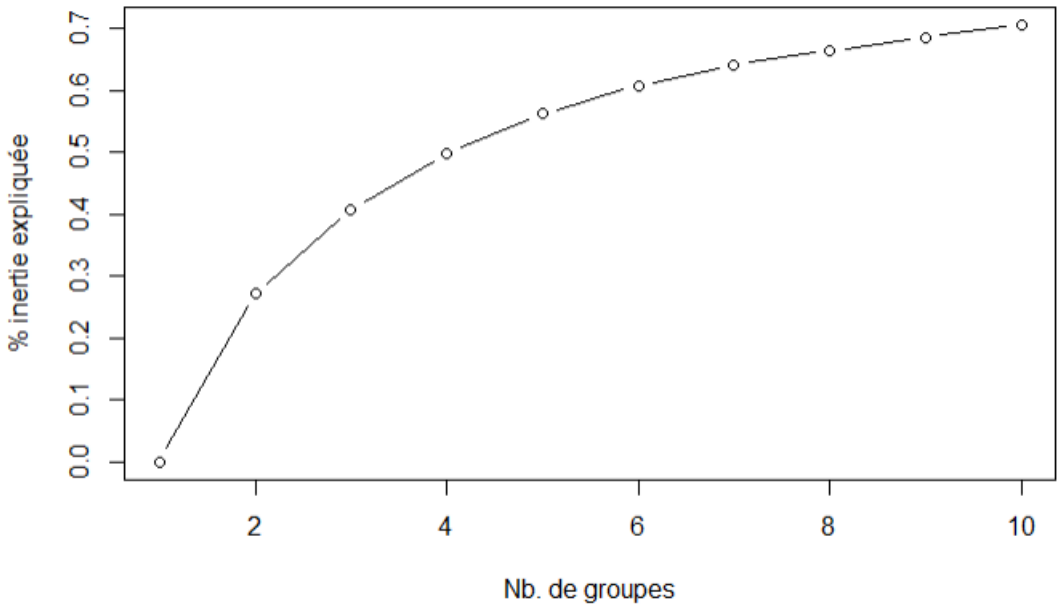

FIGURE 3.4 – Proportion d'inertie expliquée

3.2 Interprétation des groupes

L'analyse des groupes peut se faire en utilisant les barycentres des groupes mais nous allons utiliser le test de Kruskal-Wallis qu'on va compléter avec quelques statistiques descriptives pour caractériser nos groupes.

Le test de Kruskal-Wallis est un test non paramétrique que nous pouvons utiliser pour mesurer l'intensité de la liaison entre une variable qualitative et des variables quantitatives. Nous allons nous inspirer d'une application de ce test proposé par *Stéphane TUFFERY* [St8] pour interpréter les groupes que nous avons obtenus.

Nous allons, à l'aide du test Kruskal-Wallis, mesurer l'intensité puis donner avec le sens de cette liaison (positif ou négatif) grâce aux rangs de Wilcoxon. Nous allons donc calculer les liaisons pour chaque classe en utilisant une fonction indicatrice qui va transformer nos groupes en cinq variables binaires (pour chaque individu, la variable binaire correspondant à son groupe prend la valeur 1 et les quatre autres variables prennent la valeur 0). On va calculer l'intensité entre chacune de ces cinq variables et nos variables quantitatives initiales. Nous présenterons uniquement les résultats qui nous permettent d'interpréter nos groupes.

Groupe 1 : (7.4% des agents)

En se basant sur les résultats du test Kruskal-Wallis, on constate que les agents du groupe 1 ont beaucoup de réserve en budget ST(NbMoisConsoST) et ARCC (NbMoisConsoArcc). Ils n'ont donc pas de besoin en dotation. De plus, ils utilisent la dotation qu'ils reçoivent (UseARCC : proportion de dotation consommée). Ils ont donc constitué un stock de budget en réserve. Leur taux de rabais (TxRabaisAuto et TxRabaisDomP) est faible. Ils ont une sous-tarification moyenne en auto (SousTarifAuto) et une part de contrats avec réductions en auto (PartCntRedAuto) faible. Nous pouvons donc dire qu'ils font très peu de réductions. Enfin, ils ont une taille de portefeuille (PTF_IARD) faible et privilégient les mono-détenteurs lorsqu'ils utilisent du budget en auto (MonoST).

Il s'agit donc d'agents qui ne sont pas spécialisés en IARD et qui utilisent très peu les leviers à leur disposition. Nous pouvons qualifier ce groupe de **PETITS CONSOMMATEURS** en référence à leur impact sur le résultat IARD et leur faible utilisation des budgets commerciaux.

Obs	Variables	KWallis	signe
1	NbMoisConsoArcc	389	plus
2	NbMoisConsoST	372	plus
3	UseARCC	184	plus
4	TxRabaisDomP	139	moins
5	SousTarifAuto	129	moins
6	PartCntRedAuto	123	moins
7	TxRabaisAuto	117	moins
8	UseST	56	plus
9	PTF_IARD	50	moins
10	MonoST	40	plus
11	CA_IARD	35	moins
12	TxCROIS_AutoP	32	moins
13	MonoARCC	30	plus
14	TxCROIS_DomP	30	moins
15	PartCntRedDomP	26	moins
16	TermeST	26	moins
17	CliSinARCC	24	moins
18	TxTransf_Cli	23	moins
19	SousTarifDomP	21	moins
20	RC_TOT	20	moins

Table 3.2 – Test de Kruskal-Wallis Groupe 1

Groupe 2 : (19.2% des agents)

Les agents de ce groupe sont caractérisés par la taille (PTF_IARD) de leur portefeuille et leur chiffre d'affaires (CA_IARD) qui sont largement au-dessus de la moyenne nationale des agents. Leurs taux de rabais (TxRabaisAuto et TxRabaisDomP) sont élevés ainsi que leur sous-tarification moyenne (SousTarifAuto et SousTarifDomP). Ils ont un grand nombre de contrats avec réduction en portefeuille (PartCntRedAuto et PartCntRedDomP). Ils ont très peu de budget en réserve surtout en ST (NbMoisConsoST) et ne sont pas tournés vers les nouveaux produits (PartMaMaison).

Il s'agit donc d'un groupe qui est l'opposé du premier groupe en termes d'utilisation des budgets commerciaux et de taille. Les agents de ce groupe sont ceux qui portent l'activité IARD particuliers professionnels. Ils impactent également l'activité à travers leur consommation élevée en budget. Leur forte utilisation de budget augmente le risque de mauvaise utilisation du budget à leur niveau. Nous pouvons les qualifier de **GROS CONSOMMATEURS** en référence à leur taille et leur consommation.

Obs	Variables	KWallis	signe
1	CA_IARD	589	plus
2	PTF_IARD	493	plus
3	SousTarifAuto	348	plus
4	TxRabaisDomP	346	plus
5	PartCntRedAuto	333	plus
6	TxRabaisAuto	228	plus
7	SousTarifDomP	202	plus
8	PartCntRedDomP	138	plus
9	NbMoisConsoST	106	moins
10	PartMaMAISON	97	moins
11	RC_AUTO	68	plus
12	TxCROIS_DomP	68	plus
13	MonoST	52	moins
14	CTP_BASE	43	plus
15	RC_MRH	43	plus
16	RC_TOT	38	plus
17	TxRESIL_DomP	37	moins
18	NbMoisConsoArcc	37	moins
19	MonoARCC	32	moins
20	CliSinARCC	22	plus

Table 3.3 – Test de Kruskal-Wallis Groupe 2

Groupe 3 : (35.6% des agents)

Les agents de ce groupe sont caractérisés par de faibles taux de résiliation (TxRESIL_AutoP et TxResil_DomP) et ont des ratios combinés faibles (RC_TOT, RC_AUTO et RC_MRH). Ils ont donc de bons résultats techniques et font face à peu de résiliations. Ils font très peu de devis (TxDevis) mais transforment bien les devis clients (TxTransf_Cli) et prospects (TxTransf_Pros). Ils ont une sous-tarification moyenne faible en auto (SousTarifAuto) et n'ont pas beaucoup de contrats avec réductions en portefeuille. Ce qui montre qu'ils maîtrisent les coûts et ne font pas un usage abusif des budgets. Leur taux de production est faible et ils font peu de Clic and Go (Part de Clic and Go). Il s'agit d'un groupe qui se concentre sur son portefeuille et qui le développe très peu. Ce qui leur garantit de bons résultats techniques et par conséquent, ils bénéficient de bons tarifs (à travers leur CTP qui est faible). Nous pouvons les qualifier d'**AGENTS A STIMULER**.

Obs	Variables	KWallis	signe
1	TxRESIL_AutoP	565	moins
2	RC_TOT	563	moins
3	RC_AUTO	514	moins
4	PartClicGo	491	moins
5	RC_MRH	420	moins
6	CTP_BASE	416	moins
7	TxDevis	404	moins
8	SousTarifAuto	347	moins
9	TxRESIL_DomP	314	moins
10	TxTransf_Cli	296	plus
11	PartCntRedAuto	243	moins
12	TxPROD_AutoP	161	moins
13	CA_IARD	133	moins
14	TxTransf_Pros	130	plus
15	TxPROD_DomP	114	moins
16	TxRabaisDomP	109	moins
17	PartMaMAISON	87	moins
18	PTF_IARD	74	moins
19	TxRabaisAuto	67	moins
20	PartCntRedDomP	40	plus

Table 3.4 – Test de Kruskal-Wallis Groupe 3

Groupe 4 : (19.2% des agents)

Ce groupe est constitué d'agents en perte de croissance en auto (TxCROIS_AutoP) et en dommages (TxCROIS_DomP). Cela s'explique par leur faible taux de production et leur fort taux de résiliation en auto et en dommages.

Ils ont des ratios combinés élevés (RC_TOT et RC_AUTO). Ce qui montre qu'ils n'ont pas de bons résultats techniques.

Ils mettent beaucoup de budgets sur le Terme (TermeST et TermeARCC). Il s'agit donc d'un groupe d'agents qui protège son terme mais qui est en perte de production avec des résultats techniques mauvais.

Ils ont du mal à transformer les devis sur leurs clients (TxTransf_Cli). Nous pouvons les qualifier d'**AGENTS EN DIFFICULTÉ** car leurs résultats sont insuffisants.

Obs	Variables	KWallis	signe
1	TxCROIS_AutoP	546	moins
2	TxCROIS_DomP	503	moins
3	TxPROD_DomP	212	moins
4	TermeARCC	182	plus
5	VarAN_AutoP	178	moins
6	TxPROD_AutoP	158	moins
7	VarAN_DomP	157	moins
8	TxRESIL_DomP	143	plus
9	TxRESIL_AutoP	141	plus
10	CTP_BASE	124	plus
11	TxTransf_Cli	119	moins
12	RC_TOT	97	plus
13	TermeST	86	plus
14	RC_AUTO	65	plus
15	PartClicGo	63	plus
16	RC_MRH	54	plus
17	PartMaMAISON	34	plus
18	SousTarifAuto	30	plus
19	PTF_IARD	26	moins
20	TxTransf_Pros	25	moins

Table 3.5 – Test de Kruskal-Wallis Groupe 4

Groupe 5 : (18.6% des agents)

Dans ce groupe, les agents sont caractérisés par une forte production auto et dommages avec un fort taux de devis (TxDevis). Ce sont des agents actifs qui génèrent beaucoup de flux.

Malgré leur fort taux de résiliation, ils ont un apport net positif (TxCROIS_AutoP et TxCROIS_DomP). Ils sont donc dans une dynamique positive de production.

Ils sont plus tournés vers le nouveau produit auto (PartClicGo et PartMaMAISON). Ces agents sont caractérisés par leur très mauvais résultats aussi bien par produit que sur tout l'IARD (RC_TOT) mais également sur les produits auto et dommages.

Nous pouvons qualifier ces agents de **PRODUCTEURS**. Ils ont une stratégie orientée sur la production et le développement au détriment de la rentabilité.

Obs	Variables	KWallis	signe
1	TxPROD_AutoP	762	plus
2	TxDevis	708	plus
3	TxPROD_DomP	639	plus
4	TxRESIL_AutoP	327	plus
5	PartClicGo	292	plus
6	TxCROIS_DomP	273	plus
7	RC_TOT	255	plus
8	TxCROIS_AutoP	250	plus
9	VarAN_AutoP	243	plus
10	VarAN_DomP	207	plus
11	RC_AUTO	203	plus
12	TxRESIL_DomP	182	plus
13	PartMaMAISON	180	plus
14	RC_MRH	147	plus
15	PartCntRedDomP	137	moins
16	CTP_BASE	91	plus
17	TermeST	49	moins
18	TxTransf_Cli	42	moins
19	VarRes_AutoP	39	plus
20	TxRabaisDomP	38	plus

Table 3.6 – Test de Kruskal-Wallis Groupe 5

Répartition par groupe de quelques indicateurs

La répartition des indicateurs ci-dessous vient confirmer l'analyse des groupes que nous venons de faire sur la base du test de Kruskal-Wallis :

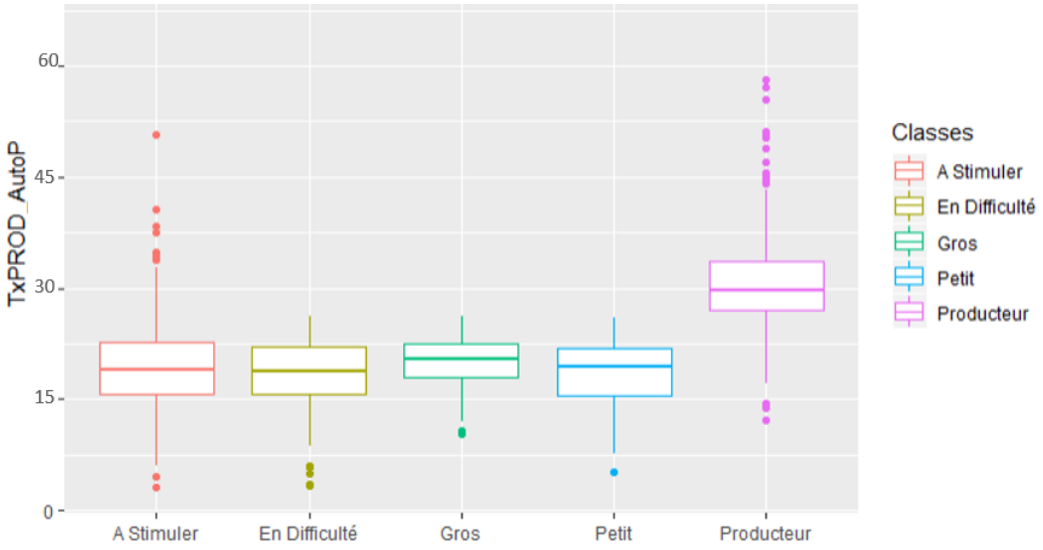

FIGURE 3.5 – Boxplot des taux de production auto par groupe

La figure 3.5 montre la répartition du taux de production auto par classe. Nous remarquons que les **Producteurs** se détachent largement des autres groupes. Ils sont donc très orientés vers la production.

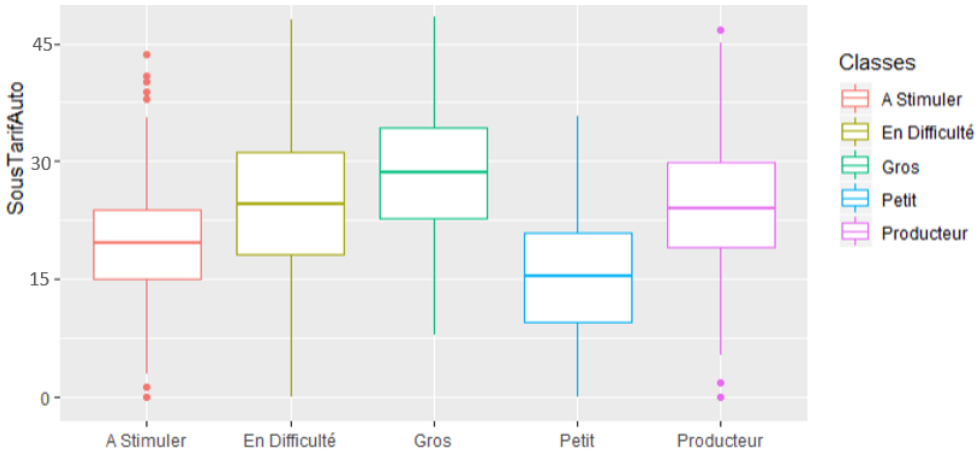

FIGURE 3.6 – Boxplot de la sous-tarification moyenne par groupe

La figure 3.6 montre la répartition de la sous-tarification moyenne auto par classe. Nous

pouvons remarquer que les **Petits Consommateurs** sont évidemment ceux qui ont une moyenne de sous-tarification la plus faible suivis des **Agents à stimuler**.

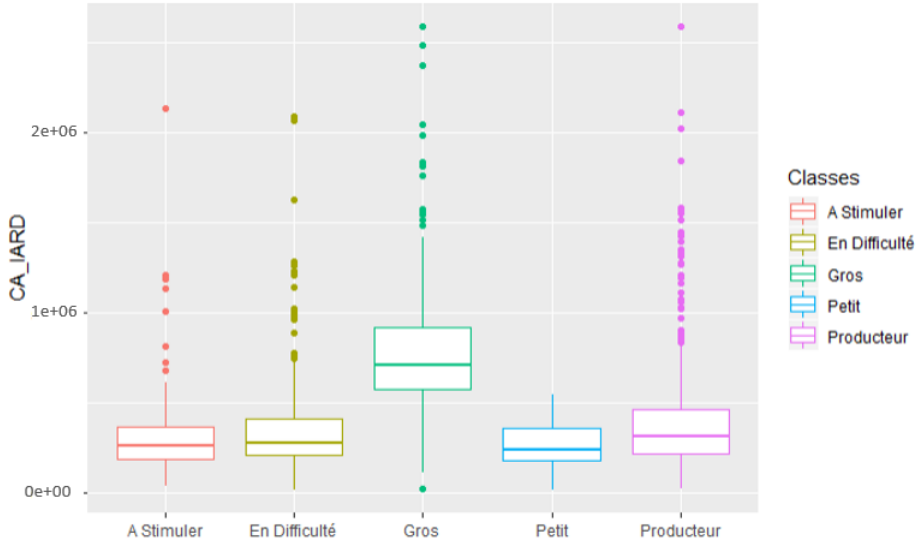

FIGURE 3.7 – Boxplot des CA par groupe

La figure 3.7 montre la répartition du chiffre d'affaires par classe. Nous remarquons que les **Gros consommateurs** se caractérisent effectivement par leur chiffre d'affaires qui est largement supérieur à la moyenne des distributeurs.

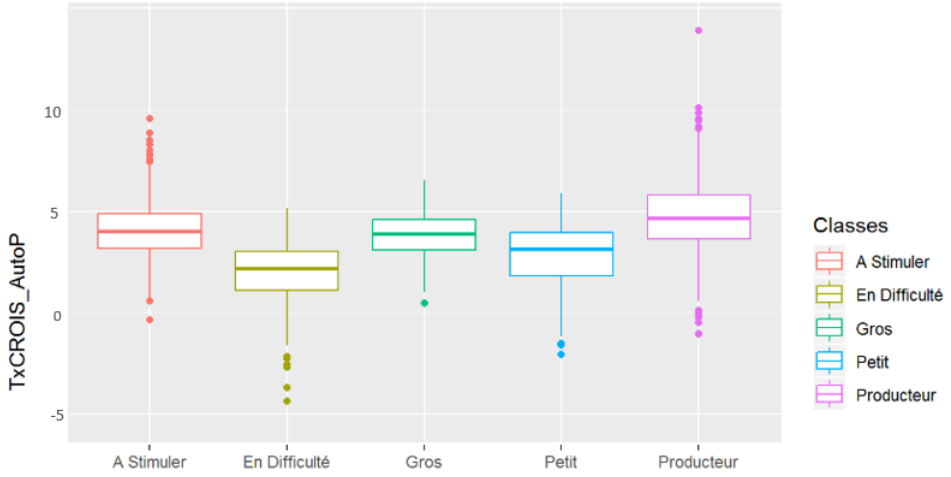

FIGURE 3.8 – Boxplot du croissance auto par groupe

La figure 3.8 montre la répartition du taux de croissance auto par classe. Nous pouvons confirmer que les **agents en difficulté** sont effectivement en perte de croissance par rapport aux autres groupes.

Dans ce chapitre, nous avons regroupé les distributeurs en groupes homogènes que nous avons interprétés à l'aide du test de Kruskal-Wallis et de quelques données statistiques sur les groupes. Mais cette classification ne garantit pas le bon classement de tous les agents. De plus, il n'est pas évident de rendre opérationnelle cette segmentation. Il nous faut donc mieux caractériser ces classes d'agents afin de disposer de critères plus opérationnels. Dans le chapitre suivant, nous allons aborder l'opérationnalisation de notre segmentation.

Chapitre 4

Opérationnalisation de la segmentation

Sommaire

4.1 Classification supervisée 49
4.1.1 Modèles linéaires généralisés : GLM 50
4.1.2 Les arbres de décisions 57
4.1.3 Conclusion sur la caractérisation des groupes 63
4.2 Généralisation de la segmentation 65
4.2.1 Généralisation non contextuelle 66
4.2.2 Généralisation contextuelle 67
4.3 A/B Testing avec une dotation ST aux groupes 68
4.4 Mesures opérationnelles 70

Pour rendre opérationnelle la segmentation, nous allons utiliser une classification supervisée pour mieux caractériser les groupes. Cette approche permettra également de généraliser la segmentation et de mesurer sa stabilité. Enfin, nous utiliserons cette caractérisation pour proposer des mesures d'accompagnement adaptées à chaque groupe.

4.1 Classification supervisée

Les groupes obtenus par la segmentation peuvent être considérés comme une variable de réponse (variable qualitative à cinq modalités). On peut donc utiliser les indicateurs présentés au chapitre 2 comme variables explicatives pour faire un apprentissage supervisé. Cet apprentissage permettra de faire une caractérisation plus simple des groupes avec moins d'indicateurs.

Nous avons fait l'option d'utiliser toute notre base de données comme base d'apprentissage car nous estimons qu'il s'agit d'une petite base de données et que toute réduction

pourrait impacter la qualité de notre apprentissage et donc notre caractérisation. Pour faire un tel apprentissage, il existe plusieurs méthodes de classification supervisée mais nous allons nous limiter aux modèles linéaires généralisés (GLM) et aux arbres de décisions pour une meilleure interprétation des résultats.

4.1.1 Modèles linéaires généralisés : GLM

4.1.1.1 Généralités :

Initiés en 1972 par Nelder et Wedderburn, puis présentés de façon complète par Mc Cullagh et Nelder en 1989, les modèles linéaires généralisés (Generalized Linear Model, GLM) sont utilisés pour modéliser une relation entre une variable à expliquer ou de réponse et des variables explicatives. Ils constituent une généralisation du modèle linéaire simple.

Du modèle linaire simple aux GLM

Un modèle linéaire simple ou régression linéaire est un modèle dont le but est d'établir une relation linéaire entre une variable à expliquer et les variables explicatives. Cela revient à exprimer la variable à expliquer sous la forme :

$$Y_i = \sum_{j=1}^{p} \alpha_j X_{ij} + \alpha_0 + \varepsilon_i \tag{4.1}$$

Avec :
- $X_{i1} \ldots\ldots X_{ip}$ les variables explicatives
- Y_i la variable à expliquer
- $\varepsilon_i \sim \mathcal{N}(0,\, \sigma^2)$ est l'écart entre le modèle et les observations : on l'appelle résidu.
- Les variables Y_i et ε_i sont indépendantes

L'objectif du modèle est donc d'établir une relation linaire entre espérance conditionnelle de la variable à expliquer et les variables explicatives. De plus, la variable à expliquer doit suivre une loi normale. Cependant, il n'est pas toujours évident, voir difficile de respecter ces conditions. Le modèle linéaire généralisé permet d'établir une relation entre ces variables sans respecter ces conditions. On peut donc s'affranchir du caractère normal de la variable à expliquer et la relation entre les variables et l'espérance conditionnelle n'est pas forcément linéaire.

Le but d'un GLM (voir annexe C pour comprendre le principe et les hypothèses du GLM) est d'établir une relation entre les variables explicatives et l'espérance de la variable à expliquer telle que :

$$g(E[Y_i]) = \eta(X_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} \tag{4.2}$$

g est appelé la fonction de lien.

Régression logistique multinomiale

Dans le cadre de notre étude, notre variable à expliquer est qualitative et comprend cinq modalités. Il s'agit donc d'une distribution multinomiale qui est la généralisation du cas binomial. En effet, si nous considérons que la variable à expliquer Y_i peut prendre les valeurs 1,2,...K avec $K \geq 2$, on note la probabilité que la i ème valeur soit de modalité k :

$$\pi_{ik} = Prob(Y_i = k)$$

Avec la contrainte :

$$\sum_{k=1}^{K} Prob(Y_i = k) = 1 \tag{4.3}$$

On a donc un modèle à K-1 vecteurs de paramètres. Dans notre cas, il suffit de connaître la probabilité d'appartenance aux quatre premiers groupes pour déduire celle du cinquième groupe.

En considérant n_i le nombre d'observations dans le groupe i et Y_{ik} le nombre d'observations dont le groupe est passé de i à k. On a alors une loi de distribution de Y qui peut se mettre sous la forme suivante :

$$Prob(Y_{ik} = y_{ik}, ..., Y_{iK} = y_{iK}) = \binom{n_i}{y_{i1}, ..., y_{iK}} \pi_{i1}^{y_{i1}} .... \pi_{iK}^{y_{iK}} \tag{4.4}$$

L'approche pour les variables multinomiales consiste à choisir une modalité de la variable à expliquer puis à modéliser les logs-odds pour les autres modalités :

$$\eta_{ik} = log \frac{\pi_{ik}}{\pi_{iK}} = \alpha_k + X_i' \beta_k \tag{4.5}$$

Dans le cadre de notre étude, nous allons choisir le groupe des **Producteurs** comme référence et modéliser les autres groupes d'agents à partir de ce groupe. Le choix de cette référence s'explique par le fait qu'il s'agit du groupe référence en termes de performance de production. Après avoir choisi la référence, il est important d'utiliser les critères de qualité pour sélectionner les variables ainsi que le modèle. Plusieurs méthodes sont utilisées pour effectuer ces choix.

4.1.1.2 Sélection des variables et choix du modèle

Le choix des variables revient à déterminer un ensemble de variables explicatives pour lesquelles le modèle a un pouvoir explicatif assez suffisant. C'est-à-dire, un modèle qui réalise le meilleur compromis entre la robustesse et le pouvoir explicatif.

En effet, si l'augmentation du nombre de variables explicatives d'un modèle permet d'avoir plus d'informations sur la variable à expliquer, elle ne garantit pas nécessairement la robustesse du modèle. Ainsi le modèle comportant toutes les variables explicatives est dit modèle saturé car il permet un ajustement parfait.

Notre problème revient donc à comparer les modèles successivement obtenus au modèle saturé. Cette comparaison est effectuée au sens des log-vraisemblances $L_{saturé}$ (pour le modèle saturé) et L (pour le modèle réalisé) et on calcule alors la déviance D :

$$D = 2(L_{saturé} - L)$$

Lorsque la déviance est proche de 0, le modèle réalisé est adéquat et à l'inverse, lorsque la déviance est importante, le modèle n'est pas adéquat.

De plus, le modèle obtenu par chaque combinaison de variables peut être sélectionné à l'aide de différents critères. Les principaux critères utilisés sont AIC (Akaike Information Criterion) et BIC (Bayesian Information Criterion). Il s'agit de critères de type log-vraisemblance pénalisée qui se définissent comme suit :

$$\begin{cases} AIC = & -2l + 2K \\ BIC = & -2l + Klog(n) \end{cases}$$

Avec
- l est la log-vraisemblance du modèle,
- K le nombre de paramètres indépendants dans le modèle
- n le nombre d'observations de l'échantillon.

Le meilleur modèle est celui qui minimise le plus ces deux critères parmi tous les modèles générés. En utilisant ces deux critères, on choisit les variables à l'aide de méthodes telles que :
- **Sélection ascendante** : à partir du modèle constitué uniquement avec la constante, on ajoute à chaque itération la variable qui permet d'avoir la plus forte baisse de l'AIC ou du BIC ; on s'arrête lorsque l'indicateur ne baisse plus.
- **Sélection descendante** : à partir du modèle avec toutes les variables, on enlève à chaque itération, la variable qui minimise le moins l'AIC ou le BIC. On s'arrête lorsqu'on atteint le minimum.

— **Stepwise** : c'est un mélange des deux approches précédentes. On effectue donc une sélection ascendante avec la possibilité de suppression d'une variable déjà sélectionnée.

Dans le cadre de notre étude, nous avons utilisé la méthode stepwise en minimisant l'AIC. La figure ci-dessous montre l'évolution de l'AIC lors de la sélection des variables.

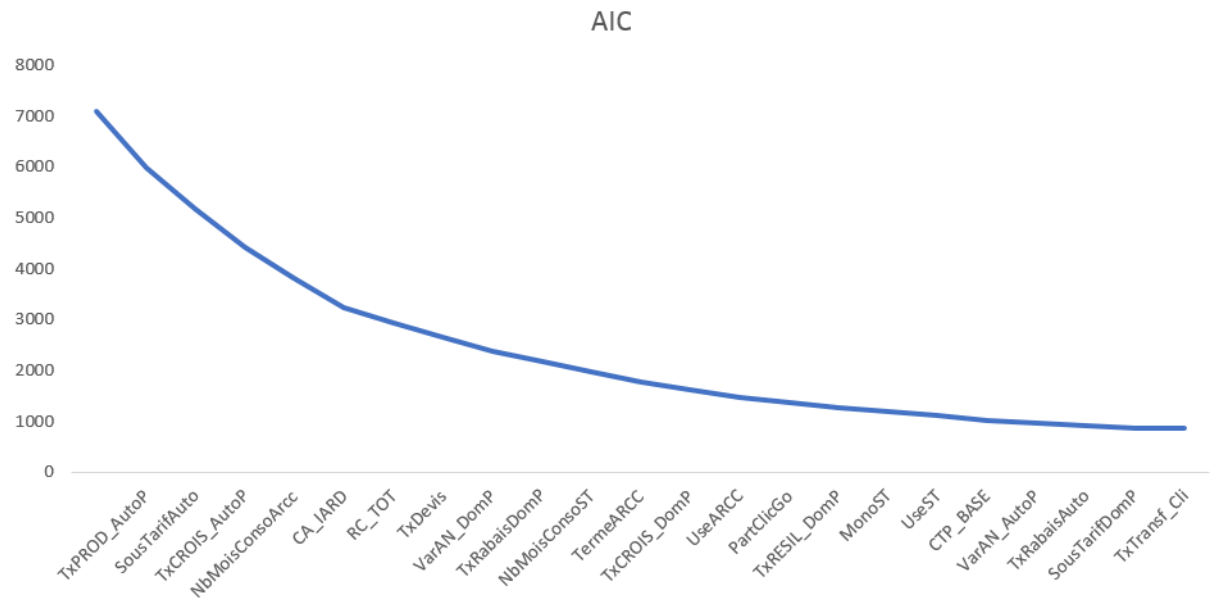

FIGURE 4.1 – Sélection des variables à l'aide de l'AIC

4.1.1.3 Validation et pouvoir explicatif

Comme nous l'avons précisé au début de ce chapitre, nous utilisons ces méthodes pour mieux caractériser nos groupes et identifier des stratégies visant à améliorer la performance des agents. Pour vérifier la fiabilité du modèle et son pouvoir explicatif pour chaque groupe, nous avons fait une validation croisée [NIE16].

La validation croisée est une méthode de validation basée sur un échantillonnage qui consiste à diviser la base en k échantillons disjoints B_1, ...,B_k de même taille. On construit alors k modèles tels que pour i $\in [1, ..., k]$,B\B_i représente l'échantillon d'apprentissage et B_i l'échantillon de test. On calcule alors la moyenne des k erreurs quadratiques moyennes qui nous permet d'estimer l'erreur de prédiction.

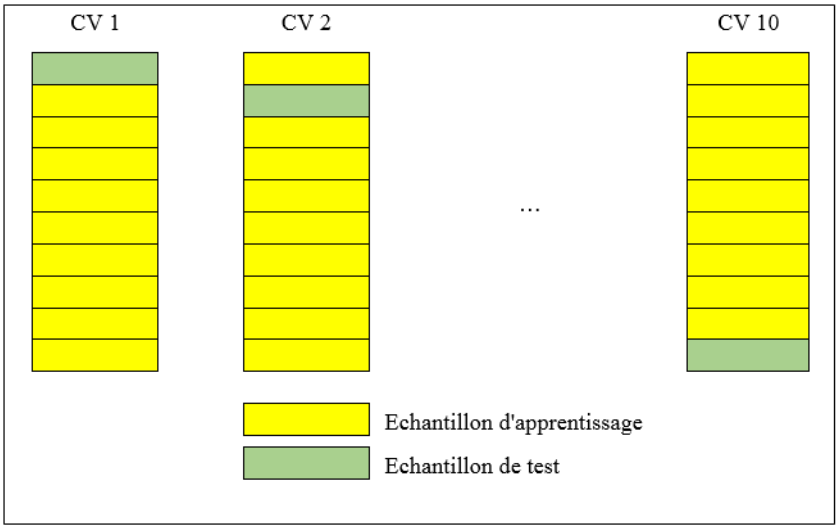

FIGURE 4.2 – Exemple de validation croisée

Pour notre étude, nous avons fait une validation croisée en divisant en 10 la base d'apprentissage. A partir de nos 10 modèles, on obtient une performance moyenne de 93.5%. La matrice de confusion ci-dessous présente les détails du classement des agents. On note donc un taux d'erreur très faible, ce qui suppose que le modèle permet donc une bonne explication de la segmentation. Il peut donc être utilisé pour caractériser nos groupes.

	Groupes k-means					
Groupes GLM	**Agents en difficulté**	**Gros consommateurs**	**Petits consommateurs**	**Producteurs**	**Agents à stimuler**	**Total**
Agents en difficulté	18,3%	0,3%	0,1%	0,1%	0,5%	**19,3%**
Gros consommateurs	0,2%	17,8%	0,0%	0,3%	0,6%	**18,9%**
Petits consommateurs	0,1%	0,0%	7,0%	0,0%	0,2%	**7,4%**
Producteurs	0,2%	0,4%	0,1%	17,9%	0,1%	**18,7%**
Agents à stimuler	0,5%	0,7%	0,2%	0,3%	34,1%	**35,8%**
Total	**19,2%**	**19,2%**	**7,4%**	**18,6%**	**35,6%**	**100,0%**

FIGURE 4.3 – Matrice de classification du GLM

4.1.1.4 Caractérisation des groupes

L'application de la procédure stepwise en utilisant la fonction *multinom* sous R nous a permis d'éliminer certaines variables et de ne retenir que les variables significatives. On peut alors utiliser les coefficients obtenus (voir le tableau 4.4) pour décrire l'impact de chaque variable significative pour chaque groupe. Pour faciliter la caractérisation, on a utilisé également le risque relatif (relative risk ratio) [Osc19] qui représente l'exponentielle du coefficient de la logit. Nous proposons de caractériser les **Petits consommateurs** par rapport au groupe référence qui est le groupe des **Producteurs**.

	A Stimuler	En Difficulté	Gros	Petit
X.Intercept.	101.2230	29.7869	50.8302	45.7671
TxPROD_AutoP	-0.8540	-0.6785	-0.6126	-0.7361
SousTarifAuto	-0.1824	-0.0044	0.1656	-0.2905
TxCROIS_AutoP	0.3164	-0.3106	0.0500	0.0641
NbMoisConsoArcc	-0.0414	-0.0753	-0.0948	0.5080
RC_TOT	-0.5231	-0.1316	-0.2139	-0.3838
TxDevis	-0.4561	-0.2020	-0.2311	-0.0302
VarAN_DomP	-0.0475	-0.0715	-0.0625	-0.0683
TxRabaisDomP	-0.2830	-0.1036	0.6384	-0.6504
NbMoisConsoST	0.0431	-0.0405	-0.0890	0.4567
TermeARCC	0.0128	0.1000	0.0203	-0.0108
TxCROIS_DomP	0.2461	-0.6092	0.2161	-0.5381
UseARCC	7.2468	5.6616	17.9334	47.1256
PartMaMAISON	-0.0615	-0.0332	-0.0991	-0.0502
PartClicGo	-0.1023	-0.0397	-0.0577	-0.0235
TxPROD_DomP	-0.8371	-0.6095	-0.7707	-0.3234
MonoARCC	-0.0007	0.0157	-0.0446	0.2194
UseST	0.8212	0.5612	-0.1093	6.3720
VarAN_AutoP	-0.0403	-0.0884	-0.0562	-0.0566
CTP_BASE	-0.3119	0.0356	-0.1664	-0.2348
TxTransf_Cli	0.0956	-0.0544	0.0481	0.0545
PointsCTP	0.1451	0.2788	-0.3847	0.1904
TxRabaisAuto	-0.2791	-0.1727	0.0616	-0.7832

FIGURE 4.4 – Coefficients du GLM

Caractérisation des Petits Consommateurs :

L'analyse des coefficients montre que les **Petits consommateurs** ont plus de chance de devenir des agents producteurs s'ils améliorent leur sous-tarification moyenne. En effet, lorsque nous supposons que les autres variables significatives sont constantes et que nous augmentons la sous-tarification auto de ces agents, la probabilité qu'ils changent de groupe pour devenir des **Producteurs** est plus élevée que celle de rester dans leur groupe.

Par contre, s'ils continuent à augmenter leur réserve en budget, leur chance de rester dans le groupe des **Petits consommateurs** est plus élevée que celle de devenir **Producteurs**. Nous pouvons noter également l'impact de l'utilisation du budget sur l'appartenance à ce groupe. En effet, une augmentation des variables UseST et UseARCC (part de budget consommée par rapport à la dotation reçue) impliquerait que ces agents n'auraient aucune chance de devenir **Producteurs**. De même, d'autres variables telles que les taux de rabais auto et dommages sont aussi importants pour ce groupe. Une forte utilisation des rabais augmenterait la probabilité de changement de groupe pour ces agents.

Nous pouvons donc conclure que les **Petits consommateurs** doivent changer leur utilisation des budgets commerciaux et augmenter leur niveau de sous-tarification pour booster leur production.

Nous pouvons caractériser par la même approche les **Gros consommateurs**, les **Agents à stimuler** et les **Agents en difficulté** par rapport aux **Producteurs**.

4.1.1.5 Conclusion

La méthode GLM nous permet de sélectionner un certain nombre d'indicateurs pertinents mais aussi d'estimer l'impact de chaque indicateur sur l'appartenance à un groupe par rapport à notre groupe référence. Nous disposons ainsi de plus d'éléments pour caractériser nos groupes et pour identifier des leviers d'amélioration adaptés à chaque groupe.

Cependant, nous estimons que le nombre d'indicateurs utilisés par le modèle reste important et ne permet pas d'avoir des critères simples pour la caractérisation de chaque groupe. De plus, la caractérisation par le modèle est basée sur les probabilités d'appartenance. Les arbres de décisions peuvent constituer une alternative avec laquelle nous pourrons avoir une caractérisation plus précise.

4.1.2 Les arbres de décisions

Les arbres de décisions constituent une méthode d'apprentissage supervisé basée sur une modélisation non linéaire. Ils permettent de définir un ensemble de règles de classification organisé en arborescence et compréhensible pour l'utilisateur. On distingue donc deux types d'arbres :

- **arbres de classification** : la variable cible est qualitative et la prédiction sera une étiquette de classe.
- **arbres de régression** : la variable cible est quantitative et la prédiction sera une valeur numérique

Les arbres de décision permettent d'établir une relation très explicite entre la variable de réponse et les variables explicatives. Ils sélectionnent automatiquement les variables pertinentes et permettent une traduction directe de l'arbre vers des règles de classification. De plus, il s'agit d'une méthode qui est robuste face aux données aberrantes et aux données manquantes.

Il existe plusieurs algorithmes qui construisent différemment l'arbre. Dans le cadre de notre étude, nous avons utilisé l'algorithme CART (Classification And Regression Tree) qui ne nécessite pas de paramétrage.

4.1.2.1 Algorithme CART

Proposé en 1984 par Breiman, Friedman, Olshen et Stone, l'algorithme CART peut être implémenté sous R avec la bibliothèque *rpart*. Il utilise des tests binaires pour construire l'arbre. En effet, l'algorithme CART partitionne à chaque nœud l'ensemble associé au nœud en deux sous ensemble à travers un test auquel la réponse est oui ou non comme le montre l'exemple 4.5.

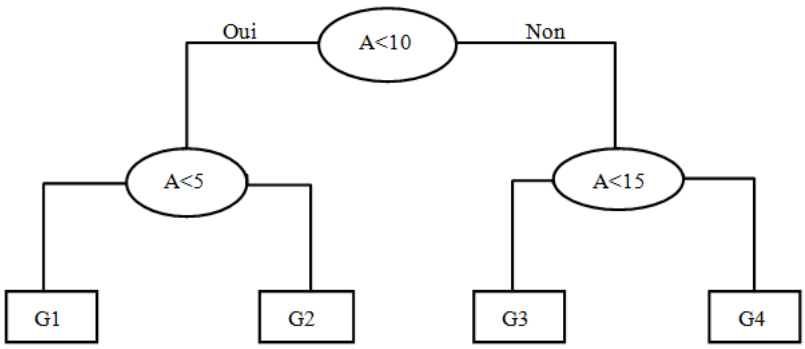

FIGURE 4.5 – Exemple d'arbre CART

L'exemple montre que l'arbre constitué à partir de l'algorithme CART part d'un nœud

initial qui contient toutes les observations puis partitionne les observations en deux groupes en s'appuyant sur l'une des variables. Ce mécanisme est répété jusqu'à l'obtention des feuilles terminales.

La démarche de l'algorithme CART se compose de deux étapes clés à savoir la construction de l'arbre maximale et l'élagage de cet arbre pour obtenir l'arbre optimal :

Construction de l'arbre maximale

Comme tout autre arbre, la construction de l'arbre maximal nécessite :

- Le choix du critère sur lequel s'appuyer pour identifier le meilleur partitionnement parmi les partitionnements proposés par chaque variable explicative. Ce critère repose sur la mesure de l'impureté associée à chaque nœud. L'impureté est calculée pour chaque nœud afin de définir quel est le partitionnement qui permet d'obtenir deux nœuds fils les plus homogènes possibles.
  L'algorithme CART utilise l'indice de Gini pour le calcul de l'impureté pour les variables cibles qualitatives.
  Pour une position p, un test t, P_g et P_d la proportion d'éléments pour respectivement les positions p_1 et p_1 ; on utilise le test qui maximise $\triangle(p,t)$.

$$\triangle\,(p,t) = Gini(p) - (P_g * Gini(p_1) + P_d * Gini(p_2)) \tag{4.6}$$

- La règle permettant de décider si un nœud est terminal, donc une feuille ou non. L'algorithme permet de définir le nombre minimum requis pour une feuille.
- L'affectation de chaque feuille à une valeur de la variable à expliquer. L'algorithme CART utilise un vote à la majorité pour un arbre de classification. La modalité de la variable la plus représentée est alors attribuée à la feuille. Pour un arbre de régression, il s'agit de la moyenne pondérée des observations de la variable de réponse.

Ces critères permettent de construire l'arbre maximal ou l'arbre saturé. C'est l'arbre le plus profond. Chaque feuille de cet arbre est constituée par une seule observation. L'arbre maximal comportant un grand nombre de feuilles terminales expose à un problème de sur-apprentissage. L'objectif de toute classification étant d'éviter le sur-ajustement, un élagage de l'arbre est fait afin de garder un niveau correct de subdivision.

Élagage de l'arbre maximal

L'élagage consiste à sélectionner successivement un nœud de décision, le transformer en une feuille terminale et ainsi supprimer les branches au-delà de ce nœud. Il permet donc de réduire la taille de l'arbre et de diminuer l'erreur de classification. L'élagage peut s'effectuer en utilisant un échantillon d'apprentissage A et un échantillon de test T.

— Construction d'une suite de sous arbres élagués $T_1...T_i...T_n$ à partir de l'arbre maximal T_0 obtenu. Le sous arbre T_{i+1} est obtenu par l'élagage T_i.

— Pour toute position p d'un sous arbre T_i, soit u_p le sous arbre associé à p. On définit

la quantité :

$$g(p) = \frac{\triangle app(p)}{|u_p| - 1} \tag{4.7}$$

Avec $\triangle app(p)$ la variation d'erreur mesurée sur l'ensemble d'apprentissage A lorsqu'on élague T_i en position p et $|u_p|$ la taille de u_p.

— $\triangle app(p)$ est donc égale à :

$$\triangle\, app(p) = \frac{MC(p) - MC(u_p)}{Card(p)} \tag{4.8}$$

Avec Card(p) le cardinal de l'ensemble des exemples de l'échantillon d'apprentissage associés à la position p
MC(p) est le nombre d'éléments de cet échantillon mal classés à la position p lorsqu'on élague T_i en position p
MC(u_p) est le nombre d'éléments de cet échantillon associés à la position p de T_i mal classés par u_p.

— On retient la position p pour laquelle g(p) est minimale et T_i est coupé en cette position p pour obtenir T_{i+1}.

— On utilise l'échantillon d'apprentissage et on calcule pour chaque T_i (de la suite construite) l'erreur apparente. L'arbre optimal est alors l'arbre qui minimise l'erreur apparente sur l'ensemble de test T.

4.1.2.2 Application des CART à notre étude :

Nous avons utilisé le package *rpart* sous R pour effectuer notre classification. Nous avons choisi de construire dans un premier temps un arbre maximal, donc de ne pas fixer des critères de restriction sur la taille d'un nœud. Cette approche nous permet d'utiliser l'estimation de la performance que propose la fonction rpart pour définir le bon nombre de feuilles. La fonction rpart procède par validation croisée pour chaque niveau de simplification. On a donc les performances par validation croisée montrées par la figure 4.6. En abscisse nous avons la complexité de l'arbre en fonction du nombre de feuilles et en ordonnée l'erreur relative associée. La courbe décrit l'erreur de classification obtenue en faisant une validation croisée. Nous retiendrons donc le nombre de feuilles pour lequel l'erreur relative est la plus faible. Dans notre cas, il s'agit de 13.

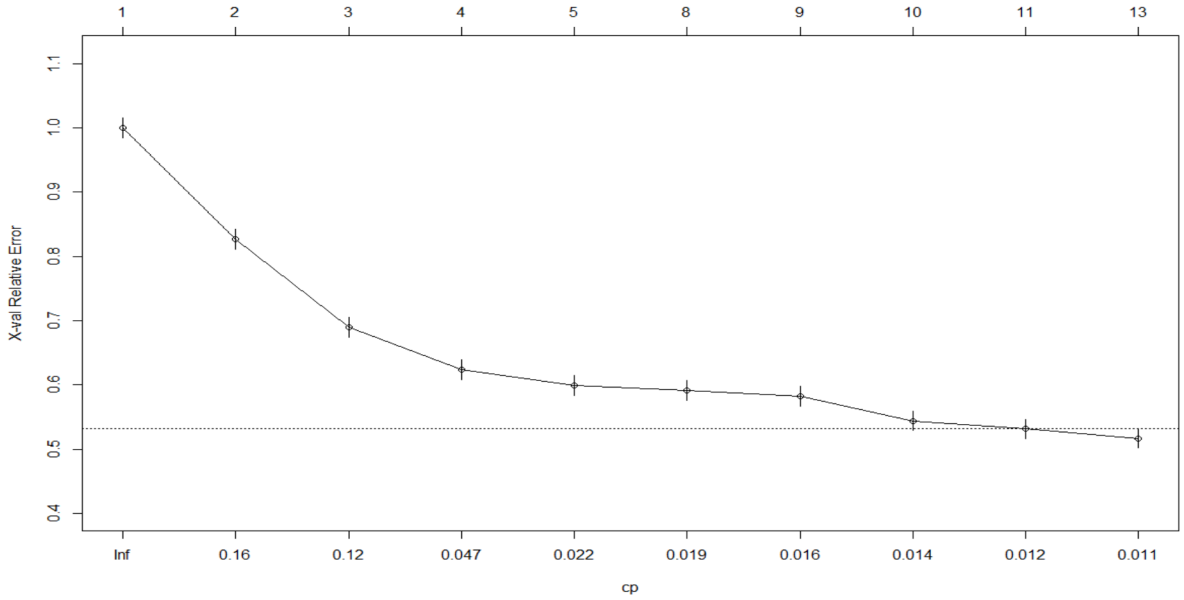

FIGURE 4.6 – Performance de l'arbre de décision

Nous avons construit un arbre optimal (voir figure 4.7) en fixant le nombre de feuilles souhaité à 13. Nous allons donc analyser les résultats de l'arbre afin de ressortir les caractéristiques de chaque groupe.

Interprétation de l'arbre :

L'arbre de décision optimal réalisé utilise 8 indicateurs pour l'élaboration des règles pouvant caractériser nos groupes. Ces huit indicateurs sont :

- le chiffre d'affaires (CA_IARD) qui constitue un indicateur important pour apprécier le poids de l'agent dans l'activité ;
- le taux de devis realisé (TxDevis) qui donne le potentiel business du distributeur ;
- le taux de production auto (TxPROD_AutoP) et le taux de résiliation auto (TxRESIL_AutoP) qui représentent la dynamique commerciale en auto du distributeur ;
- le taux de croissance dommages particuliers (TxCROIS_DomP) qui représente la dynamique commerciale en dommages particuliers du distributeur ;
- le ratio combiné total IARD (RC_TOT) qui est le seul indicateur technique parmi les huit mais reste assez représentatif des résultats techniques du distributeur ;
- la sous-tarification moyenne appliquée par le distributeur en auto (SousTarifAuto) ;
- le nombre de mois de réserve de crédit commercial dont dispose l'agent en ARCC (NbMoisConsoArcc).

Nous proposons d'aborder l'analyse de l'arbre à travers nos cinq groupes.

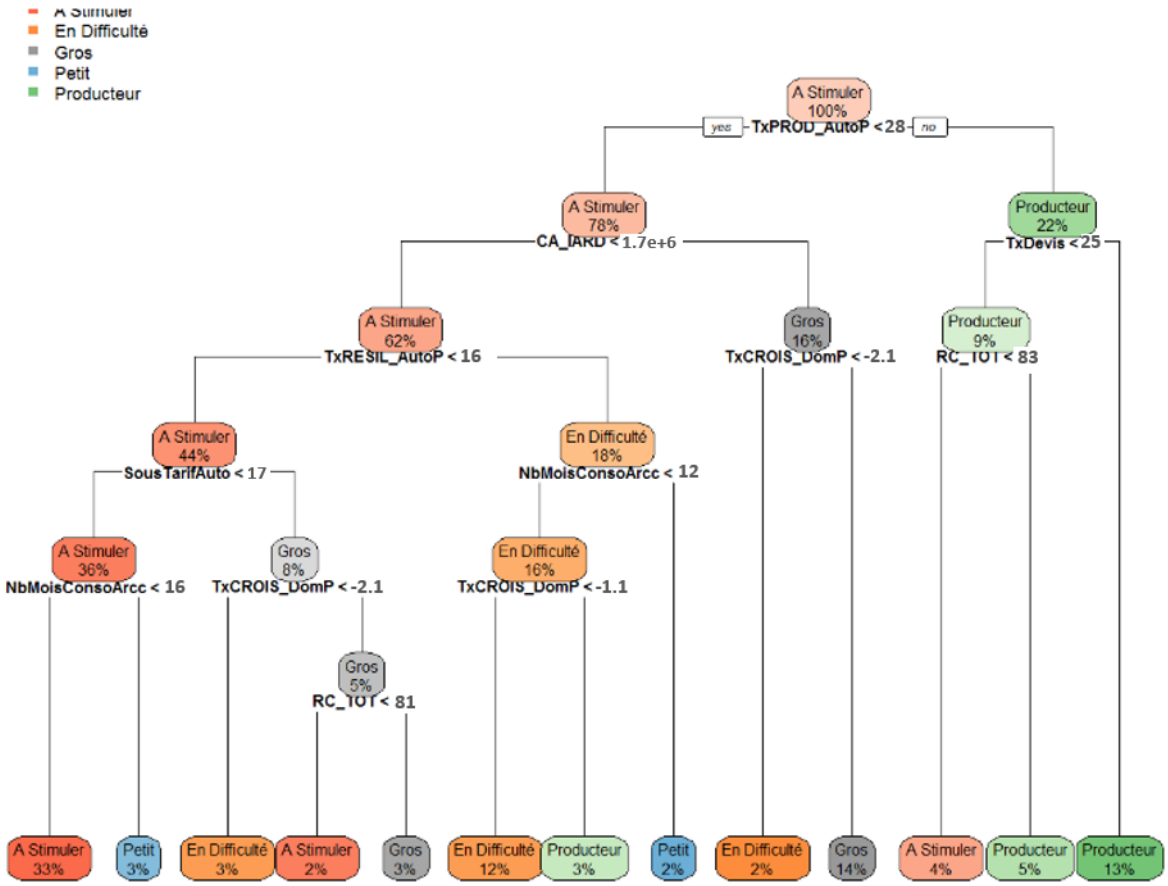

FIGURE 4.7 – Arbre de décision de classification des agents

Les Producteurs :

Dans notre arbre, les **Producteurs** sont présents au niveau de trois feuilles avec plus de la moitié d'entre eux sur une seule feuille. Nous avons donc trois règles qui définissent les **Producteurs** :

— La feuille regroupant 13% des distributeurs est constituée d'agents ayant un taux de devis supérieur ou égale à 25% (soit un taux de devis supérieur au troisième quartile) et une production auto supérieur à 28% (supérieur au troisième quartile). Ce sont donc les distributeurs qui ont beaucoup d'opportunités avec une forte production en auto.

— La deuxième feuille représentant les **Producteurs** (soit 5% des distributeurs) est constituée de distributeurs ayant un ratio combiné total supérieur ou égale à 83 (soit un ratio combiné supérieur à la moyenne des distributeurs) et leur production auto est également supérieure à 28%. Cette règle montre que les **Producteurs** qui ne font pas beaucoup de devis n'ont pas de bons résultats techniques.

— La dernière feuille faisant apparaître les **Producteurs** ne comporte que 3% des distributeurs. Les producteurs de cette feuille ont une croissante en dommages particuliers supérieure à la moyenne nationale et ont moins de 12 mois de budget commercial ARCC en réserve. Ils ont un taux de résiliation auto supérieur à 16% (plus que la moyenne nationale) avec un chiffre d'affaires pas très élevé (moins que le troisième quartile) et un taux de production auto inférieur à 28%. Il s'agit des **Producteurs** en difficulté en auto mais qui s'appuient sur les dommages particuliers pour améliorer leur production.

Ces trois règles montrent que les **Producteurs** sont ceux qui produisent beaucoup en auto et qui font beaucoup de devis au détriment de la rentabilité. Mais il n'y a pas que l'auto qui définit les **Producteurs**, la dernière règle nous montre qu'il y a une petite partie des **Producteurs** qui sont plutôt orientés dommages particuliers.

Les autres groupes :

L'analyse des règles des autres groupes se fait de façon similaire. En résumé,

— **les Agents à stimuler** sont caractérisés par de bons résultats techniques et une sous-tarification moyenne pas élevée (moins que la médiane) ;

— **les Agents en difficulté** ont un apport net en dommages particuliers (taux de croissance) très négatif (inférieur à la médiane) ;

— **les Gros consommateurs** sont surtout définis par leur chiffre d'affaires élevé ou leur niveau de sous-tarification auto élevé lorsque leur ratio combiné n'est pas très bon.

— **les Petits consommateurs** sont caractérisés par leur grand réserve (soit plus que le dernier quartile) en budgets commerciaux.

4.1.2.3 Conclusion sur les CART

Au regard des règles issues de notre arbre de décision, nous avons noté que la définition de nos groupes n'est pas différente de celle que nous avions faite à partir de notre segmentation k-means. De plus, l'arbre nous a permis d'obtenir des règles qui permettent un découpage plus simple. Néanmoins, notre arbre ne classe que 76% des agents correctement, ce qui reste moins important que la régression logistique multinomiale.

	Groupes k-means					
Groupes CART	**Agents en difficulté**	**Gros consommateurs**	**Petits consommateurs**	**Producteurs**	**Agents à stimuler**	**Total**
Agents en difficulté	**13,9%**	1,1%	0,7%	0,4%	2,8%	18,9%
Gros consommateurs	1,2%	**13,3%**	0,5%	0,8%	2,0%	17,8%
Petits consommateurs	0,4%	0,1%	**4,9%**	0,2%	0,8%	6,3%
Producteurs	1,2%	2,4%	0,9%	**15,0%**	1,2%	20,8%
Agents à stimuler	2,5%	2,3%	0,5%	2,1%	**28,9%**	36,2%
Total	19,2%	19,2%	7,4%	18,6%	35,6%	100,0%

FIGURE 4.8 – Matrice de classification des CART

4.1.3 Conclusion sur la caractérisation des groupes

Après avoir appliqué ces deux méthodes pour mieux caractériser nos groupes, nous pouvons récapituler les deux méthodes comme décrit dans le tableau ci-dessous. Sur la base de cette comparaison et par rapport à notre besoin d'avoir des critères assez compréhensibles pour distinguer nos groupes, nous allons choisir l'arbre de décision avec les réajustements qui en découlent. Nous obtenons ainsi des groupes basés sur les règles de l'arbre CART.

Ces groupes sont caractérisés par les huit indicateurs qu'utilisent l'arbre et permettent de faire ressortir les indicateurs de production, de maîtrise de coût et d'utilisation de budgets commerciaux.

	GLM	CART
Avantages	-Permet d'évaluer l'importance de chaque indicateur sur chaque groupe et comment il faut faire évoluer cet indicateur pour changer de groupe	-Définit des règles simples d'appartenance à chaque groupe -Utilise seulement 8 indicateurs pour expliquer les groupes d'agents -Permet de définir des objectifs quantifiables à chaque groupe sur la base des critères de l'arbre
Faiblesses	-Explication des groupes avec plusieurs variables -Pas de règles simples mais uniquement des probabilités - Difficile de s'appuyer sur ces critères pour expliquer à chaque agent son groupe	-Beaucoup d'agents (23%) doivent être réajuster pour rester cohérent avec les règles établies par l'arbre de décision

TABLE 4.1 – Comparaison GLM vs CART pour la caractérisation des groupes

Après l'ajustement des groupes avec l'arbre CART réalisé, la figure 4.9 présente quelques chiffres clés par groupe.

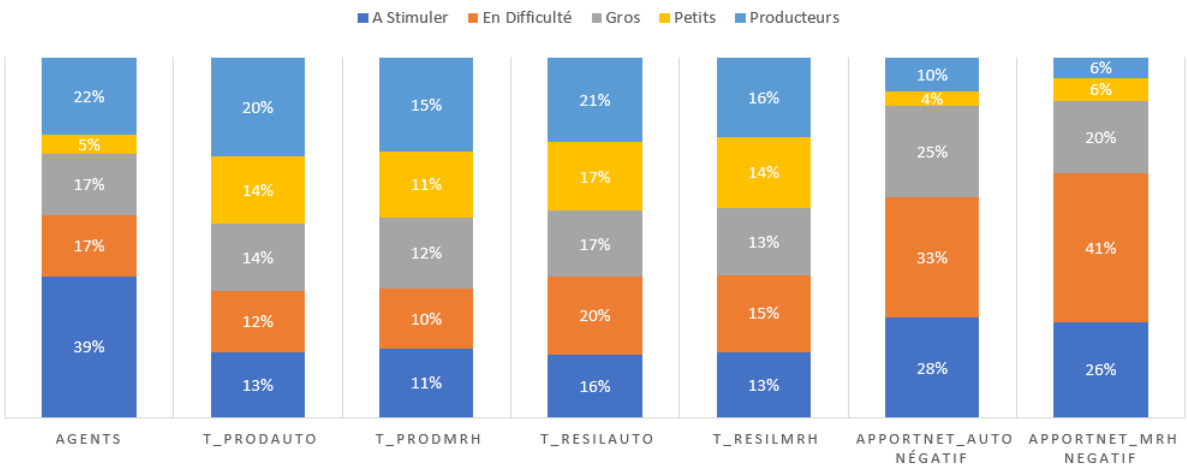

FIGURE 4.9 – Statistiques des groupes sur la dégradation de l'apport net

Nous constatons que les **Producteurs** contribuent très peu à la dégradation de l'apport net. Ce qui signifie que leur taux de résiliation élevé est fortement compensé par leur fort taux de production. A l'opposé les **Gros consommateurs** et les **Agents en difficulté**

contribuent fortement à la dégradation de l'apport net. Les deux groupes contribuent à près de 60% de l'apport net négatif alors qu'ils ne représentent que 34% des agents. Ce qui confirme la nécessité d'avoir une segmentation des agents et de pouvoir le mettre à jour en fonction de l'évolution des distributeurs afin d'adapter les stratégies d'accompagnement.

4.2 Généralisation de la segmentation

Après avoir caractérisé nos groupes, il faut vérifier la stabilité de la segmentation et proposer une généralisation. Il s'agit d'un des objectifs de l'étude. En effet, la segmentation doit être mise à jour d'année en année pour que les mesures d'accompagnement s'adaptent au nouveau groupe de chaque agent. Il faut donc que le modèle de segmentation soit stable et cohérent pour qu'on puisse l'utiliser pour classer les agents lorsqu'on met à jour les indicateurs (utilisés pour la construction de l'arbre).

Nous proposons d'utiliser la base de données de test que nous avons construit à partir des données 2015-2016 (comme expliqué au chapitre 2) pour mesurer la stabilité de la segmentation. Cette base de test servira pour identifier le groupe qu'aurait eu chaque agent en 2015-2016.

Sachant que les indicateurs qui ont permis la construction de l'arbre de décision dépendent pour la plupart de l'exercice, le changement d'exercice engendre un changement de taux de production ou de taux de devis. Un taux de devis qui représente un bon taux une année peut ne pas l'être une autre année compte tenu du contexte et des performances de chaque agent. L'enjeu est donc de savoir s'il faut utiliser une généralisation de la segmentation basée sur les indicateurs normalisés ou non. Nous allons comparer les résultats des deux approches pour décider quelle approche correspond le mieux à nos données.

Enfin, pour une meilleure appréciation de ce test de stabilité, il faut tenir compte des changements dans le temps. Les agents peuvent subir des mouvements de portefeuilles, ce qui peut modifier les indicateurs d'un distributeur d'une année à l'autre. En effet,

— un distributeur peut reprendre le portefeuille d'un autre distributeur. Ce qui va augmenter son chiffre d'affaires et changer son ratio combiné. Un tel changement peut affecter également son taux de production ainsi que son utilisation des budgets commerciaux.
— un distributeur peut se mettre en association avec un autre distributeur. Ils deviennent alors une association. L'association qu'ils forment aura des indicateurs qui seront différents de ceux de chacun d'eux.

Ces changements affectent 15% à 20% des distributeurs. Il est donc normal que certains agents changent de groupe.

4.2.1 Généralisation non contextuelle

Dans cette approche, nous ne tenons pas compte du contexte de chaque année. Nous utilisons les règles identifiées à partir des données de notre base 2016-2017 comme référence. Nous allons utiliser les indicateurs non normalisés et classer les agents directement à partir de la base de test et des critères de l'arbre de décision. Les résultats de la classification sont présentés dans la matrice de confusion 4.10.

	Groupes 2017					
Groupes 2016	**Agents en difficulté**	**Gros consommateurs**	**Petits consommateurs**	**Producteurs**	**Agents à stimuler**	**Total**
Agents en difficulté	**7,0%**	0,6%	0,6%	1,6%	2,3%	12,0%
Gros consommateurs	2,9%	**14,0%**	0,2%	1,4%	1,5%	20,0%
Petits consommateurs	0,3%	0,1%	**3,2%**	0,6%	1,1%	5,2%
Producteurs	2,2%	0,9%	0,4%	**13,1%**	1,8%	18,4%
Agents à stimuler	4,8%	1,4%	1,1%	4,9%	**32,1%**	44,4%
Total	17,2%	16,9%	5,4%	21,7%	38,7%	100,0%

FIGURE 4.10 – Matrice de confusion de classification avec des indicateurs non normalisés

L'approche non contextuelle avec des données non normalisées montre que 71% des distributeurs n'ont pas changé de groupe. Seuls 29% ont changé de groupe. Un grand nombre de changements concerne les **Agents en difficulté** et les **Producteurs**. L'appartenance à ces groupes est basée sur le taux de croissance et le taux de production. Il s'agit d'indicateurs qui sont très volatiles.

Enfin, on peut noter une forte diminution du nombre d'**Agents en difficulté** au profit des **Agents à stimuler**. Ce qui s'explique à la lecture des règles de l'arbre par une baisse de la moyenne de sous-tarification auto au niveau de la base de test. En d'autres termes, les données 2015-2016 montrent une sous-tarification moyenne en auto plus faible des **Agents en difficulté**. C'est ce qui explique leur passage au groupe d'**Agents à stimuler**.

Cette généralisation est donc sensible aux tendances des indicateurs, ce qui peut conduire à la disparition d'un groupe.

4.2.2 Généralisation contextuelle

Dans cette approche, nous faisons l'hypothèse forte d'utiliser des données normalisées pour les critères de l'arbre de décision et ainsi de faire une classification avec des données normalisées. Cette approche donne 72% de distributeurs qui ont gardé le même groupe. Mais ces résultats sont légèrement différents des résultats que nous avons obtenus dans l'approche précédente. Les groupes sont plus stables et les **Agents en difficulté** sont moins affectés.

Cette stabilité s'explique par le fait que les indicateurs utilisés sont normalisés. La classification permet donc d'avoir des groupes plus homogènes. Il s'agit donc d'une classification au sein des agents qui ne dépend pas d'une référence fixée à partir des performances d'un exercice donné.

	Groupes 2017					
Groupes 2016	**Agents en difficulté**	**Gros consommateurs**	**Petits consommateurs**	**Producteurs**	**Agents à stimuler**	**Total**
Agents en difficulté	**9,6%**	1,4%	0,6%	2,7%	4,6%	18,9%
Gros consommateurs	1,7%	**13,3%**	0,1%	1,2%	1,1%	17,4%
Petits consommateurs	0,3%	0,1%	**3,3%**	0,6%	1,1%	5,5%
Producteurs	2,4%	1,1%	0,5%	**13,3%**	2,0%	19,2%
Agents à stimuler	3,3%	1,0%	0,9%	3,8%	**29,9%**	39,0%
Total	17,2%	16,9%	5,4%	21,7%	38,7%	100,0%

FIGURE 4.11 – Matrice de confusion de classification avec des indicateurs normalisés

Nous pouvons donc conclure que l'approche de généralisation contextuelle garantit une meilleure stabilité et permet de garder des groupes plus homogènes. Nous préconisons donc une généralisation contextuelle même si la normalisation peut engendrer une perte d'informations.

4.3 A/B Testing avec une dotation ST aux groupes

L'objectif de la segmentation étant de proposer des mesures adaptées à chaque groupe, les A/B tests constituent un bon moyen pour mesurer l'impact d'un accompagnement sur chacun des groupes. En effet, l'A/B test est une technique utilisée dans le marketing pour tester différentes hypothèses (ou différentes variantes) d'une solution afin d'en choisir la plus optimale. Nous l'avons adapté à notre étude pour tester deux variantes d'une mesure d'accompagnement. Ce test a permis d'analyser l'effet de cet accompagnement sur chaque groupe et de conclure sur l'importance de différencier les mesures pour nos groupes.

La caractérisation des groupes montre que les agents n'utilisent pas les budgets commerciaux de la même façon et leurs besoins en budgets ne sont pas les mêmes. A partir des données de 2018, nous avons construit deux scénarii qui nous permettent d'effectuer un A/B test. En effet, une dotation exceptionnelle a été faite à certains agents afin de booster leur production. Cette dotation nous a servi de base pour identifier deux hypothèses à tester avec un échantillon de 900 agents pour chaque cas :

— Hypothèse 1 : les agents n'ayant pas reçu de dotation
— Hypothèse 2 : les agents ayant reçu une dotation

L'A/B test a servi alors à comparer la production de chaque groupe suivant chaque hypothèse. La dotation ayant été faite à la fin du premier trimestre, nous avons choisi de comparer la production sur le deuxième trimestre de chaque agent par rapport à sa production à la même époque l'année dernière. Le deuxième trimestre correspond à la période juste après la dotation et nous permet de capter l'effet immédiat de cette dotation. De plus, la comparaison avec le deuxième trimestre de l'année précédente s'explique par le fait que chaque période de l'année est spécifique et il est plus judicieux de comparer des périodes identiques. Nous avons donc analysé la variation en affaires nouvelles de chaque agent.

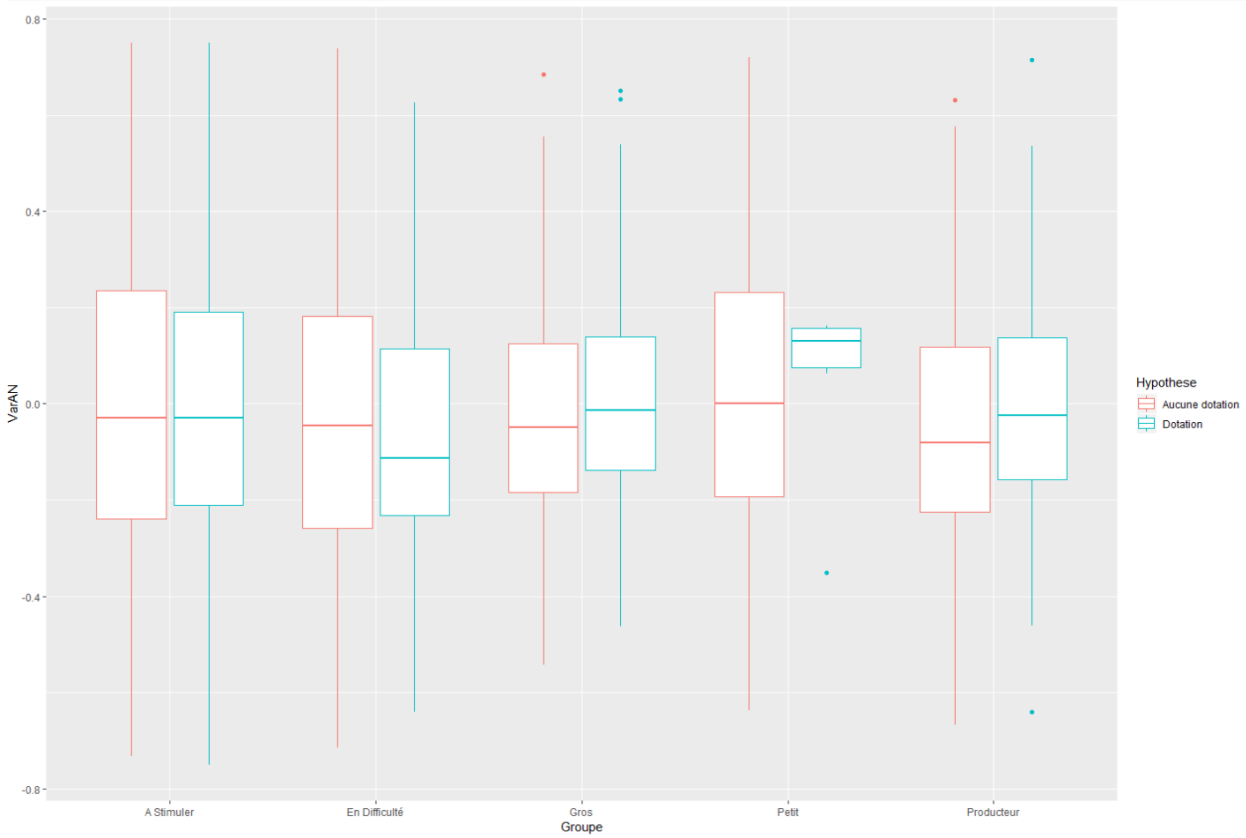

FIGURE 4.12 – AB Test : Variation des affaires nouvelles par groupe

La dotation ayant été faite suivant des critères de production et de réserve en budget ST, le choix des agents n'est donc pas aléatoire. Cela représente donc un biais pour notre analyse. C'est ce qui explique aussi le faible nombre d'agents **Petits consommateurs** dans le groupe des agents ayant bénéficié de cette dotation. Il est donc difficile d'analyser les résultats du test pour les **Petits consommateurs**.

La variation des affaires nouvelles observée sur la figure 4.12 permet de tirer trois conclusions :

— La dotation n'a eu aucun effet sur la production des **Agents à stimuler** . Ces agents étant concentrés sur leur portefeuille et orientés résultats, ils sont donc très peu sensibles à une dotation supplémentaire. Nous pouvons en déduire qu'il faut les inciter ou aider à se créer des devis afin de pouvoir utiliser cette dotation.

— Les **Agents en difficulté** qui ont reçu une dotation ont moins produit par rapport à ceux qui n'ont pas reçu. On a donc l'effet contraire à ce que l'on souhaitait. Il faut donc identifier des mesures autres que la dotation pour relancer leur production.

— Les **Gros consommateurs** et les **Producteurs** ont su utiliser cette dotation pour améliorer leur production par rapport à l'année précédente. On constate une nette différence entre les **Producteurs** qui ont reçu une dotation et ceux qui n'en n'ont

pas eu. L'incitation avec une dotation a donc eu un effet positif sur la production de ces groupes.

Ce Test montre l'importance d'adapter les mesures aux besoins de chaque groupe. Il s'agit d'une mesure adaptée uniquement aux **Gros consommateurs** et aux **Producteurs**. Le budget alloué aux **Agents à stimuler** et **Agents en difficulté** est inefficace et constitue une perte de ressources. L'expérience a montré également que la dotation doit être accompagnée d'autres mesures pour espérer un impact dans le cas des **Agents à stimuler**. Il est donc important de trouver les accompagnements qui permettront à chaque groupe de se développer.

4.4 Mesures opérationnelles

A l'issue de la segmentation et des règles de classifications permettant de définir chaque groupe, nous pouvons proposer quelques mesures pour améliorer la rentabilité de chaque groupe tout en développant leur production :

Mesures adaptées aux Petits consommateurs :

Ils constituent le groupe qui utilise le moins les budgets commerciaux pour développer leur chiffre d'affaires, il faut :

— Former ces agents sur la connaissance des produits IARD particuliers professionnels ainsi que sur les latitudes commerciales qui existent. Cette formation aura pour but de les aider à développer l'activité IARD.

— Privilégier leur tarification en affaires nouvelles pour les accompagner à attirer de nouveaux clients afin de développer l'activité.

— Initier un accompagnement personnalisé par les directeurs de vente ou les inspecteurs pour accompagner ces agents. Ces accompagnements permettront d'identifier le potentiel commercial de l'emplacement géographique de ces agents pour définir une stratégie de développement avec eux.

— Inciter ces agents à se mettre en association avec d'autres agents afin de constituer de grands ensembles. Cette mesure leur permettra d'avoir un portefeuille plus grand et donc une clientèle plus large à multi-équiper.

Ils représentent le groupe ayant le moins d'impact sur les résultats techniques IARD particuliers professionnels.

Mesures adaptées aux Gros consommateurs

C'est le groupe qui impacte le plus le développement de l'activité. Pour améliorer leur production et réduire leur consommation en budget, il faut :

— Surveiller leur utilisation des budgets commerciaux.
— Inciter ces agents à produire davantage en améliorant leur tarification en affaires nouvelles.
— Orienter ces agents vers les nouveaux produits où ils n'ont pas besoin de mettre des budgets commerciaux. Ce qui permettra de réduire leur consommation en budget.
— Éviter qu'ils achètent des points CTP pour avoir plus de budgets commerciaux.

L'objectif pour ce groupe sera de relancer la production tout en réduisant les budgets commerciaux à la main de l'agent.

Mesures adaptées aux Producteurs

Les **Producteurs** sont des agents dynamiques qui se créent beaucoup d'opportunités et les transforment en affaires. Ils privilégient le développement mais au détriment des coûts et des résultats techniques. Pour les accompagner vers la rentabilité, il faut :

— Surveiller leur sélection du risque. En effet, il faut former ces agents sur la nécessité de faire une meilleure sélection du risque afin d'améliorer leur résultat technique. Il faut les inciter à mettre du rabais uniquement sur les clients les moins risqués.
— Analyser davantage leur portefeuille pour les aider à limiter le risque de leur portefeuille.
— Surveiller leur dotation et l'utilisation des budgets commerciaux en affaires nouvelles. En effet, leur forte propension à mettre du budget pour attirer de nouveaux clients agit sur leur rentabilité et ne garantit pas une bonne duration des contrats qu'ils souscrivent. C'est ce qui explique leur fort taux de résiliation.
— Inciter ces agents à multi-équiper leurs clients pour baisser le taux de résiliation.
— Privilégier leur tarification au renouvellement (terme) afin de s'assurer d'une baisse de leur taux de résiliation. Cette mesure aura pour but de les aider à protéger leur portefeuille et d'éviter les résiliations au terme.

Les **Producteurs** doivent veiller à améliorer leur résultat technique et à conserver leur portefeuille.

Mesures adaptées aux agents à stimuler

Ces agents présentent de bons résultats mais produisent très peu. Pour les inciter à produire, il faut :

— Orienter plus de leads (devis effectués en ligne) vers ces agents afin qu'ils disposent de plus d'opportunités pour transformer en affaires nouvelles.
— Améliorer leur capacité à créer des opportunités (devis) en s'inspirant des techniques utilisées par les **Producteurs**.
— Proposer une dotation en budget commercial basée sur le nombre de devis réalisés pour les inciter à se créer davantage d'opportunités.
— Analyser le potentiel de leur position géographique et la concurrence dans leur voisinage pour définir une stratégie de relance de leur production.
— Améliorer leur tarification en affaires nouvelles afin de leur permettre d'attirer de nouveaux clients.

Mesures adaptées aux agents en difficulté

Ces agents sont en perte de croissance et non pas de bons résultats techniques. Il faut :
— Travailler leur CTP pour améliorer leur tarification.
— Améliorer leur capacité à transformer les devis qu'ils font.
— Réduire la part de dotation qu'ils peuvent mettre sur le terme afin d'éviter qu'ils continuent à utiliser du budget commercial sur des clients qui ne sont pas rentables.
— Surveiller la qualité de leur souscription afin de les aider à améliorer leur rentabilité.
— Réduire leur délégation élargie à une délégation standard pour mieux contrôler leur souscription.
— Développer leur production en les incitant à mettre plus de budgets en affaires nouvelles tout en privilégiant le multi-équipement.

Conclusion et perspectives

Les mesures d'accompagnement constituent un coût non négligeable et leur optimisation est nécessaire afin qu'elles soient orientées vers la production et la rentabilité. Une manière de les optimiser est de les adapter aux besoins de chaque distributeur. Cette étude a proposé une identification des besoins de chaque agent en s'appuyant sur la production de l'agent, sa maîtrise technique et son utilisation des budgets commerciaux. La création de plusieurs indicateurs a permis de cerner le comportement de l'agent à travers ces trois dimensions et de créer cinq groupes d'agents. La caractérisation et l'analyse de ces groupes a permis de proposer des mesures d'accompagnement qui sont adaptées de la situation de chaque agent.

La segmentation des agents proposée dans cette étude a montré :

— l'importance des devis dans la production. En effet, les agents qui font le plus de devis sont ceux qui produisent le plus.
— la difficulté pour les agents de concilier la production et la rentabilité.
— la nécessité de surveiller et de mieux contrôler l'utilisation des budgets commerciaux pour certains agents.

L'étude permet également de faire une répartition plus efficace des budgets commerciaux en fonction des besoins de chaque groupe mais aussi d'améliorer la surveillance des portefeuilles agents. De plus, les résultats de l'étude serviront pour faire des ajustements du tarif au terme par agent afin de protéger les agents ayant une bonne rentabilité et dont le portefeuille est sensible au tarif au renouvellement. Enfin, Nous pouvons utiliser les résultats de cette segmentation comme socle pour définir les délégations de souscription de chaque agent.

Si cette étude a permis de segmenter les agents, elle est toutefois basée sur le comportement passé des agents. Elle ne tient donc pas compte de leur potentiel futur. La valeur future des affaires nouvelles peut permettre d'apprécier le potentiel de chaque agent. Ce potentiel peut être utilisé pour une meilleure segmentation. Une nouvelle dimension pour compléter cette étude serait d'ajouter l'ELR (Expected Loss Ratio) qui est indicateur de la rentabilité avancée des contrats de chaque agent.

De plus, la segmentation n'a pas pris en compte les comportements déviants des agents. En effet, certains comportements, comme la déclaration incorrecte des risques à la souscription, peuvent expliquer les mauvais résultats techniques. En guise de perspectives, nous proposons donc d'analyser ces comportements déviants et d'identifier de nouveaux indicateurs qui permettront une segmentation plus fine.

Pour finir, les résultats de cette segmentation ont suscité quelques interrogations concer-

nant les transferts de portefeuille entre agents et les associations tels qu'ils se font actuellement. Quelle est la pertinente et l'impact sur la rentabilité de céder le portefeuille d'un agent d'un groupe donné à un agent d'un autre groupe ? Quelle est l'impact de mettre en association les portefeuilles d'un **Agent en difficulté** et d'un **Agent producteur** ou d'un autre groupe en général ? Quel groupe est le plus adéquat pour reprendre le portefeuille provenant d'un groupe donné afin de favoriser le développement du portefeuille tout en veillant sa rentabilité ?

Annexe A

Définition des indicateurs de l'étude

Taux de Devis :
C'est le nombre de devis réalisés par l'agent rapporté à la taille de son portefeuille de l'année précédente. Cet indicateur met en exergue les possibilités qu'à l'agent de faire de la production. Le taux de Devis montre le potentiel qu'à une agence pour se développer.

$$TxDevis = \frac{Nombre\ de\ devis\ réalisés\ l'année\ n}{Nombre\ de\ contrats\ en\ portefeuille\ année\ (n-1)}$$

Taux de Transformation prospects :
C'est le nombre de devis prospects concrétisés (affaires nouvelles) sur le nombre de devis effectués pour des prospects . Cette variable montre la capacité de l'agent à attirer de nouveaux clients.

$$TxTransf_Pros = \frac{Nombre\ de\ devis\ prospects\ concrétisés}{Nombre\ de\ devis\ prospects\ réalisés}$$

Taux de Transformation clients :
C'est les devis clients concrétisés sur le nombre de devis effectués pour des clients de l'agence. Le taux de transformation clients d'un agent donne sa capacité à multi-équipé ses clients.

$$TxTransf_Cli = \frac{Nombre\ de\ devis\ clients\ concrétisés}{Nombre\ de\ devis\ clients\ réalisés}$$

Taux de production :
C'est le nombre d'affaires nouvelles réalisées sur l'année rapporté au portefeuille de l'année précédente de l'agent. Le taux de production caractérise la progression du portefeuille de l'agent d'une année à une autre.

$$TxPROD = \frac{Nombre\ d'affaires\ nouvelles\ réalisées\ l'année\ n}{Nombre\ de\ contrats\ en\ portefeuille\ année\ (n-1)}$$

Taux de résiliation :

C'est le nombre de résiliations de l'année rapporté au portefeuille de l'année précédente. Le taux de résiliation caractérise la part de portefeuille qu'à perdu l'agent sur les douze derniers mois.

$$TxRESIL = \frac{Nombre\ de\ résiliations\ de\ l'année\ n}{Nombre\ de\ contrats\ en\ portefeuille\ année\ (n-1)}$$

Taux de croissance :
C'est l'apport net rapporté au portefeuille de l'année précédente. Le taux de croissance se différencie du taux de production et du taux de résiliation par sa capacité à mesurer sa croissance réelle. Il caractérise la diminution ou l'augmentation du portefeuille.

$$TxCROIS = TxPROD - TxRESIL$$

Le portefeuille :
C'est le nombre de contrats IARD en portefeuille de l'agent. Il caractérise la taille du portefeuille de l'agent. La plupart des indicateurs étant des taux, cette variable permettra d'analyser les indicateurs en tenant compte également de la taille du portefeuille.

Le chiffre d'affaires :
Il caractérise le poids de l'agent dans la production globale et pourrait aider à savoir à quel point l'agent influence la production globale.

La part de Clic and Go :
C'est le nombre d'affaires nouvelles "Clic and GO" rapporté au nombre d'affaires nouvelles auto. Cette variable caractérise la propension de l'agent à utiliser le nouveau produit auto. Nous nous intéressons à ce produit car il s'agit d'un produit pour lequel l'utilisation du budget commercial par l'agent est revue et limitée.

$$PartClicAndGo = \frac{Nombre\ d'affaires\ nouvelles\ ClicAndGo}{Nombre\ d'affaires\ nouvelles\ en\ Auto}$$

La part de ma maison :
C'est le nombre d'affaires nouvelles "Ma Maison" rapporté aux affaires nouvelles MRH. Ma maison est également le nouveau produit MRH qui limite l'utilisation du budget par l'agent.

$$PartMaMaison = \frac{Nombre\ d'affaires\ nouvelles\ MaMaison}{Nombre\ d'affaires\ nouvelles\ en\ MRH}$$

Sous-tarification moyenne :
C'est le rapport entre la cotisation moyenne après rabais et la cotisation moyenne avant

rabais. Cette variable caractérise le niveau de rabais moyen qu'effectue le distributeur.

$$Sous_Tarif = \frac{Tarif\ moyen\ après\ réduction\ appliqué\ par\ l'agent}{Tarif\ moyen\ avant\ réduction}$$

Part de contrats avec réduction :
C'est le nombre de contrats en portefeuille ayant bénéficié d'un rabais rapporté au nombre de contrats en portefeuille. Cet indicateur nous permet de mesurer la nécessité du rabais dans la production de l'agent.

$$PartCntRed = \frac{Nombre\ de\ contrat\ avec\ réduction}{Nombre\ de\ contrat\ en\ portefeuille}$$

La part du budget utilisée sur le terme :
C'est la part du ponctuel et du reconductible qui sont utilisées sur le terme. Cette variable permet d'avoir la proportion du budget qui est utilisée par l'agent pour agir sur le Terme.

$$Terme = \frac{Budget\ commercial\ utilisé\ pour\ le\ terme}{Budget\ commercial\ total\ utilisé\ par\ l'agent}$$

La part de budget utilisée pour les mono-détenteurs :
Cette variable permet de capter la part du budget que l'agent utilise pour des clients mono-détenteur. Il s'agit d'une variable qui caractérise une utilisation pas très efficiente du budget alloué. Car l'utilisation du budget pour des mono détenteurs ne favorise pas le multi-équipement.

$$Mono = \frac{Budget\ commercial\ utilisé\ pour\ les\ mono-détendeurs}{Budget\ commercial\ total\ utilisé\ par\ l'agent}$$

La part de budget utilisée pour les clients sinistrés :
C'est la part du budget qui est utilisé pour faire du rabais à des clients déjà sinistrés. Cette variable caractérise une utilisation du budget pour favoriser des profils risqués.

$$CliSin = \frac{Budget\ commercial\ utilisé\ pour\ les\ clients\ sinistrés}{Budget\ commercial\ total\ utilisé\ par\ l'agent}$$

L'impact du budget commercial sur le chiffre d'affaires :
C'est le rapport entre la part du budget commercial utilisée par l'agent sur le chiffre d'affaires de l'agent. C'est une variable qui met en évidence si l'agent utilise beaucoup de

budget ou non.

$$TxRabais = \frac{Budget\ commercial\ utilisé\ par\ l'agent}{Chiffre\ d'affaires + Budget\ commercial\ total\ utilisé\ par\ l'agent}$$

La part de budget consommée :
C'est la consommation de l'agent en budget commercial rapporté à la dotation qu'il a reçu. Cet indicateur permet de savoir le niveau de budget que l'agent utilise par rapport à ce qu'il reçoit.

$$Use = \frac{Budget\ commercial\ consommé}{Budget\ commercial\ doté}$$

Le réserve de budget en mois :
C'est le budget disponible rapporté à la consommation moyenne de l'agent. Cette variable exprime le nombre de mois pour lequel l'agent dispose de budget. Ainsi plus cette variable est grand, moins l'agent a besoin de nouvelle allocation de budget.

$$NbMois = \frac{Solde\ budget\ commercial}{Consommation\ mensuelle\ moyenne\ de\ l'agent}$$

Annexe B

Réduction de dimensions avec l'ACP

Cette partie décrit l'utilisation de l'ACP pour la réduction de dimension sur les bases des indicateurs de maîtrise de coût et ceux d'utilisation de budget.

B.1 ACP sur les indicateurs de maîtrise de coût

Dans cette section, nous allons faire une ACP normée sur les indicateurs de maîtrise des coûts par l'agent. En utilisant la même approche que pour les indicateurs de production pour analyser et choisir les composantes à retenir à l'issue de l'ACP, nous allons choisir les composantes ayant une valeur propre supérieure à 1. On constate que seules les trois premières composantes peuvent faire l'objet d'une interprétation. Ces composantes représentent 74.77% de l'inertie totale.

	Valeurs propres	Variances (%)	Variances cumulées(%)
Composante1	3.37	36.33	36.33
Composante2	2.42	26.89	63.23
Composante3	1.09	11.54	74.77
...	...	...	...
Composante9	0.06	0.71	100

Table B.1 – Valeurs propres et variances

Interprétation des composantes :

Première composante :

La figure B.1 de contribution montre les variables SousTarifAuo, RC_TOT,RC_Auto,RC_DomP (ratio combiné), CTP_BASE(coefficient technique personnalisé) et PartCntRedAuto participent à la construction de cette première composante. Les autres variables ne contribuent pas assez à cette composante.

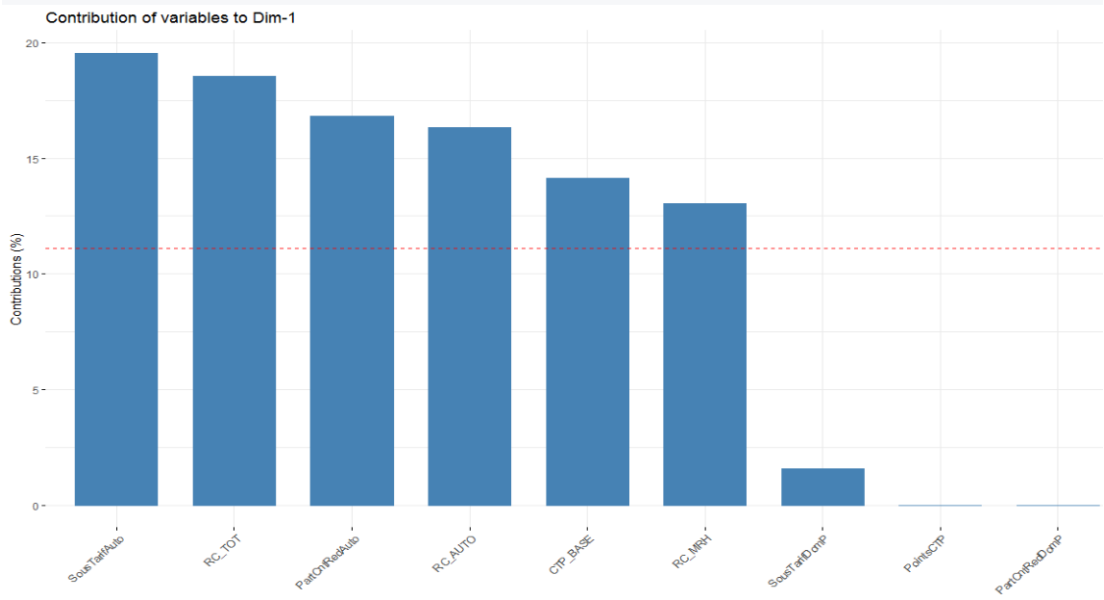

FIGURE B.1 – Contribution à la composante 1

Le cercle de corrélation du premier plan (voir figureB.2) montre également que toutes ces variables sont bien représentées et sont toutes corrélées négativement par rapport à cette première composante. Il s'agit d'un effet de taille qui montre que les agents ayant un mauvais résultat (RC élevé) ont également un CTP élevé et font beaucoup de réduction et de sous-tarification en auto. Il s'agit d'une composante qui exprime **la rentabilité et la réduction auto**.

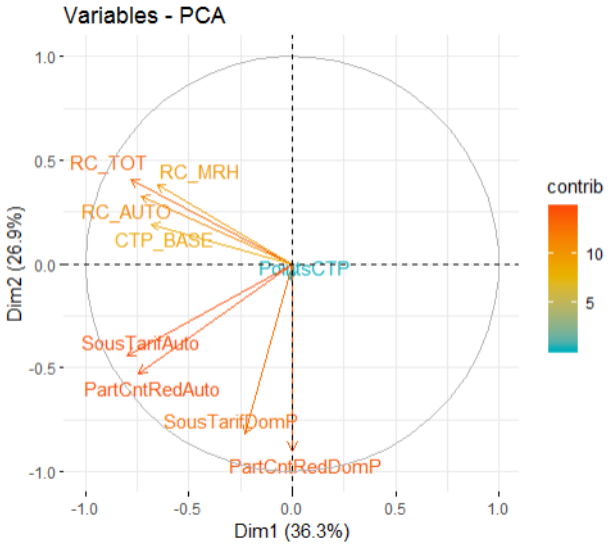

FIGURE B.2 – Cercle de corrélation du premier plan

Deuxième composante :

La figure B.3 montre la contribution des variables à la deuxième composante. On constate que PartCntRedDomP (part de contrat avec réduction) et Sous_Tarif_DomP (sous-tarification) sont les variables qui contribuent le plus à la composante. En utilisation le cercle de corrélation du premier plan, on constate que ces variables sont toutes négativement corrélées avec la composante et sont bien représentées. On peut donc conclure qu'il s'agit d'une composante qui exprime le niveau de **réduction dommages particuliers** par l'agent.

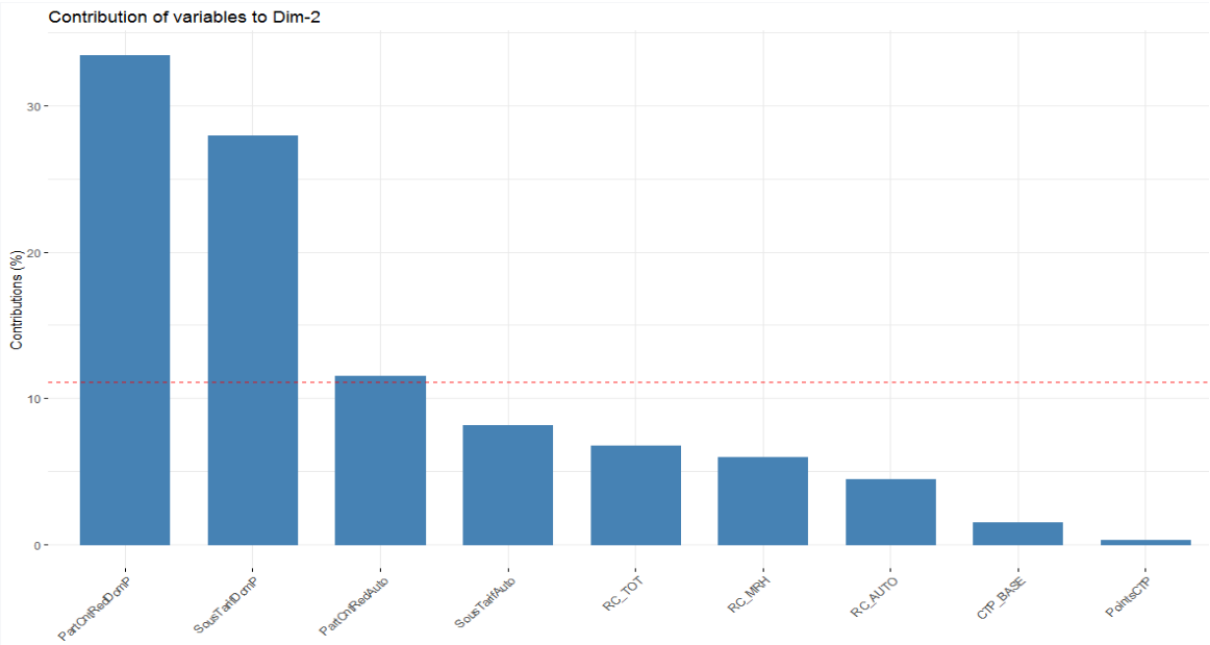

FIGURE B.3 – Contribution à la deuxième composante

En se basant sur les variables contribuant à la troisième composante et à leurs représentations sur le cercle de corrélation, il est difficile d'expliquer cette composante. Nous allons donc retenir uniquement les deux premières composantes pour notre étude. Ce qui représente 63% inertie expliquée. Ces deux composantes sont :

- Composante 1 : la rentabilité et la réduction (sous-tarification moyenne et part de contrats avec réduction) en auto
- Composante 2 : la réduction en dommages particuliers

B.2 ACP sur les indicateurs d'utilisation de budgets commerciaux

En utilisant les résultats de R notamment sur les valeurs propres de chaque composante, nous constatons que le premier plan permet d'expliquer que 35,55% de la variance totale. A partir du critère de Kaiser, nous allons nous focaliser sur les quatre premières composantes.

	Valeurs propres	Variances (%)	Variances cumulées(%)
Composante1	2.51	20.96	20.96
Composante2	1.75	14.59	35.55
Composante3	1.55	12.93	48.48
Composante4	1.30	10.88	59.36
...	...	...	...
Composante12	0.08	0.6	100

Table B.2 – Valeurs propres et variances

Interprétation des composantes :

Premiere composante :

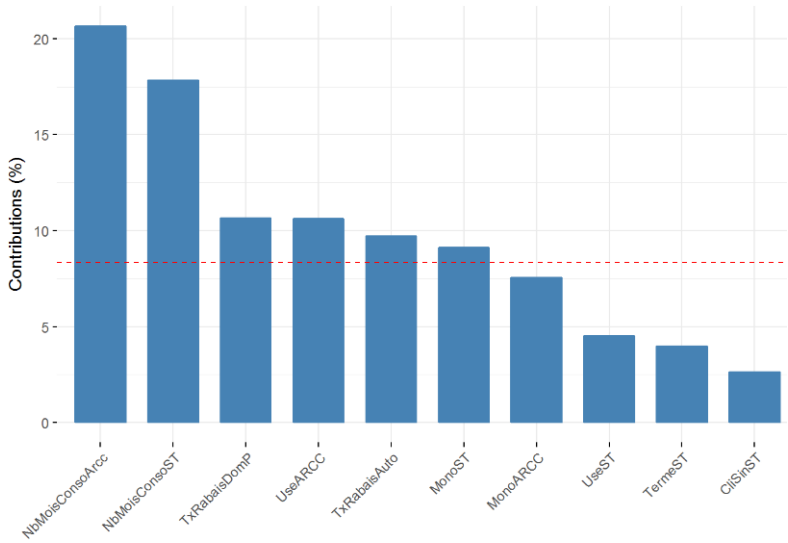

Figure B.4 – Contribution a la composante 1

Les variables NbreMois_ST et NbreMois_ARCC sont les deux qui influencent le plus cette composante et sont toutes deux corrélées positivement. Cette composante étant très dominée par les deux variables du nombre de mois de consommation en réserve, on peut conclure qu'il s'agit d'une composante qui exprime **la réserve en nombre de mois de consommation**

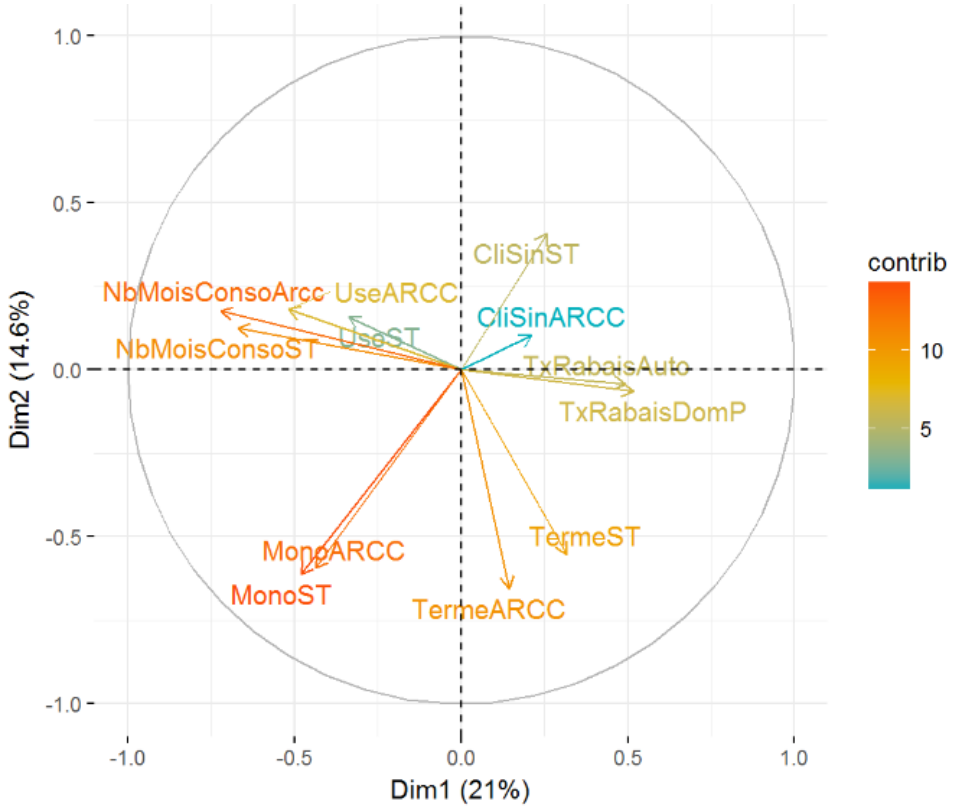

FIGURE B.5 – Cercle de corrélation du premier plan

Deuxième composante :

Cette composante est constituée grâce aux variables TermeST,TermeARCC (la part du budget utilisée pour le terme) et variables MonoST, MonoARCC (la part de budget allouée aux mono-détenteurs) , comme le montre le graphe B.6 de contribution à la composante. La matrice de corrélation du premier plan (voir figure B.5) montre que ces quatre variables sont bien représentées. De plus, on constate que les variables TermeST, TermeARCC, MonoST et MonoARCC ont un effet taille . Elle exprime **la répartition du budget**.

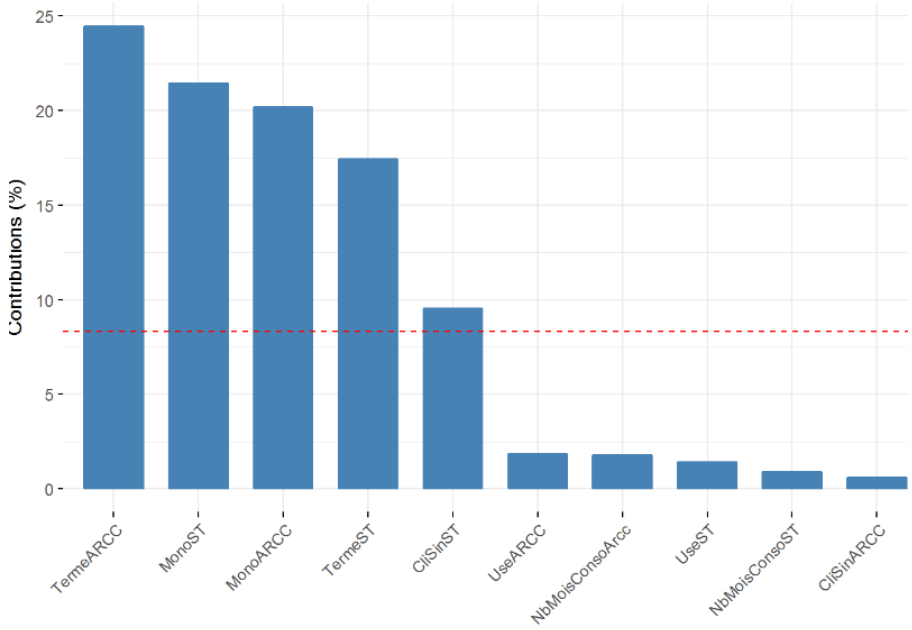

FIGURE B.6 – Contribution composante 2

Troisième composante :

Les variables TermeST et TermeARCC sont les deux variables qui contribuent le plus à cette composante. On peut en conclure qu'il s'agit d'une composante qui exprime **la part de budget utilisée pour le terme**

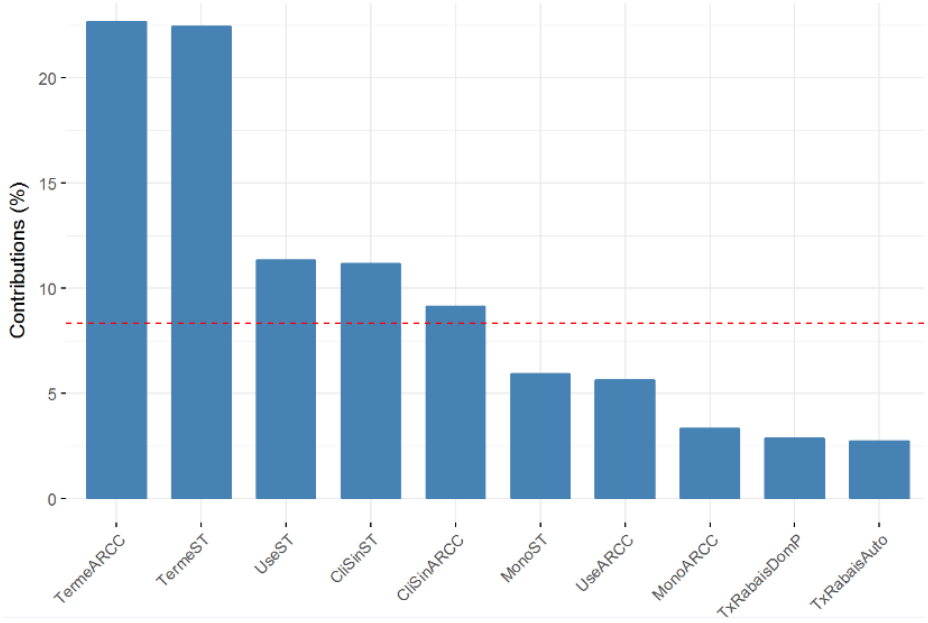

FIGURE B.7 – Contribution composante 3

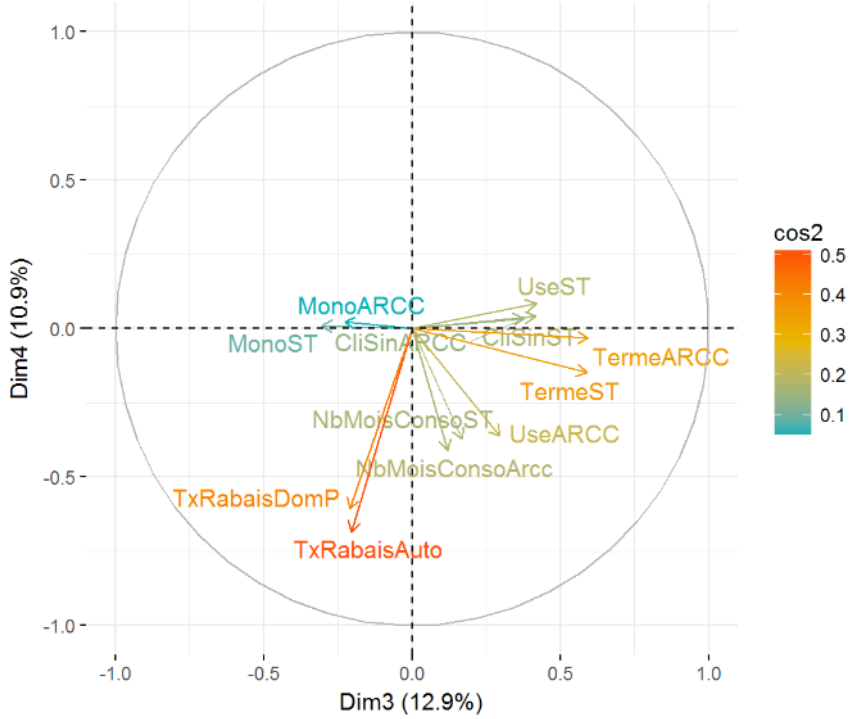

FIGURE B.8 – Cercle de corrélation du plan (composante 3 et composante 4)

Quatrième composante :

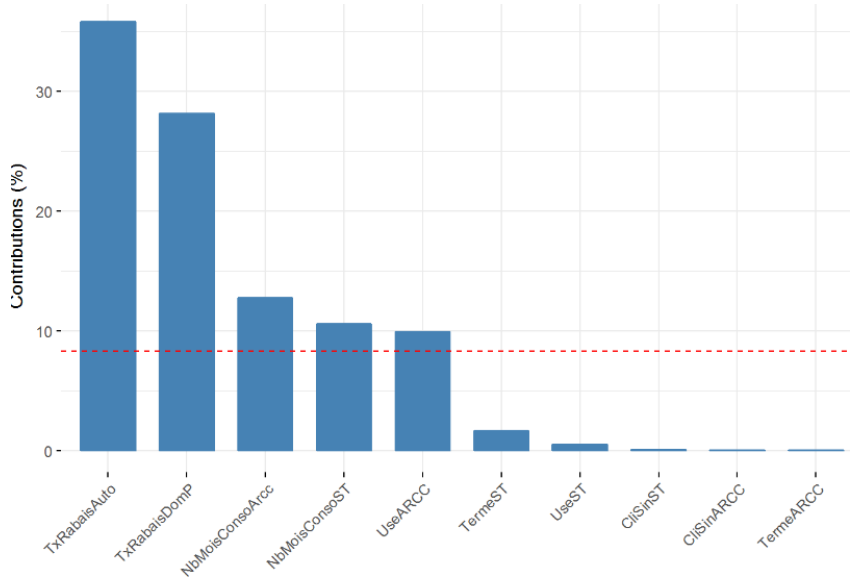

FIGURE B.9 – Contribution composante 4

En se basant sur les sorties de R, les variables TxRabais_ST et TxRabais_ARCC (taux

de rabais) sont les variables qui ont participé à la construction de la composante. Cette composante exprime **le taux de rabais**.

En somme, les quatre composantes seront retenues pour notre étude et ils expliquent 61.87% de la variance totale. Ces quatre composantes sont :

- Composante 1 : la réserve en nombre de mois de consommation
- Composante 2 : la répartition du budget
- Composante 3 : la part de budget utilisée pour le terme.
- Composante 4 : le taux de rabais

Annexe C

Modèles Linéaires Généralisés

Les modelés linéaires généralisés (Generalized Linear Model, GLM) sont utilisés pour modéliser une relation entre une variable à expliquer ou de réponse Y_i et les variables explicatives X_{ij}.

C.1 Hypothèses et principes des GLM

Pour modéliser la relation entre Y_i et les X_{ij} , les GLM s'appuient sur trois les conditions suivantes[Wik] :

- **Une composante aléatoire :** Cette composante permet de caractériser la distribution de la variable à expliquer. Il faut donc trouver une famille de distribution qui permet d'autoriser d'autres lois conditionnelles que la loi gaussienne. La famille exponentielle est donc utilisée pour caractériser la densité de la distribution de cette variable. On suppose donc que les observations sont des variables aléatoires $Y1.....$ Y_n indépendantes et de lois de probabilité appartenant à la famille exponentielle. Leur densité peut se mettre sous la forme :

$$f_{\alpha_i}(Y_i) = exp(\frac{\alpha_i Y_i - b(\alpha_i)}{a(\phi)} + c(Y_i, \phi)) \tag{C.1}$$

  Avec :
  — α_i est le paramètre canonique ou paramètre de la moyenne,
  — a est une fonction non nulle définie sur R,
  — b est une fonction définie sur R, deux fois dérivable
  — c est une fonction définie sur R x R, deux fois dérivable
  — ϕ est le paramètre de dispersion
  La famille exponentielle regroupe un grand nombre de lois à savoir la loi Normale, la loi Gamma, Log-Normale, Bernouilli :
  Cas de la loi de Bernouilli :
  Considérons une variable aléatoire Y qui suit la loi de Bernouilli de paramètres n et

p. La fonction de densité est alors :

$$P(Y = y) = p^y(1-p)^{1-y}$$

On peut le mettre sous la forme d'une fonction appartenant à la famille exponentielle :

$$P(Y = y) = exp^{(yln(p)+(1-y)ln((1-p))}$$

Avec cette diversité de lois qui appartiennent à la famille exponentielle, les GLM peuvent être utilisés pour résoudre un grand nombre de problèmes.

- **Une composante déterministe :** Cette composante permet de mettre en avant l'interaction entre les variables explicatives afin de savoir quels sont les prédicteurs. En considérant X la matrice d'observations et β un vecteur de p paramètre, la composante déterministe (prédicteur lineaire) s'exprime sous la forme d'une combinaison linéaire :

$$\eta(x_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \tag{C.2}$$

- **Le lien :** Il permet d'établir le lien entre la composante aléatoire et la composante déterministe. En effet, le but d'un GLM est d'établir une relation entre les variables explicatives et l'espérance de la variable à expliquer. Il faut donc trouver une fonction qui modélise une relation l'espérance mathématique de Y et η : il s'agit de la fonction de lien. Cette fonction permet d'établir la relation suivante :

$$g(E[Y_i]) = \eta(x_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \tag{C.3}$$

g est la fonction de lien et est fonction de la variable à expliquer Y : Dans la pratique,

Distribution de Y	**Fonctions de lien**	**Formule**
Binomiale	Logit	$g(x) = ln(\frac{x}{1-x})$
Poisson	Ln	$g(x) = ln(x)$
Normale	Probit	$g(x) = \Phi^{-1}(x)$
Gamma	Puissance	$g(x) = \frac{1}{x}$

Table C.1 – Exemples de fonctiosn de lien

les fonctions de lien d'identité sont utilisées lorsqu'il s'agit d'un modèle additif. En effet, pour une fonction de lien identité, on a :

$$E[Y_i] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Quant au modèle multiplicatif, c'est plutôt la fonction de lien logarithme qui est

privilégiée :

$$ln(E[Y_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Ce qui donne :

$$E[Y_i] = exp^{(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}$$

Par conséquent :

$$E[Y_i] = exp^{\beta_0} \prod_{j=1}^{p} exp^{\beta_j x_{ij}}$$

C.2 Estimation des paramètres

Dans la pratique c'est le maximum de vraisemblance qui est implémenté par la plupart des logiciels pour estimer les paramètres.Il s'agit d'un estimateur statistique qui est le produit des fonctions de densité. Dans notre cas, considérons toujours notre variable à expliquer Y qui suit une loi de la famille exponentielle et $Y_1....Y_n$ des observations indépendantes de cette variable.
Pour une fonction de lien g(μ_i)=θ_i avec θ qui dépend de β, la log vraisemblance du modèle est :

$$L(\beta) = \sum_{i=1}^{n} lnf(y_i, \theta_i, \phi) = \sum_{i=1}^{n} l(y_i, \theta_i, \phi)$$

$$L(\beta) = \sum_{i=1}^{n} \frac{\alpha_i y_i - b(\alpha_i)}{a(\phi)} + c(y_i, \phi) = \sum_{i=1}^{n} \ell_i$$

Pour obtenir le maximum, il faut calculer la dérivée $\frac{\partial \ell}{\partial \beta}$ et l'annuler :

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta} \frac{\partial \theta}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta i} \frac{\partial \eta i}{\partial \beta i}$$

On obtient alors pour chaque terme :

$$\begin{cases} \frac{\partial \ell_i}{\partial \theta} = & \frac{Y_i - b\prime(\theta)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)} \\ \frac{\partial \theta}{\partial \mu_i} = & \frac{\partial^2 b(\theta)}{\partial \theta^2} = Var[Y_i]/a(\phi) \\ \frac{\partial \mu_i}{\partial \eta i} = & \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} = (g^{-1})\prime(\eta_i) \\ \frac{\partial \eta i}{\partial \beta i} = & \frac{\partial X_i \beta}{\partial \beta_j} = x_{ij} \end{cases}$$

Les équations de vraisemblance donnent alors pour tout j=1....n :

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{Var[Y_i]} x_{ij} \frac{\partial \mu_i}{\partial \eta i} = 0$$

Ces équations ne sont pas linéaires et donc on ne peut pas les résoudre.La résolution requiert l'utilisation de méthodes itératives dans lesquels interviennent la matrice de Hessien.On peut donc mettre notre équation sous la forme d'une condition du premier ordre d'une régression linéaire pondérée :

$$\sum_{i=1}^{n} W_i \frac{\partial \eta i}{\partial \mu_i} \frac{y_i - \mu_i}{Var[Y_i]} x_{ij} = 0$$

Avec **W** la matrice diagonale de pondération :

$$W_i = \frac{1}{Var[Y_i]} (\frac{\partial \mu_i}{\partial \eta i})^2$$

En utilisant la methode de Newton Raphson, à partir de β_k, on peut obtenir β_{k+1}. Puis par itération successif, il est possible d'atteindre cette convergence :

$$\hat{\beta} \xrightarrow{P} \beta$$
$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{loi} \mathcal{N}(0, I(\beta)^{-1})$$

Avec I(β) la matrice d'informations de Fisher approximée par sa valeur au point $\hat{\beta}$. Ce qui nous ramène au cas classique de la loi normale dont on connaît la forme des intervalles de confiance. On a donc un intervalle de confiance au seuil α pour les coefficients β_j est :

$$[\widehat{\beta_j} - q_{\alpha/2}\sqrt{v_{jj}}; \widehat{\beta_j} + q_{\alpha/2}\sqrt{v_{jj}}]$$

Avec

- v_{jj} est j ème élément diagonal de la matrice I
- $q_{\alpha/2}$ est le quartile d'ordre 1 de la loi normale centrée réduite

Bibliographie

[Bac10] Alain Baccini. *Statistique Descriptive Multidimensionnelle*. Institut de Mathematiques de Toulouse. 2010 (cf. p. 21).

[Eur14] CGPA Europe. *Observatoire européen des intermédiares d'assurances*. `http://cgpa.fr/sites/default/files/cgpa_observatoire_vf_web_bat_protect.pdf`. 2014 (cf. p. 4).

[JHS12] Monsieur Jacques-Henri SUBLEMONTIER. « Classification non supervisée :de la multiplicité des données à la multiplicité des analyses ». Thèse de doct. Université d'Orléans, 2012 (cf. p. 37).

[NIE16] Marine NIEDZWIEDZ. *Rapport d'études Club ALGO*. Institut des actuaires. 2016 (cf. p. 53).

[Osc19] Torres-Reyna Oscar. *Logit, Probit and Multinomial Logit models in R*. `https://www.princeton.edu/~otorres/LogitR101.pdf`. 2019 (cf. p. 55).

[SAJO05] Mingoti Sueli A. et Lima Joab O. « Comparing SOM neural network with Fuzzy c-means,K-means and traditional hierarchical clustering algorithms ». In : *ScienceDirect* (2005) (cf. p. 37).

[St8] TUFFERY Stéphane. *DATA MINING et STATISTIQUE DECISIONNELLE*. TECHNIP, 2008 (cf. p. 37, 40).

[TRI13] Thi Diep TRIEU. *Une analyse exploratoire et une tarification spécifique aux seniors sur le portefeuille de Swiss Life dans le cadre du programme de recherche « Conséquence du vieillissement de la population sur la sinistralité des assurances »*. 2013 (cf. p. 22).

[Wik] WikiStat. *Introductions au modèle linéaire général*. `https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modlin-mlg.pdf` (cf. p. 87).

Table des figures

1.1 Étude comparative des canaux de distribution 5

1.2 Répartition CA IARD PP par canal de distribution 7

1.3 Utilisation de la ST 9

1.4 Utilisation de l'ARCC 9

1.5 Calcul du ratio combiné 10

1.6 CTP de base par rapport au CTP effectif 12

2.1 Matrice de corrélation 26

2.2 Contribution composante 1 28

2.3 Cercle de corrélation du premier plan 28

2.4 Contribution composante 2 29

2.5 Contribution composante 3 30

2.6 Cercle de corrélation du second plan (composante 3 et composante 4) . . . 30

2.7 Contribution composante 4 31

2.8 Construction de la base de classification 33

3.1 Inertie intra-classe vs inertie inter-classe 36

3.2 Classification k-means 37

3.3 Nombre de clusters optimal par la méthode de silhouette 39

3.4 Proportion d'inertie expliquée 40

3.5 Boxplot des taux de production auto par groupe 46

3.6 Boxplot de la sous-tarification moyenne par groupe 46

3.7 Boxplot des CA par groupe 47

3.8 Boxplot du croissance auto par groupe 47

4.1 Sélection des variables à l'aide de l'AIC 53

4.2 Exemple de validation croisée 54

4.3 Matrice de classification du GLM 54

4.4 Coefficients du GLM 55

4.5 Exemple d'arbre CART 57

4.6 Performance de l'arbre de décision 60

4.7 Arbre de décision de classification des agents 61

4.8 Matrice de classification des CART 63

4.9 Statistiques des groupes sur la dégradation de l'apport net 64

4.10 Matrice de confusion de classification avec des indicateurs non normalisés . 66

4.11 Matrice de confusion de classification avec des indicateurs normalisés . . . 67

4.12 AB Test : Variation des affaires nouvelles par groupe 69

B.1 Contribution à la composante 1 80

B.2 Cercle de corrélation du premier plan 80

B.3 Contribution à la deuxième composante 81

B.4 Contribution a la composante 1 82

B.5 Cercle de corrélation du premier plan 83

B.6 Contribution composante 2 84

B.7 Contribution composante 3 84

B.8 Cercle de corrélation du plan (composante 3 et composante 4) 85

B.9 Contribution composante 4 85

Liste des tableaux

2.1 Indicateurs de production 19

2.2 Indicateurs de maîtrise des coûts 20

2.3 Indicateurs d'utilisation des budgets commerciaux 21

2.4 Valeurs propres et variances 27

3.1 Groupes k-means 39

3.2 Test de Kruskal-Wallis Groupe 1 41

3.3 Test de Kruskal-Wallis Groupe 2 42

3.4 Test de Kruskal-Wallis Groupe 3 43

3.5 Test de Kruskal-Wallis Groupe 4 44

3.6 Test de Kruskal-Wallis Groupe 5 45

4.1 Comparaison GLM vs CART pour la caractérisation des groupes 64

B.1 Valeurs propres et variances 79

B.2 Valeurs propres et variances 82

C.1 Exemples de fonctiosn de lien 88