Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du
Diplôme d'Actuaire EURIA
et de l'admission à l'Institut des Actuaires

7 Septembre 2023

Par : MIROLO Annie
Titre : Identification de profils types en assurance emprunteur orientée sur la transformation de devis en contrats

Confidentialité : Oui Durée : 2 ans

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

Membre présent du jury de l'Institut des Actuaires :
DUBOIS David
QUERE Yann
STOCKSIEKER Samuel
Signature :

Entreprise :
GALEA & Associés
Signature :

Membres présents du jury de l'EURIA :
BAILLEUL Ismaël

Directeur de mémoire en entreprise :
BALLOT Charline
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 :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Remerciements

Avant de commencer la présentation de ce mémoire, j'adresse tout d'abord mes remerciements à l'ensemble des membres du cabinet GALEA & Associés, représenté par le Président, Monsieur Norbert GAUTRON, pour leur bienveillance et leur disponibilité.

Je remercie Monsieur Léonard FONTAINE, associé du cabinet, pour sa participation à la collecte des données essentielles à ce mémoire.

Je tiens à remercier en particulier ma responsable, Madame Charline BALLOT, pour sa franchise constructive, son soutien et son intention portée au bon déroulement du mémoire.

J'adresse également mes précieux remerciements au Datalab du cabinet, notamment à l'implication d'Alexandre EBY, Etienne RAYNAL et Irchad VALJEE MAMODE. Leurs contributions de compétences techniques ont été essentielles dans la réalisation de ce mémoire. Je remercie aussi Nina BOUCHE pour ses conseils d'experte en assurance emprunteur.

Enfin, je remercie ma tutrice académique Floriane PIVETEAU pour le temps accordé à la relecture du mémoire ; ainsi que Monsieur Franck VERMET, Directeur de l'EURIA, pour les enseignements liés à la data science en lien avec le métier d'Actuaire.

Je dois finalement mes remerciement à mon père, Laurent MIROLO, pour son accompagnement tout au long de mes études.

Résumé

L'assurance emprunteur est régulièrement assujettie à de nouvelles réglementations, en témoigne notamment la loi Lemoine qui permet de modifier plus facilement son assurance emprunteur. A cet effet, les comparateurs d'assurances en ligne sont susceptibles de faire face à une augmentation du nombre de devis simulés.

Ce contexte pourrait susciter l'intérêt des actuaires à analyser la tendance du marché devenue favorable à l'augmentation des devis qui se transformeraient en contrats. En identifiant ces profils types, les compagnies d'assurances pourraient par exemple mieux comprendre le comportement des nouveaux prospects et ajuster leurs produits en conséquence ; à même égard que les comparateurs d'assurances pourraient cibler leurs démarches commerciales envers les individus les moins risqués ou les plus susceptibles de souscrire un contrat d'assurance spécifique.

Actuellement, les méthodes d'apprentissage non supervisé permettent de regrouper les individus qui se ressemblent selon leurs caractéristiques. Toutefois, le groupe de risque auquel appartient l'individu n'est pas pris en compte, le risque étant ici la transformation ou non du devis en contrat. Or, la connaissance de cette information peut jouer un rôle important lors du regroupement d'individus.

L'étude présentée dans ce mémoire propose donc une approche alternative à la méthode classique qui consiste à intégrer l'information sur la connaissance préalable de la transformation du devis en contrat, et ainsi obtenir des profils types homogènes en caractéristiques et en risque.

Mots clés : assurance emprunteur - apprentissage non supervisé - profils types - transformation - devis - caractéristiques

Abstract

Mortgage insurance is regularly subject to new regulations, as evidenced by the Lemoine law, which allows for easier modification of one's mortgage insurance. As a result, online insurance comparators may potentially encounter an increase in the number of simulated quotes.

This context could spark the interest of actuaries to analyze the market trend that has become favorable for an increase in quotes that would convert into contracts. By identifying these typical profiles, insurance companies could, for instance, gain a better understanding of the behavior of new prospects and adjust their products accordingly. Similarly, insurance comparators could tailor their marketing efforts towards individuals with lower risk profiles or those most likely to subscribe to a specific insurance policy.

Currently, unsupervised learning methods enable the grouping of individuals with similar characteristics. However, the risk group to which the individual belongs is not considered, with risk being defined as the conversion or non-conversion of the quote into a contract. Yet, the knowledge of this information can play a significant role in the grouping of individuals.

The study presented in this thesis proposes an alternative approach to the conventional method, involving the incorporation of information regarding prior knowledge of the conversion of quotes into contracts. This allows for the creation of homogeneous profiles in terms of characteristics and risk.

Keywords : mortgage insurance - unsupervised learning - profile types - conversion - quotes - characteristics

Synthèse

Présentation du contexte

L'assurance emprunteur, secteur majeur de l'assurance de personnes, joue un rôle de protection pour l'assuré en cas d'imprévu tels que le décès, l'invalidité, l'incapacité ou la perte d'emploi en offrant une garantie d'indemnisation à l'organisme prêteur.

En France, plusieurs réformes ont émergé pour encourager la concurrence sur ce marché, permettant aux emprunteurs de trouver la meilleure offre correspondant à leur profil. La récente Loi Lemoine a notamment permis aux emprunteurs la résiliation à tout moment de leur contrat d'assurance. Cette évolution a engendré un environnement dans lequel les comparateurs d'assurances en ligne pourraient constater une hausse des simulations de devis.

Problématique

Face à la potentielle augmentation du nombre de devis, les comparateurs et compagnies d'assurances pourraient adopter une démarche proactive dans l'optique de cibler les comportements les plus propices à transformer leur devis en contrat.

En identifiant ces profils types, les comparateurs auront la possibilité d'améliorer la pertinence de leurs stratégies commerciales tout en optimisant l'utilisation de leurs ressources. De plus, cette démarche peut être avantageuse pour les compagnies d'assurances qui pourraient ajuster leurs produits en conséquence.

Par conséquent, l'idée est d'étudier les profils types les plus ou moins susceptibles de transformer leur devis en contrat afin de répondre à ces enjeux commerciaux.

Méthode d'apprentissage non supervisé

La classification ascendante hiérarchique, notée CAH, est une méthode d'apprentissage non supervisé qui va permettre de regrouper les devis ayant des caractéristiques proches, formant ainsi des *clusters*. Actuellement, cette méthode n'intègre pas la composante de la variable à expliquer Y, appelée ci-après la variable cible. Or, l'objectif du mémoire est de trouver les profils qui se ressemblent tant en termes de caractéristiques que de probabilité de transformer leur devis en contrat.

Le coeur de ce mémoire est donc la présentation de l'application actuarielle d'un algorithme, qui vise à pallier une contrainte inhérente à la CAH lorsqu'elle est appliquée à l'étude de la transformation des devis en contrats.

Présentation de l'approche alternative proposée en réponse à la problématique
Les approches mises en œuvre dans ce mémoire issues de la CAH classique, appelées CAH1 et CAH2, proposent de prendre en compte l'information de la variable Y par connaissance de la transformation du devis en contrat. Les deux approches sont les suivantes :

1. La CAH1 : ajout d'une seconde matrice de dissimilarité

Cette première approche peut être vue comme un compromis entre deux classifications ascendantes hiérarchiques obtenues par deux sources d'informations : d'une part les variables caractéristiques du devis et de la personne, et d'autre part la variable cible.

2. La CAH2 : pondération par l'importance des variables caractéristiques

Cette seconde approche basée sur l'importance des variables consiste à attribuer un poids représentatif du pouvoir explicatif des variables dans la prédiction de la variable cible. La méthode utilisée, *Permutation Feature Importance*, repose sur l'hypothèse que si une variable est considérée comme importante alors ses variations auront un impact significatif sur le modèle.

Application des méthodes aux données

Les trois approches suivantes sont appliquées aux données issues de la jointure entre la base de données des devis et des contrats d'un comparateur d'assurance :

— CAH0 : approche classique
— CAH1 : ajout d'une seconde matrice de dissimilarité
— CAH2 : pondération par l'importance des variables caractéristiques

En raison du nombre important de devis étudiés, certaines modalités doivent être regroupées pour permettre l'application des modèles sur R.

Regroupement des variables
Le regroupement des modalités s'effectue par trois approches indépendantes et en fonction de la nature des variables : par un simple regroupement des modalités, par une catégorisation des variables continues à l'aide d'un algorithme CART ou par un historique de taux d'emprunt.

Après ce regroupement, l'utilisation de la fonction nécessaire à l'implémentation des modèles est rendue possible en combinant l'agrégation des données et la construction de trois scénarios différents.

Sélection du *cluster* optimal
Dans l'objectif de comparer les trois approches entre elles, la recherche d'un nombre de *clusters* optimal est étudiée grâce à des indicateurs de qualité interne et externe tels que, respectivement, l'indice de Silhouette et l'indice de Rand.

Étude des *clusters*
Les *clusters* sont ensuite représentés visuellement par leur barycentre et leur inertie. Cette représentation permet d'avoir une vue d'ensemble des groupes formés et synthétise l'interprétation des résultats. Une analyse de leur densité est également présentée.

Comparaison avec une approche supervisée classique
Des prédictions sont également effectuées pour confronter les performances des algorithmes de regroupement. Ici, l'objectif n'est pas de prédire des libellés mais plutôt d'attribuer un nouveau devis dans un *cluster* en fonction de sa similitude avec les devis existants. À cette fin, l'algorithme des k plus proches voisins a été utilisé.

Profils types & conclusion

Les algorithmes de *clustering* présentés dans ce mémoire ont permis la formation de profils types homogènes en caractéristiques et en risque. Ils servent à distinguer et à catégoriser les

différentes situations qui se présentent lors de la transition d'un devis vers un contrat.

Deux profils types sont présentés : le premier représente l'un des profils types (parmi les huit obtenus) le plus susceptible de transformer son devis en contrat ; le deuxième représente l'un des profils types (parmi les quatre obtenus) le moins susceptible de transformer son devis en contrat.

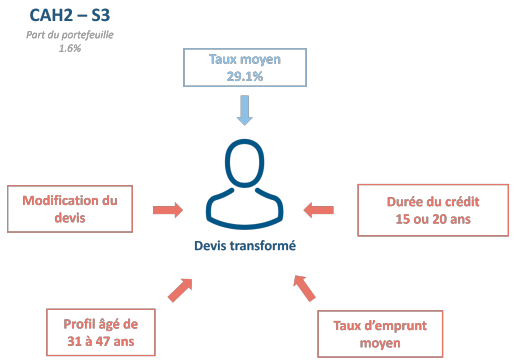

FIGURE 1 – Profil type d'un devis transformé - CAH2 Scénario 3 (*cluster 4*)

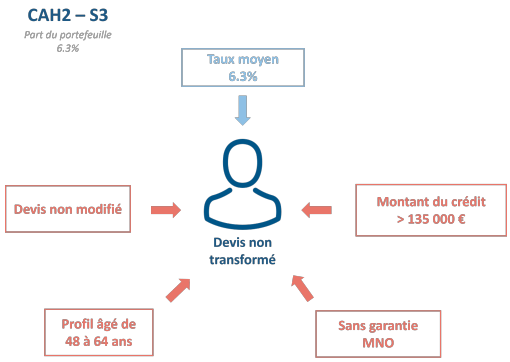

FIGURE 2 – Profil type d'un devis non transformé - CAH2 Scénario 3 (*cluster 2*)

En étudiant ces profils types, les acteurs de l'assurance emprunteur pourraient renforcer leur compréhension des comportements clients, ce qui leur permettrait d'ajuster leurs produits et leurs stratégies commerciales en conséquence.

A noter que la période d'observations des données, 2020 à 2023, est marquée par d'importants événements économiques et financiers et peut complexifier l'identification de certaines caractéristiques.

Limites

Deux limites évoquées ci-dessous auraient pu enrichir l'analyse des profils types obtenus et potentiellement conduire à des résultats différents :

Omission des caractéristiques du second emprunteur
L'analyse a pris en considération les attributs de l'emprunteur principal désigné alors que les contrats d'assurance emprunteur sont principalement souscrits de manière conjointe.

Absence de distinction entre nouvelle assurance et changement d'assurance
Les données disponibles ne permettent pas de distinguer les devis d'assurance qui sollicitent une couverture pour un nouveau prêt, de ceux qui envisagent de changer leur assurance actuelle.

Ouvertures

Ajout de la composante sinistralité : tarification et notion du risque
L'étude présentée dans ce mémoire pourrait être approfondie par une jointure supplémentaire avec une base de données sur les sinistres. En effet, l'ajout de cette donnée pourrait permettre de ressortir des profils plus ou moins risqués parmi ceux qui sont plus susceptibles de transformer leur devis en contrat. Cette analyse permettrait aux actuaires de tarifer au plus juste les produits proposés en fonction des profils.

Élargissement à d'autres secteurs de l'assurance
Les approches mises en oeuvre dans ce mémoire pourraient trouver leur application aussi bien en assurance emprunteur que pour des contrats Santé ou Prévoyance.

Synthesis

Context Introduction

Mortgage insurance, a major sector in personal insurance, serves as a protective measure for the insured in case of unexpected events such as death, disability, incapacity, or loss of employment, providing compensation assurance to the lending institution.

In France, several reforms have emerged to encourage competition in this market, enabling borrowers to find the best offer that matches their profile. The recent Lemoine Law, in particular, has allowed borrowers to terminate their insurance contracts at any time. This development has led to an environment in which online insurance comparison platforms could observe an increase in quote simulations.

Problem Statement

Faced with the potential increase in the number of quotes, both insurance comparison platforms and insurance companies could adopt a proactive approach aimed at targeting behaviors most likely to convert their quotes into contracts.

By identifying these typical profiles, comparison platforms will have the opportunity to enhance the relevance of their commercial strategies while optimizing the utilization of their resources. Furthermore, this approach can be advantageous for insurance companies, allowing them to tailor their products accordingly.

Therefore, the idea is to study the profiles that are more or less likely to convert their quotes into contracts in order to address these business challenges.

Unsupervised learning method

Hierarchical Ascendant Classification, hereinafter referred to as HAC, is an unsupervised learning method that facilitates the grouping of quotes with similar characteristics, thereby forming clusters. Currently, this method does not incorporate the explanatory variable Y, referred to here as the target variable. However, the objective of the thesis is to identify profiles that resemble each other both in terms of characteristics and in terms of the likelihood of converting their quotes into contracts.

The core of this thesis lies in the presentation of the actuarial application of an algorithm, which aims to address a limitation inherent to HAC when applied to the study of the conversion of quotes into contracts.

Introduction of the alternative approach proposed in response to the problem statement
The approaches implemented in this thesis, derived from classical HAC and referred to as HAC1 and HAC2, aim to incorporate the information from variable Y through knowledge of the conversion of quotes into contracts. The two approaches are as follows :

1. HAC1 : Addition of a second dissimilarity matrix

This first approach can be seen as a compromise between two hierarchical ascendant classifications obtained from two sources of information : on one hand, the characteristic variables of the quote and the individual, and on the other hand, the target variable.

2. HAC2 : Weighting based on the importance of characteristic variables

This second approach, based on variable importance, involves assigning a weight that represents the explanatory power of variables in predicting the target variable. The method employed, Permutation Feature Importance, is built on the assumption that if a variable is deemed important, then its variations will have a significant impact on the model.

Application of the methods to the data

The following three approaches are applied to the data resulting from the merging of the quotes and contracts database from an insurance comparison platform :

— HAC0 : Classic approach
— HAC1 : Addition of a second dissimilarity matrix
— HAC2 : Weighting based on the importance of characteristic variables

Due to the significant number of quotes studied, certain categories must be grouped together to facilitate the application of models in R.

Variable grouping
The grouping of categories is done through three independent approaches, based on the nature of the variables : through a straightforward consolidation of categories, by categorizing continuous variables using a CART algorithm, or by utilizing historical borrowing rates.

Following this grouping, the use of the function required for implementing the models is made possible by combining data aggregation and the construction of three different scenarios.

Selection of the optimal cluster
In order to compare the three approaches, the search for an optimal number of clusters is investigated using internal and external quality indicators such as the Silhouette index and the Rand index, respectively.

Study of the clusters
The clusters are then visually represented by their centroids and inertia. This representation provides an overview of the formed groups and synthesizes the interpretation of the results. An analysis of their density is also presented.

Comparison with a classic supervised approach
Predictions are also made to compare the performance of the clustering algorithms. In this context, the goal is not to predict labels but rather to assign a new quote to a cluster based on its similarity to existing quotes. For this purpose, the k-nearest neighbors algorithm has been utilized.

Typical Profiles & Conclusion

The clustering algorithms presented in this thesis have facilitated the creation of homogeneous profiles in terms of characteristics and risk. They are employed to differentiate and categorize the various scenarios encountered during the transition from a quote to a contract.

Two typical profiles are presented : the first represents one of the (among the eight obtained) profiles most likely to convert their quotes into contracts ; the second represents one of the (among the four obtained) profiles least likely to convert their quotes into contracts.

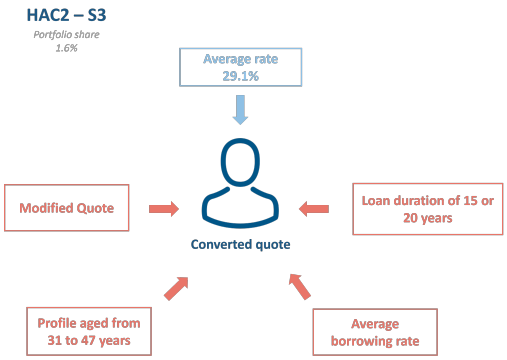

FIGURE 3 – Typical Profile of a Converted Quote - HAC2 Scenario 3 (*cluster 4*)

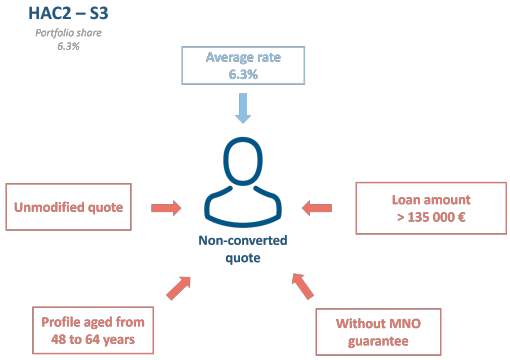

FIGURE 4 – Typical Profile of a Non-Converted Quote - HAC2 Scenario 3 (*cluster 2*)

By studying these typical profiles, stakeholders in mortgage insurance could enhance their understanding of customer behaviors, enabling them to adapt their products and business strategies accordingly.

It is noteworthy that the period of data observation, from 2020 to 2023, is marked by significant economic and financial events, which could complicate the identification of certain characteristics.

Limitations

Two limitations mentioned below could have enriched the analysis of the obtained typical profiles and potentially led to different outcomes :

Omission of Second Borrower's Characteristics
The analysis considered attributes of the designated primary borrower, even though mortgage insurance contracts are primarily subscribed jointly.

Lack of Differentiation Between New and Changed Insurance
The available data does not allow for differentiation between insurance quotes seeking coverage for a new loan versus those considering a change in their current insurance.

Future Directions

Incorporation of Claims Data : Pricing and Risk Assessment
The study presented in this thesis could be further enhanced by an additional integration with a claims database. Incorporating this data could enable the identification of profiles with varying levels of risk among those more likely to convert their quotes into contracts. This analysis would provide actuaries with the ability to price the offered products more accurately based on these profiles.

Expansion to Other Insurance Sectors
The approaches implemented in this thesis could be applicable not only to mortgage insurance but also to Health or Life Insurance contracts.

Table des matières

Introduction

L'assurance emprunteur occupe une place dominante dans le secteur de l'assurance de personnes, offrant une protection financière aux emprunteurs en cas de décès, d'incapacité, d'invalidité ou de perte d'emploi. Sur le marché français, elle a généré en 2019 un chiffre d'affaires de 10 milliards d'euros dont 70%[1] étaient spécifiquement dédiés à la couverture des prêts immobiliers.

En 2022 avec l'émergence de la loi Lemoine, trois changements majeurs sont intervenus dans le paysage de l'assurance emprunteur notamment la possibilité de pouvoir résilier à tout moment son assurance pour se tourner vers une autre offre, sous certaines conditions. Dans ce contexte, les compagnies et les comparateurs d'assurances doivent rester vigilants quant à la compétitivité des offres proposées sur le marché. En effet, cette liberté de pouvoir modifier son assurance emprunteur, à tout moment, peut amener le nombre de personnes qui rechercherait de nouvelles offres plus avantageuses et adaptées à leurs besoins à augmenter. Ainsi, la concurrence sur ce marché pourrait s'accroître et inciter les acteurs de l'assurance à proposer des produits plus attractifs.

L'assouplissement de la modification de l'assurance emprunteur permis par la Loi générera probablement une augmentation des devis simulés pour les comparateurs d'assurances qui agissent en tant que courtiers. Leur vocation principale consiste à dénicher pour leurs clients la meilleure assurance.

Dans l'optique de canaliser les devis qui montrent une tendance élevée à se convertir en contrats, un intérêt stratégique est mis en avant : en identifiant les devis qui manifestent une forte probabilité de transformation, les courtiers et les comparateurs d'assurances peuvent concentrer leurs efforts en ciblant les opportunités de conversion les plus prometteuses. Cette démarche proactive viserait non seulement à maximiser les ressources des courtiers et des comparateurs d'assurances et permettrait aussi aux compagnies d'assurance d'adapter leurs produits en conséquence ; au même titre que les comparateurs qui commercialiseraient leurs propres produits.

Face à l'opportunité d'identifier des profils types les plus susceptibles de transformer leur devis en contrat, des méthodes de *data science* en apprentissage statistique non supervisé peuvent être utilisées. Elles permettent de regrouper des individus possédants des caractéristiques communes entre eux, formant ainsi des *clusters*. Toutefois, ces algorithmes traditionnels présentent leur limite et ne permettent pas de prendre en compte la connaissance préalable de la transformation du devis en contrat. Or, cette information est importante car elle permettrait de segmenter les prospects de manière plus précise tout en répondant à la problématique.

C'est pourquoi, l'approche proposée dans ce mémoire consiste à intégrer cette information préalablement connue, permettant ainsi de trouver les profils qui se ressemblent tant en termes de caractéristiques que de probabilité de transformer leur devis en contrat. En analysant dans quelles mesures ces modèles alternatifs peuvent améliorer la qualité des regroupements, les actuaires et les compagnies d'assurance pourront adapter leurs produits, cibler leurs offres d'assurance emprunteur et optimiser leurs stratégies commerciales.

1. Source : Assemblée Nationale N-4699, Rapport 2021, p7

Ce mémoire vise à explorer comment de nouvelles approches en apprentissage statistique peuvent être utilisées dans le cadre d'analyses actuarielles en assurance emprunteur. Le secteur de l'assurance emprunteur et son marché actuel sont présentés en première partie. Ensuite, il sera rappelé de manière théorique les méthodes actuelles de l'apprentissage statistique permettant de rebondir sur les nouvelles approches utilisées. Enfin, l'application de ces algorithmes est mise en œuvre sur la base de données dans un objectif de sélectionner les profils types des devis qui se transforment le plus souvent et le moins souvent en contrat.

Première partie

Présentation de l'assurance emprunteur

Chapitre 1

La souscription d'une assurance emprunteur

Ce chapitre introduit les fondamentaux de l'assurance emprunteur, puis son évolution au travers des grandes réformes la menant à son marché actuel.

L'assurance emprunteur est "un contrat d'assurance qui va garantir le remboursement du capital restant dû d'un prêt, ou de ses échéances, à l'établissement prêteur, en cas de survenance de certains évènements comme le décès, l'invalidité, l'incapacité ou la perte d'emploi de l'assuré."[1]

L'assureur intervient auprès de la banque en cas de décès, d'arrêt de travail lié à une maladie ou une perte d'emploi en remboursant la totalité ou une partie du capital emprunté par l'emprunteur. Généralement, les individus qui souscrivent à ce type d'assurance ont pour objectif d'acquérir un bien immobilier ou de contracter un prêt à la consommation. L'assurance emprunteur joue donc un rôle de protection pour l'assuré en cas d'imprévu en offrant une garantie d'indemnisation à la banque.

L'assurance emprunteur n'est pas obligatoire d'après la loi. Toutefois, elle est demandée par les organismes prêteurs d'argent pour la majorité des prêts car en pratique aucune banque ne va accepter d'accorder un prêt à un client sans être couvert par une assurance.

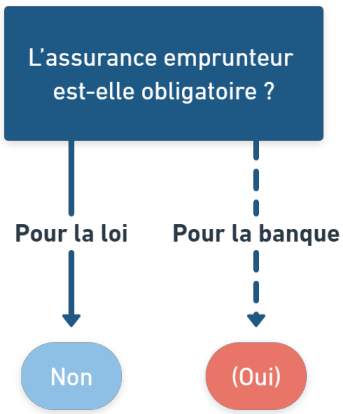

FIGURE 1.1 – L'assurance emprunteur obligatoire ?

1. Définition de la Banque de France : Assurance emprunteur

Cependant, il existe des exceptions, d'après le *graphique présenté ci-dessous 1.2*, montrant que la part des emprunteurs couverts par un contrat d'assurance emprunteur est inférieure à 100%, sur un historique des années 2010 à 2021. Ceci soulève qu'aucun contrat n'est souscrit pour certains emprunts.

Une explication plausible provient de l'application du taux d'usure. Le taux d'usure est le taux d'intérêt maximum légal que la banque peut appliquer au crédit de l'emprunteur, y compris le taux de l'assurance. Le respect de ce taux peut rapidement devenir une limite pour le client, en raison de l'augmentation générale des taux plus rapide que celle du taux d'usure.

Ainsi, sous certaines conditions, les banques peuvent accorder des prêts à des emprunteurs sans être couverts par une assurance. Par exemple, la banque peut appliquer un nantissement ou une hypothèque sur les biens des clients possédant déjà un grand nombre de propriétés.

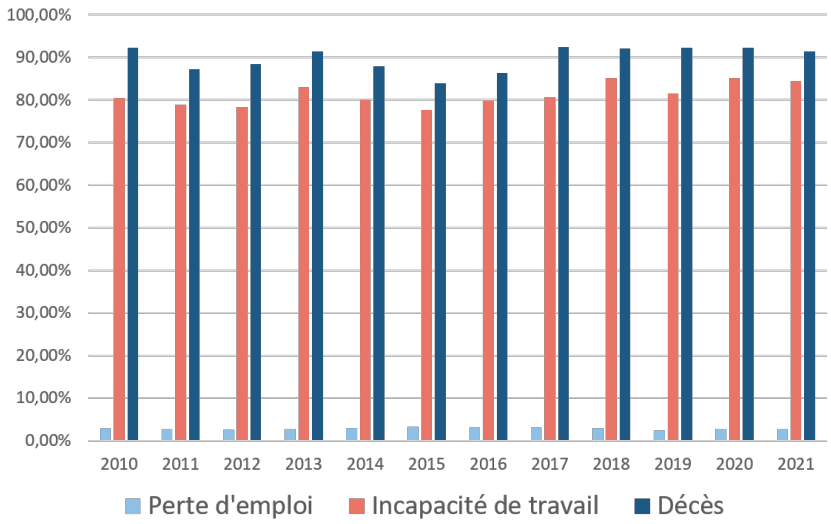

FIGURE 1.2 – Part des emprunteurs couverts par un contrat d'assurance *(Source : Enquête annuelle sur le crédit à l'habitat 2021, Banque de France, dernière mise à jour 19.01.2023)*

À partir de ce graphique, il est identifié que la garantie décès est présente dans 9 cas sur 10, à l'opposé de la garantie perte d'emploi qui représente moins de 5% de la part des emprunts. Ceci est notamment dû à son coût élevé.

1 Les acteurs

En assurance emprunteur, il existe différents acteurs intervenant dans la mise en place du contrat d'assurance. Chacun d'entre eux joue un rôle bien précis.

L'organisme prêteur
L'organisme prêteur est l'institution financière qui accorde le prêt à l'emprunteur. Il peut notamment s'agir d'une banque mais aussi de tout autre établissement financier habilité à fournir des prêts. Son rôle est de prêter de l'argent à un individu qui souhaite souscrire un prêt. En échange de ce prêt, l'organisme prêteur va faire payer des intérêts à l'emprunteur.

L'intermédiaire d'assurance
L'intermédiaire d'assurance est l'acteur qui joue le rôle d'apporteur d'affaires entre les emprunteurs et les compagnies d'assurance. Il s'agit généralement des courtiers, des agents généraux ou même des comparateurs d'assurances en ligne. L'intermédiaire d'assurance perçoit des commissions de la part de l'assureur en contrepartie de son apport d'affaires.

L'assuré
L'assuré est la personne physique ou morale porteuse du risque qui est couverte par le contrat d'assurance. Il s'agit de l'emprunteur.

L'assureur
L'assureur représente la compagnie d'assurance qui propose, produit et émet le contrat d'assurance emprunteur. Ce contrat est destiné à couvrir les risques auprès de l'organisme prêteur en cas de non-remboursement du prêt par l'emprunteur. Son rôle est de fournir une protection financière à l'emprunteur en cas de survenance d'un sinistre lié au décès, à l'incapacité, à l'invalidité ou à la perte d'emploi, et également à l'organisme prêteur d'être couvert dans cette situation.

Le souscripteur
Le souscripteur est la personne physique ou morale qui souscrit au contrat d'assurance. Son rôle est d'initier le processus de souscription du contrat d'assurance en le signant et en s'engageant à payer les primes correspondantes. La plupart du temps, le souscripteur est l'emprunteur, c'est-à-dire la personne qui contracte le prêt et qui bénéficie de la protection par l'assureur.

Le bénéficiaire
Le bénéficiaire est la personne physique ou morale destinée à recevoir les indemnités prévues par le contrat d'assurance en cas de survenance d'un sinistre pour l'assuré. Pour l'assurance emprunteur, le bénéficiaire est l'organisme prêteur, très souvent la banque.

2 Le vocabulaire associé

Après avoir introduit les différents acteurs interagissant dans la mise en place du contrat d'assurance, voici d'autres définitions utiles :

La délégation d'assurance
La délégation d'assurance se définit comme la possibilité pour l'emprunteur de choisir librement une assurance auprès d'un organisme autre que celui proposé par l'organisme prêteur. Il peut se tourner vers une compagnie d'assurance de son choix.

Néanmoins, cette délégation d'assurance est encadrée par la réglementation. La nouvelle offre d'assurance choisie par l'assuré doit proposer des garanties équivalentes ou supérieures à celles fixées par l'établissement prêteur.

La quotité
Par définition, la quotité exprime en pourcentage la part du prêt couvert par l'assurance en cas de sinistre. Lorsqu'un emprunt est réalisé seul, la quotité doit être de 100%. Cela signifie que la totalité du prêt est couverte par l'assurance. Dans le cadre d'une co-assurance, la somme minimale de la quotité des deux souscripteurs doit être de 100%, et peut aller jusqu'à 200%.

La quotité a un impact sur les primes : plus la quotité est élevée, plus les primes sont élevées puisque l'assureur prend en charge un risque plus important. A l'inverse, une quotité plus faible entraîne des primes plus basses.

La fiche standardisée d'information (FSI)
La fiche standardisée d'information est un document qui présente de manière claire et concise les principales caractéristiques d'un contrat d'assurance emprunteur. Elle vise à faciliter la comparaison des offres d'assurances émises par différentes compagnies pour permettre aux emprunteurs de prendre des décisions éclairées. Cette fiche présente a minima les caractéristiques suivantes :

— Les garanties (obligatoires et optionnelles)
— Les exclusions

— Le délai de carence (période pendant laquelle les garanties ne s'appliquent pas)
— Le délai de franchise (période à l'issue de laquelle interviendra la prise en charge du sinistre)[2]
— La quotité
— Le coût de l'assurance

Les assureurs ont l'obligation de remettre cette fiche à l'emprunteur lors de la présentation d'une offre. Cette fiche sert notamment à comparer des offres à celle exigée par la banque, dans le cas où l'emprunteur souhaite opter pour la délégation d'assurance.

Le questionnaire médical
L'emprunteur doit remplir un questionnaire médical permettant à l'assureur de collecter des informations sur son état de santé. Ce questionnaire vise donc à évaluer le risque porté par l'assureur vis-à-vis de son assuré. Il comprend des questions du type : les antécédents de santé sur les 10 dernières années, l'Indice de Masse Corporelle (IMC), la tension artérielle, le statut de travailleur handicapé, le cas échéant.

À la suite de ce questionnaire l'assureur peut appliquer une surprime si l'emprunteur est jugé à risque. Il peut également refuser de couvrir l'emprunteur si le risque est jugé trop important. Il est à noter qu'en cas de fausses déclarations l'assureur pourra refuser d'indemniser l'emprunteur lors d'un sinistre.

Cependant, depuis la mise en place de la Loi Lemoine en 2022 (*détaillée au paragraphe : 1.2*), ce questionnaire est supprimé sous certaines conditions d'emprunts.

3 Les contrats

Maintenant que les acteurs et le vocabulaire spécifique sont définis, les contrats d'assurance peuvent être présentés. Sur le marché actuel, il existe deux grandes catégories de contrat en assurance emprunteur : le contrat individuel et le contrat de groupe.

Le contrat de groupe
Les contrats de groupe sont souscrits auprès de la banque commerciale entre l'emprunteur et l'organisme prêteur. Ils sont standardisés par des grilles préétablies par les assureurs pour regrouper des emprunteurs qui se ressemblent. A noter que ces contrats restent moins segmentés que les contrats individuels mais permettent de proposer un contrat adapté à l'emprunteur. Les contrats de groupe opèrent selon le principe de la mutualisation des risques : les primes versées par les emprunteurs sont toutes réunies dans un "pot commun" et constituent une réserve pour régler les sinistres qui surviennent.

Le contrat de groupe réunit trois acteurs : l'assureur (assureur partenaire de la banque), le souscripteur (l'organisme prêteur : la banque) et l'adhérent (l'emprunteur). Le tarif est communiqué par les assureurs auprès de la banque et respecte des grilles tarifaires. Cela permet notamment à la banque de jouer le rôle d'apporteur d'affaires. C'est la banque qui prendra les décisions si une dérogation devait avoir lieu.

L'adhésion au contrat groupe est facultative, l'emprunteur peut choisir librement de s'orienter vers un contrat individuel.

Le contrat individuel
Les contrats individuels sont souscrits directement auprès d'un assureur et les échanges se font de manière indépendante de la banque, soit directement entre l'assuré et l'assureur. Il est aussi possible de passer par l'intermédiaire d'un courtier en assurance. Les contrats individuels offrent

2. Exemple : pour la garantie ITT, si la franchise est de 60 jours, l'indemnité ne commencera à être versée qu'à compter du 61ème jour.

une plus grande flexibilité et un ajustement sur-mesure pour l'assuré. En effet, ils lui permettent de choisir les garanties et le niveau de couverture correspondant au mieux à ses besoins et à son profil de risque. Il en découle une tarification segmentée car les tarifs s'adaptent à chaque profil de risque. Ce contrat est donc avantageux pour un emprunteur considéré comme "bon risque" car il pourrait payer sa prime moins chère qu'avec un contrat groupe.

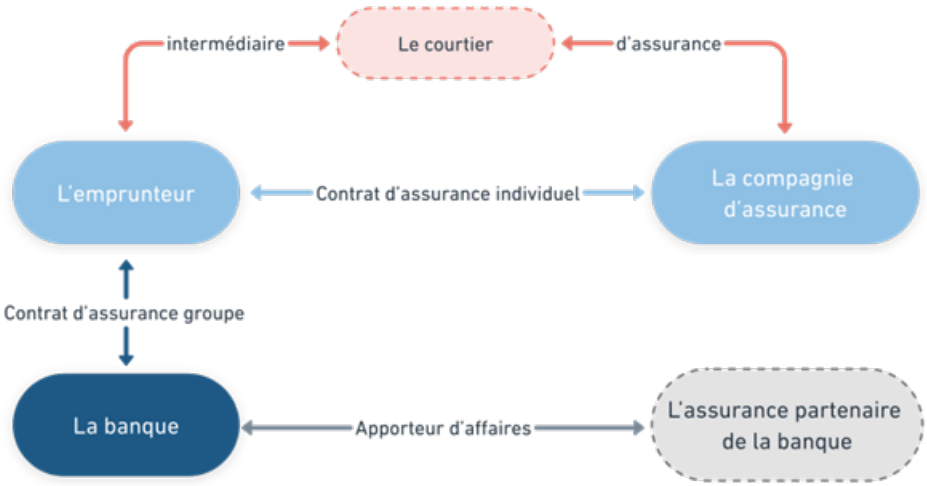

FIGURE 1.3 – Acteurs et liens d'un contrat groupe et individuel

4 Les garanties

Ces contrats d'assurance proposent différentes garanties qui visent à protéger les emprunteurs en cas de situations telles que le décès, l'incapacité, l'invalidité ou la perte d'emploi. Ces situations peuvent être protégées par des garanties comme présentées ci-dessous.

Les garanties se définissent comme l'engagement pris par l'assureur vis-à-vis de l'assuré. Elles sont définies dans le contrat d'assurance et présentent les indemnisations perçues par le bénéficiaire lors d'un sinistre incombant l'emprunteur. Pour que le risque soit assurable, il doit vérifier certaines conditions et doit être :

— aléatoire
— licite
— futur
— réel
— quantifiable (pour calculer sa probabilité d'apparition)

Pour présenter les garanties, il est ici définit les deux termes suivants :

1. Les garanties "primaires", présentes dans la majorité des contrats. Elles comprennent la garantie couvrant le décès ainsi que la garantie couvrant la perte totale et irréversible de la perte d'autonomie. Ces garanties ne sont pas obligatoires mais sont très souvent exigées par les banques lors de la souscription d'un emprunt.
2. Les garanties "secondaires", souvent exigées pour des prêts immobiliers. Elles comprennent toutes les autres garanties : l'incapacité, l'invalidité et la perte d'emploi.

Le choix d'inclure ou non certaines garanties est laissé à la discrétion de l'organisme prêteur responsable de la souscription du contrat.

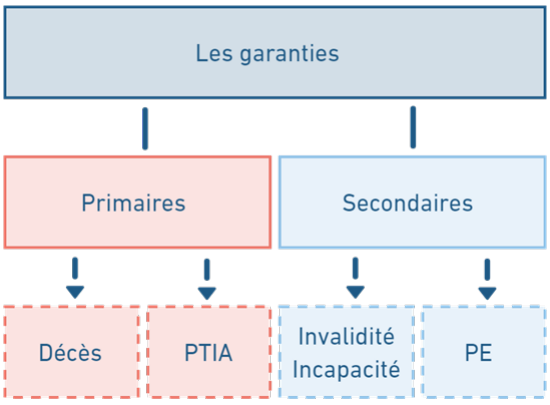

FIGURE 1.4 – Les garanties en assurance emprunteur

4.1 Les garanties primaires

Deux garanties sont majoritairement incluses dans les contrats d'assurance emprunteur, quel que soit le profil de l'emprunteur ou le type de prêt.

La garantie décès
La garantie décès est incluse dans la plupart des contrats (à l'exception de certains dossiers comme précédemment évoqué dans la figure 1.2). Cette garantie assure le remboursement du capital restant dû de l'emprunteur, en cas de décès avant l'âge limite fixé dans les conditions contractuelles.

La garantie Perte Totale et Irréversible d'Autonomie (PTIA)
La garantie PTIA, également incluse dans la plupart des contrats, prend en charge les mensualités de remboursement du prêt en cas d'invalidité lorsque celle-ci est de niveau 3 sur l'échelle de la Sécurité Sociale.[3]

Les contrats prévoient généralement des cas d'exclusions pouvant être :
— Liés à une activité physique dangereuse (parachute, équitation, sports de combat)
— Liés à un métier à risque (pompier, militaire, policier...)
— Liés à l'âge (limite d'âge pouvant être fixée de 60 à 70 ans en moyenne)
— Liés à des évènements extérieurs (faits de guerres, émeutes, actes de terrorisme, suicide...)

Ces deux garanties prévoient la plupart du temps un délai de carence pouvant aller de 1 à 12 mois.

3. La qualification d'une invalidité de niveau 3 signifie que l'emprunteur ne peut réaliser seul, c'est-à-dire sans l'aide d'un tiers, 3 sur 4 des Actes de la Vie Quotidienne (AVQ) suivants : se nourrir, se laver, se vêtir/se dévêtir et se déplacer.

4.2 Les garanties secondaires

De plus, si l'emprunteur contracte un crédit immobilier dans le but d'acheter une résidence principale ou secondaire, il pourra être amené à souscrire les garanties supplémentaires suivantes, selon l'exigence de l'organisme prêteur :

Les garanties d'invalidité et d'incapacité

— La garantie Incapacité Temporaire de Travail (ITT) : désigne l'impossibilité totale d'exercer son travail à la suite d'un accident ou d'une maladie. A la différence de la garantie PTIA, la situation d'invalidité est provisoire et non permanente. Cette période d'invalidité correspond à une période de convalescence pendant laquelle l'organisme d'assurance prend en charge les mensualités de remboursement du prêt bancaire, et ce jusqu'à la reprise d'activité du souscripteur.

— La garantie Invalidité Permanente Totale (IPT) : désigne l'impossibilité d'exercer son travail dans les mêmes conditions qu'avant la survenance d'un accident ou d'une maladie. La garantie s'applique seulement lorsque les conditions de retour au travail après la consolidation du sinistre reconnaissent un taux d'invalidité supérieur ou égal à 66%. La prise en charge d'indemnisation par l'assureur se fait selon le même principe que la garantie ITT.

— La garantie Invalidité Permanente Partielle (IPP) : fonctionne selon le même principe que la garantie précédente IPT pour un taux d'invalidité compris entre 33% et 66%.

— La garantie Invalidité Permanente Professionnelle (IP PRO) : fonctionne selon le même principe que la garantie précédente IPT pour les professions médicales, paramédicales et vétérinaires. Cette garantie spécifique a la particularité d'accorder des montants de pourcentages de taux d'invalidité plus élevés car certaines capacités humaines sont spécifiques à la réalisation du métier et peuvent affecter considérablement l'exercice de la profession.

Les garanties invalidité et incapacité des contrats d'assurance ne couvrent pas les maladies du dos et les affections psychologiques et psychiatriques. Or, certains organismes prêteurs exigent que ces options soient assurées. La garantie suivante permet ainsi la couverture de ce risque.

La garantie Maladies Non Objectivables (MNO)[4]
La garantie MNO est une garantie complémentaire souscrite à la signature du contrat d'assurance. Elle dépend des déclarations de l'assuré en indiquant son état de santé général et notamment la présence de douleurs telles que le mal de dos et/ou un suivi psychologique. Ces pathologies peuvent être à l'origine d'un arrêt de travail.

La garantie Perte d'Emploi (PE)
La garantie perte d'emploi permet de prendre en charge le remboursement par l'assureur des annuités de remboursement du prêt lorsque l'assuré est au chômage. Pour ce faire, l'assuré doit être en mesure de prouver sa perte d'emploi ainsi que ses indemnités de chômage de Pôle Emploi. Cette garantie facultative est très peu utilisée dans le cadre de l'assurance emprunteur car son coût est élevé. De plus, l'assureur applique souvent une franchise de 3 mois avant l'indemnisation de la perte d'emploi et une durée maximale de remboursement en moyenne fixée à 18 mois. Le nombre de mensualités total de remboursement, en cas de récurrence de perte d'emploi, peut également être limité.

4. Une maladie non objectivable est une maladie non "mesurable" et non "quantifiable" par un médecin, reposant presque exclusivement sur les déclarations du patient.

5 Les types de prêts

Comme évoqué précédemment, les garanties présentes dans les contrats peuvent varier selon le type de prêt choisi par l'emprunteur.

A chaque besoin personnel de financement, il existe un prêt adapté, ce qui induit l'existence de différents types de prêts détaillés ci-dessous. Les principaux sont les prêts immobiliers représentant la plus grande part du marché, ensuite les prêts à la consommation et enfin les prêts professionnels.

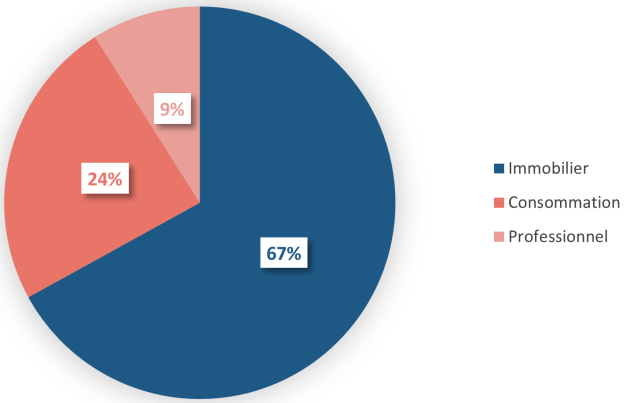

FIGURE 1.5 – Répartition des cotisations par type de prêt *(source : "L'assurance Française données clés 2021", FFA, septembre 2022, p49)*

Les prêts immobiliers
Les prêts immobiliers sont souscrits notamment dans le cas :

— De l'achat d'une résidence principale
— De l'achat d'une résidence secondaire
— D'un l'investissement locatif

Chaque emprunt immobilier donne lieu à un tableau d'amortissement du capital dans lequel figure : les périodes, le capital restant dû (CRD), les intérêts, l'amortissement du capital et les annuités (égales à la somme du remboursement partiel du capital et des intérêts de la période).

Période	Annuités	Intérêts	Amortissements	CRD début de période

FIGURE 1.6 – Tableau d'amortissement d'un prêt immobilier

Il existe cinq grands types de prêts immobiliers qui diffèrent de par leur mode de remboursement :

1. Le remboursement in fine

Pour un remboursement in fine, l'emprunteur s'engage à rembourser la totalité du capital initial en un seul versement à l'échéance du prêt à la différence du paiement des intérêts qui s'effectuera à la fin de chaque période. Ce prêt est qualifié de prêt non amortissable car le capital n'est pas remboursé via les mensualités mais en une seule fois à l'échéance.

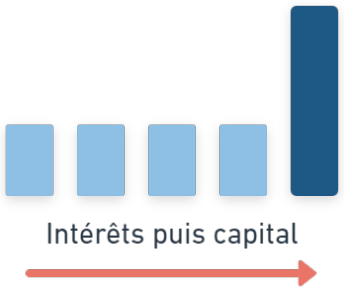

FIGURE 1.7 – Les flux du remboursement par amortissement constant

Ce prêt concerne principalement les investisseurs disposant déjà d'un apport suffisant pour soutenir le remboursement du prêt par l'épargne dont ils disposent. Il ne représente donc qu'une faible partie de la population. La durée maximale de ce type de prêt est de 15 ans.

2. Le remboursement par annuités constantes

Quand l'emprunteur choisit un prêt avec remboursement à annuités constantes, il s'engage à rembourser chaque année une somme fixe représentant une partie d'intérêts et une partie du capital. Ce montant d'annuité constante est fixe dans le temps et déterminé à partir du capital emprunté C.

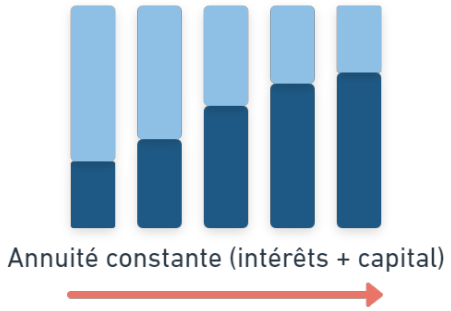

FIGURE 1.8 – Les flux du remboursement par annuité constante

Par la démonstration figurant en annexe (1), le montant de l'annuité pour chaque période est déterminé par :

$$a = C\frac{i}{((1-(1+i)^{-T})}$$

Avec :
— i l'intérêt composé du prêt
— C le capital initial emprunté
— T la durée totale de l'emprunt, en période
— a le montant de l'annuité constante d'une période

3. Le remboursement par amortissement constant du capital

Quand l'emprunteur choisit un prêt avec remboursement constant du capital, il s'engage à rembourser le capital emprunté à hauteur d'une somme constante de $\frac{C}{T}$ à chaque période. Le calcul des intérêts porte sur le capital restant dû. Or, au fur et à mesure du remboursement du prêt le capital restant dû diminue, ce qui fait également diminuer le montant des intérêts.

Ainsi, les annuités payées en fin de chaque période prennent la forme d'une progression arithmétique de raison $\frac{-iC}{T}$ car elles sont en diminution à chaque période de la part d'intérêts

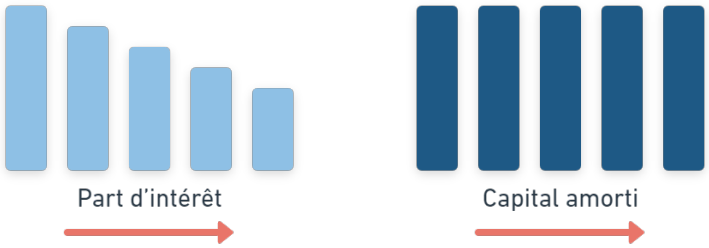

FIGURE 1.9 – Les flux du remboursement par amortissement constant

associée au montant $\frac{C}{T}$. La valeur de la somme de cette suite arithmétique représente la somme totale des intérêts versés (en valeur absolue) et vaut :

$$(\text{nombre de terme} + 1)\,\frac{\text{premier terme} + \text{dernier terme}}{2}$$

Soit

$$(T+1)\,\frac{\frac{iC}{T} + (T-1)\frac{iC}{T}}{2} = \frac{iC(T+1)}{2}$$

4. Remboursement avec prêt relais

Ce type de prêt concerne les emprunteurs déjà propriétaires d'un bien immobilier qui sera mis en vente ou qui est déjà mis en vente. Ce bien ne sera pas vendu avant l'achat d'une autre résidence. L'organisme prêteur accorde donc un prêt à l'emprunteur, lui permettant d'acquérir un nouveau logement en attendant la vente de sa propriété actuelle. C'est un prêt de courte durée (maximum 2 ans).

5. Prêt à taux zéro

Le prêt à taux zéro ne comporte aucun intérêt et l'emprunteur est seulement soumis au remboursement du capital initial C. C'est l'Etat qui prend en charge le remboursement des intérêts. Ce prêt s'applique dans des conditions particulières pour l'achat d'un logement neuf ou à réhabiliter.

Les prêts à la consommation
Les prêts à la consommation sont les prêts accordés aux particuliers pour financer leurs dépenses personnelles de consommation. Les besoins sont généralement des besoins pour l'achat d'une voiture, le financement de vacances, le paiement de frais médicaux ou l'aménagement d'une maison. Le financement des études par un prêt étudiant fait aussi partie de cette catégorie de prêt. Le montant de ces prêts est compris entre 200€ et 75 000€ avec une période de remboursement a minima de 3 mois. En générale, la durée maximale accordée est de 7 ans à 10 ans.

Les prêts professionnels
Les prêts professionnels sont à destination des entreprises et des entrepreneurs pour répondre à leurs besoins de financement liés à leur activité professionnelle. Ce type de prêt peut notamment aider : au démarrage d'une entreprise, à l'achat de matériel pour rénovation des bureaux, à l'acquisition de biens immobiliers, au recrutement de personnel . . .

6 La tarification

La prime d'assurance peut se tarifer selon trois modes de calcul et s'ajoute aux mensualités de remboursement du prêt.

Les trois modes d'expressions rencontrés sur le marché sont :

— En Capital Initial (CI) : le montant de la prime est fixe pendant toute la durée de vie du prêt et est exprimée en pourcentage du capital initial

— En Capital Restant Dû (CRD) : le montant de la prime diminue progressivement chaque année pendant la durée de vie du prêt et est exprimé en pourcentage du capital restant dû

— Selon un taux variable exprimé en pourcentage du CRD : le montant de la prime suit l'évolution théorique du risque de l'assuré. Par exemple l'évolution du risque en fonction de l'âge et du capital restant dû de l'assuré.

Les primes d'assurance correspondent aux cotisations que l'emprunteur paie périodiquement à la banque ou à la compagnie d'assurance pour bénéficier en contrepartie de la couverture de l'assurance. La prime d'assurance emprunteur peut varier selon plusieurs facteurs : l'âge, le capital emprunté, la durée de remboursement, la profession, le tabagisme et le questionnaire de santé.

En plus de ces primes d'assurance, l'emprunteur doit aussi rémunérer au taux d'intérêt l'organisme prêteur (hormis le cas du prêt à taux zéro). Il existe :

— Le taux d'intérêt simple, qui concerne essentiellement les opérations inférieures à 1 an et qui est peu utilisé en général. Il se calcule toujours à partir du Capital Initial (CI) emprunté et est proportionnel à ce capital. Il se comporte donc comme un taux proportionnel.
— Le taux d'intérêt composé, calculé à la fin de chaque période (mensuelle en général) pour être ajouté au capital restant dû. Par définition, il est décrit comme l'intérêt capitalisé car il produit à son tour des intérêts pour les périodes futures. Il se comporte donc comme un taux actuariel.

Chapitre 2

Le contexte de l'assurance emprunteur

Après avoir abordé les principales caractéristiques et définitions de l'assurance emprunteur, il est pertinent d'examiner les évolutions réglementaires, le contexte économique du marché ainsi que la notion de comparateurs.

1 Les réglementations de l'assurance emprunteur

En France, l'assurance emprunteur a connu plusieurs réformes permettant de promouvoir la concurrence et d'offrir aux emprunteurs un plus grand choix quant à la possibilité de comparer les offres du marché. Cela permet aux emprunteurs de trouver la meilleure offre correspondant à leur profil.

Jusqu'en 2010, l'emprunteur était obligé de souscrire l'assurance proposée par la banque octroyant le prêt sous peine de voir sa demande de prêt refusée. Depuis cette date, plusieurs lois ont permis d'assouplir les conditions de changement de contrat en assurance emprunteur. Voici les réglementations les plus marquantes ayant impacté le marché de l'assurance emprunteur :

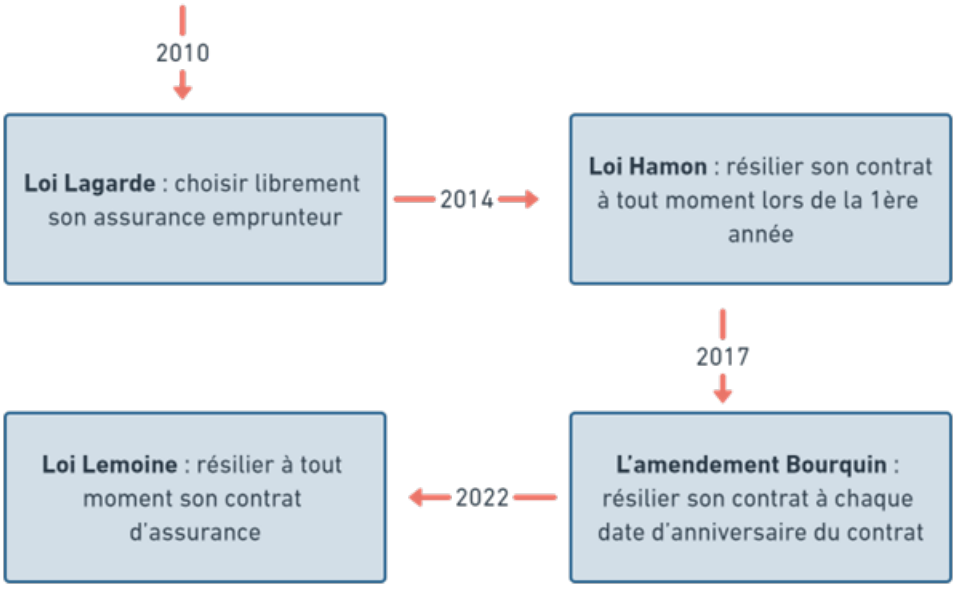

FIGURE 2.1 – Les grandes dates de l'assurance emprunteur

1.1 L'historique des réglementations

Loi Lagarde 2010

La Loi Lagarde de 2010 a permis aux emprunteurs de choisir librement leur assurance emprun-

teur. Ils peuvent alors faire appel à :

— Des compagnies d'assurance (assureur, mutuelle, institution de prévoyance)
— D'autres organismes financiers
— Des courtiers en assurance

Cette loi a permis la délégation d'assurance. Toutefois, le contrat d'assurance emprunteur proposé doit comporter des garanties équivalentes à celles de l'assurance proposée par l'organisme prêteur.

Loi Hamon 2014
"La loi Hamon est une loi relative à la consommation. Elle vise à renforcer les droits et la protection des consommateurs face aux vendeurs en leur permettant un gain en pouvoir d'achat."[1]

La Loi Hamon permet aux emprunteurs de résilier à tout moment leur contrat emprunteur dans les 12 mois suivant la signature du contrat. Auparavant, les contrats d'assurance emprunteur étaient souscrits pour la durée de vie du contrat sans la possibilité de résilier.

Loi Sapin 2 : l'amendement Bourquin 2017
La Loi Sapin 2 ne s'applique pas directement à l'assurance emprunteur mais vise à renforcer la régulation et la surveillance du secteur financier. Son application a introduit des mesures visant à renforcer les pouvoirs de l'Autorité de Contrôle Prudentiel et de Résolution (ACPR), l'organisme de surveillance des assurances en France. La finalité de cette nouvelle mesure est de renforcer la protection des assurés pour la stabilité financière des marchés.

Cette loi est également connue sous le nom de l'Amendement Bourquin qui a le même objectif que les lois précédentes : faciliter le changement d'assurance emprunteur. Cet amendement offre ainsi la possibilité à l'assuré de résilier son contrat à chaque date anniversaire, en respectant un niveau de garanties au moins égal à celui du contrat en vigueur.

1.2 La nouveauté 2022 : loi Lemoine

Toutes les réglementations précédentes ont permis d'aider l'emprunteur à changer d'assurance mais aucune n'a réellement permis de libéraliser le marché de l'assurance emprunteur pour les prêts immobiliers. Ce marché est toujours monopolisé par les banques assurances.

FIGURE 2.2 – Les nouveautés de la Loi Lemoine

1. Le guide de l'assurance de prêt : meilleurtaux.com

La loi Lemoine a pour but de rendre l'assurance emprunteur plus accessible en réduisant le délai du droit à l'oubli, en supprimant le questionnaire de santé (sous certaines conditions) et en favorisant la concurrence par la baisse des prix, grâce à la possibilité de résilier le contrat en dehors de l'échéance annuelle.

Voici les trois grandes nouveautés de la Loi Lemoine entrées en vigueur en mars 2022 :

1. La résiliation à tout moment du contrat d'assurance souscrit par l'emprunteur. Avant la Loi Lemoine, le changement de contrat d'assurance pouvait se faire : soit lors de la première année suivant la signature du contrat (Loi Hamon), soit une fois par an à la date anniversaire du contrat (amendement Bourquin).

2. La suppression du questionnaire de santé. Le questionnaire de santé permet d'estimer le risque que représente l'assuré pour l'organisme assureur. Dorénavant, s'il remplit les conditions suivantes, l'assuré n'est plus dans l'obligation de remplir ce questionnaire :
   — montant du capital inférieur à 200 000 €
   — échéance du prêt antérieure aux 60 ans de l'assuré

3. Le droit à l'oubli. Avant la loi Lemoine, le droit à l'oubli concernait que les cancers. Les règles étaient les suivantes :
   — Être diagnostiqué avant 21 ans et être en rémission depuis plus de 5 ans ;
   — Être diagnostiqué après 21 ans et être en rémission depuis plus de 10 ans.
   Avec la loi Lemoine, le droit à l'oubli s'est étendu à l'hépatite C. Les règles sont désormais les suivantes :
   — Être diagnostiqué avant 21 ans et être en rémission depuis plus de 5 ans ;
   — Être diagnostiqué après 21 ans et être en rémission depuis plus de 5 ans.
   Le modification apportée concerne donc les emprunteurs diagnostiqués après 21 ans et dont la rémission date de 5 à 10 ans.

Ces grandes nouveautés ont déjà entraîné des répercussions visibles sur le marché de l'assurance emprunteur. Puisque l'assurance devient davantage accessible aux personnes à risque, le risque pour l'assureur peut augmenter. En contrepartie, certains assureurs ont ajusté leurs tarifs afin d'anticiper une augmentation de la fréquence des sinistres. Cette situation affecte principalement les contrats d'assurance individuels, tandis que les contrats collectifs bénéficient encore de la mutualisation des risques et du niveau de marge appliqué.

2 Le marché actuel de l'assurance emprunteur et ses comparateurs

FIGURE 2.3 – Chiffres clés 2021 et 2022

Actuellement, le marché de l'assurance emprunteur est largement dominé par les banques qui détiennent 87%[2] des parts du marché. Malgré la crise du Covid, l'assurance emprunteur à

2. Chiffres clés 2022 : Assurance prêt immobilier : le marché bousculé par la loi Lemoine 2022, MAGNOLIA, novembre 2022

su maintenir sa croissance avec une augmentation de +4,6% du total des cotisations entre 2020 et 2021.[3]

L'ouverture à la concurrence faisant suite aux différentes réglementations attire de nouveaux assureurs alternatifs principalement positionnés sur les contrats individuels. En plus des assureurs historiques comme Axa ou Generali, des acteurs jusqu'à présent moins importants distribuent maintenant plus de contrats : ce sont les mutuelles d'assurance, en confère Malakoff Humanis (*tableau 2.5*).

Rang	Organisme	Cotisations Brutes de réassurance (en Md€)	Variation 2021/2020
1	CNP Assurances	2 694	1,7%
2	Crédit agricoles Assurances	2 284	9,5%
3	Groupe des Assurances du Crédit mutuel	1 615	5,8%
4	Assurance du groupe BPCE	904	12,7%
5	BNP Paribas Cardif	833	0,4%

FIGURE 2.4 – *Source : "Classement de l'assurance emprunteur 2022 : les bancassureurs chiffres 2021", par l'Argus de l'assurance, septembre 2022 (argusdelassurance.com)*

Rang	Organisme	Cotisations Brutes de réassurance (en Md€)	Variation 2021/2020
1	AXA France	678	3,7%
2	Generali	297	6,8%
3	Allianz	165	-11,0%
4	MetLife	114	2,3%
5	**Malakoff Humanis**	**102**	**53,9%**

FIGURE 2.5 – *Source : "Classement de l'assurance emprunteur 2022 : les alternatifs chiffres 2021", par l'Argus de l'assurance, septembre 2022 (argusdelassurance.com)*

Cette concurrence sur le marché de l'assurance emprunteur a notamment permis aux emprunteurs de réaliser des économies sur le coût total du prêt en sélectionnant l'offre la plus avantageuse en terme de prix. Les comparateurs d'assurance emprunteur en ligne peuvent être utiles pour faciliter la comparaison des offres.
Par ailleurs, l'application de la Loi Lemoine dont l'objectif principal est de favoriser la libéralisation du marché de l'assurance emprunteur devrait faire augmenter les flux sur les plateformes de comparateurs en ligne.

3. Chiffres clés 2021 "L'assurance Française données clés 2021", FFA, septembre 2022, p49

Deuxième partie

Présentation des méthodes actuelles de l'apprentissage statistique

Cette partie présente les bases théoriques des algorithmes d'apprentissage statistique utilisés dans la suite de ce mémoire.

L'apprentissage statistique, communément appelé *Machine Learning*, est une branche de l'intelligence artificielle fondée sur des approches mathématiques et statistiques. Elles permettent d'exploiter une grande quantité de données à travers des algorithmes génériques avec une phase d'apprentissage qui permet à l'algorithme d'apprendre à résoudre un problème spécifique sans être explicitement programmé pour.

L'apprentissage statistique est utilisé dans le domaine de l'actuariat pour analyser des données complexes, modéliser des risques et prendre des décisions éclairées. Les résultats issus de ces algorithmes peuvent être utilisés pour réaliser des prédictions, comprendre et anticiper les comportements des assurés. Par exemple, l'apprentissage statistique peut être un outil d'aide à la gestion des risques assurantiels, tels que : la détection de la fraude, les prévisions de mortalité ou de survie, la gestion des sinistres, la tarification des contrats...

Deux grands types d'apprentissage statistique existent et sont utilisés dans les travaux qui font l'objet de ce mémoire ; l'apprentissage supervisé et l'apprentissage non-supervisé.

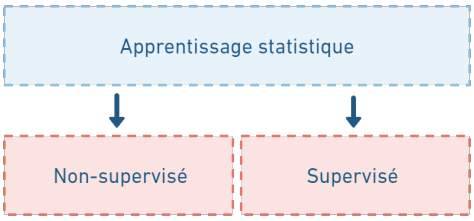

FIGURE 2.6 – Algorithmes d'apprentissage statistique (liste non-exhaustive) utilisés dans le cadre de ce mémoire

— Sur un ensemble de données composées de paires d'entrées (caractéristiques) et de sorties attendues (étiquettes), **l'apprentissage supervisé** consiste à apprendre la relation entre les données en entrée et en sortie. Les caractéristiques représentent les informations sur lesquelles le modèle doit se baser pour prendre une décision, tandis que les étiquettes sont les réponses correctes correspondant à ces caractéristiques. L'apprentissage supervisé est couramment utilisé dans des tâches telles que la classification (assigner des catégories à des données), la régression (prédire des valeurs numériques), et divers autres problèmes où les sorties attendues sont connues.
De manière formelle, soit $Y = (Y_1, \ldots, Y_m)$ le vecteur des variables de sorties, on parle aussi de variables cibles, d'étiquettes, de *labels*, m est la dimension des variables cibles. Soit $X^\top = (X_1, \ldots, X_p)$ le vecteur des caractéristiques, des prédicteurs, des *features*, p est la dimension des caractéristiques. $x_i^\top = (x_{i1}, \ldots, x_{ip})$ correspond au vecteur des caractéristiques pour la ligne i et y_i une valeur parmi Y. Le modèle apprend sur les couples $(x_1, y_1), \ldots (x_N, y_N)$, N est le nombre de données. L'apprentissage supervisé peut être formellement décrit comme un problème d'estimation de densité où il s'agit de déterminer les propriétés de la densité conditionnelle $Pr(Y|X)$.

— **L'apprentissage non-supervisé** consiste en l'organisation d'individus en groupes homogènes au sein d'une population. Le nombre de groupes et leur nature ne sont au préalable pas définis. L'objectif de l'apprentissage non-supervisé est de découvrir des structures ou des regroupements d'individus similaires dans le jeu de données. En apprentissage on se donne un jeu de données de taille N, d'observations $(x_1, \ldots, x_N)$ d'un $p-$vecteur X ayant pour densité $Pr(X)$. L'objectif de l'apprentissage non supervisé est d'inférer de manière automatique et non guidée les propriétés de $Pr(X)$.

L'apprentissage non-supervisé est d'abord présenté en raison de son utilisation centrale dans le mémoire.

Chapitre 3

Le *clustering* : apprentissage non-supervisé

Le *clustering* est une méthode d'apprentissage non-supervisé qui consiste à regrouper les individus qui ont des caractéristiques proches autour d'un même centre. Ce centre est l'individu représentatif du *cluster* et représente son profil type. En conséquence, les points de données d'un groupe en particulier présentent des propriétés similaires. Dans le même temps, les points de données de différents groupes ont des caractéristiques différentes.

Dans le cadre de ce mémoire, deux algorithmes connus et très utilisés parmi la famille des algorithmes de *clustering* sont présentés :

1. **La classification ascendante hiérarchique**, appellation réduite à "CAH", se base sur une mesure de similarité entre les individus pour créer une hiérarchie de regroupement. Cette méthode est dite hiérarchique car elle produit des groupes de plus en plus vastes incluant des sous-groupes en leur sein. Cette méthode est aussi qualifiée "d'ascendante", car au départ, tous les individus sont seuls dans une classe pour être ensuite rassemblés par groupes de plus en plus nombreux.

2. **Le partitionnement en K-moyenne**, appelé *K-means*, regroupe les individus en un nombre K de *clusters*. Cette méthode a pour finalité de diviser l'ensemble des individus dans les K groupes renseignés, selon un critère de division.

1 La classification ascendante hiérarchique

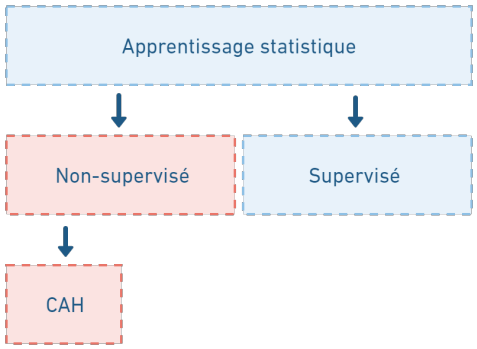

FIGURE 3.1 – Apprentissage statistique non-supervisé : CAH

1.1 Principe de la CAH

La classification ascendante hiérarchique est une méthode d'apprentissage non-supervisé utilisée en analyse de données. Son objectif principal est de regrouper les individus les plus proches deux à deux pour former des *clusters*.

Cet algorithme regroupe progressivement les individus en classes de plus en plus grandes. C'est un procédé dit itératif ou agglomératif qui se termine lorsque tous les individus ont été regroupés dans une classe spécifique. Ces regroupements successifs peuvent être représentés visuellement par un dendrogramme. Le dendrogramme est le graphique propre à la représentation des individus issus du résultat d'une classification hiérarchique. Il se présente souvent comme un arbre binaire dont les feuilles sont les individus alignés sur l'axe des abscisses. Des traits verticaux partant de l'abscisse de deux groupes ou de deux individus sont dessinés lorsqu'ils se rejoignent. La longueur de chaque trait vertical correspond à la distance séparant les individus ou groupes d'individus regroupés. Puis ces deux groupes sont reliés par un segment horizontal qui est d'autant plus grand que les groupes sont éloignés.

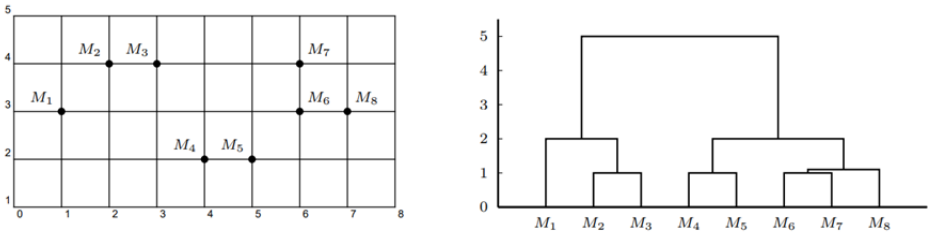

FIGURE 3.2 – Représentation du dendrogramme à partir de la matrice des distances

D'après cette figure, la branche gauche du dendrogramme est construite ainsi : les individus M_2 et M_3 sont initialement regroupés en raison de leur proximité, de la même manière que les individus M_4, M_5 et M_5, M_6, M_6. Puis, ce groupe (composé des deux individus M_2 et M_3) est regroupé avec l'individu le plus proche, M_1.

1.2 Paramètres de la CAH

La CAH classique se définit à partir de deux paramètres :

1. Une mesure de dissimilarité (ou de distance)
2. Une méthode de regroupement (ou dissimilarité inter-classe)

Pour décider quels *clusters* doivent être agglomérés, il est nécessaire de disposer à la fois d'une mesure de dissimilarité inter-classe permettant le regroupement, qui correspond à un critère de liaison (*linkage criterion*), et d'une mesure de dissimilarité entre les points d'observations qui correspond à la distance appropriée au type de données. Le critère de liaison détermine la distance entre les ensembles d'observations en fonction des distances entre les observations elles-mêmes.

Le choix de la distance et du critère de liaison peut avoir un impact majeur sur le résultat de la classification.

1.2.1 La mesure de dissimilarité

Le premier paramètre à définir pour réaliser une classification ascendante hiérarchique est la matrice de dissimilarité. Elle indique la ressemblance, deux à deux, entre tous les individus décrits par le jeu de données. Pour regrouper les individus qui se ressemblent (et séparer ceux qui ne se ressemblent pas), il faut un "critère de ressemblance". Pour cela, l'algorithme a

besoin des distances entre tous les individus. Cette distance est calculée à partir des variables caractéristiques X. Elle dépend du type de variable à étudier.

Distance entre variables quantitatives

Pour les variables uniquement quantitatives, en notant x_i un point de l'espace représentant un individu du jeu de données pour $i \in [1, N]$, N le nombre d'individus total, et p le nombre de variables caractéristiques ; la ressemblance entre deux individus $x_{1,j}$ et $x_{2,j}$, avec $j \in [1, p]$, peut être calculée à partir de distances mathématiques, notées $d(x_1, x_2)$, telles que (liste non-exhaustive) :

— La distance euclidienne : $\sqrt{\sum_{j=1}^{p}(x_{1,j} - x_{2,j})^2}$

— La distance maximale : $\underset{j\in[1,p]}{max}|x_{1,j} - x_{2,j}|$

— La distance de Manhattan : $\sum_{j=1}^{p}|x_{1,j} - x_{2,j}|$

— La distance de Canberra : $\sum_{j=1}^{p}\frac{|x_{1,j}-x_{2,j}|}{|x_{1,j}|+|x_{2,j}|}$

Distance entre variables quantitatives et qualitatives

Pour un mélange de variables qualitatives et quantitatives, ou uniquement de variables qualitatives, la distance de Gower est utilisée [4]. Elle se construit par l'intermédiaire d'un indice de similarité noté S_g. Cet indice consiste à mesurer la ressemblance entre deux individus décrits par des variables qualitatives et quantitatives. L'indice de similarité varie entre 0 et 1. Il vaut 1 pour des individus identiques et 0 pour des individus considérés sans aucun point commun. L'indice de similarité de Gower entre deux individus x_1 et x_2 se calcule de la manière suivante :

$$S_g(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p} s_{12,j} \tag{3.1}$$

Avec :

— p le nombre total de variables qualitatives et quantitatives
— $s_{12,j}$ représente la similarité partielle entre les individus x_1 et x_2 concernant la variable j qui se calcule de manière différente selon que la variable soit qualitative ou quantitative

La similarité partielle de l'indice de Gower entre deux individus x_1 et x_2 :

— pour une variable qualitative elle vaut 1 si la variable j prend la même valeur pour les deux individus ; 0 sinon
— pour une variable quantitative elle est égale au rapport entre la différence absolue entre les deux valeurs numériques $y_{1,j}$ et $y_{2,j}$ et l'écart maximal E_j observé sur l'ensemble des données soit $s_{12,j} = \frac{|y_{1,j}-y_{2,j}|}{E_j}$

Par suite, la distance de Gower notée D_g s'obtient simplement de la manière suivante :

$$D_g(x_1, x_2) = 1 - S_g(x_1, x_2) \tag{3.2}$$

La distance de Gower comprend donc des valeurs comprises entre 0 et 1. Elle est nulle entre deux individus identiques et est égale à 1 entre deux individus totalement différents.

$\longrightarrow$ Par conséquent, une fois la distance choisie, la matrice de dissimilarité D s'obtient en calculant la distance entre les individus deux à deux, selon l'une des méthodes de distance citée ci-dessus.

$$\begin{bmatrix} 0 & d(x_1, x_2) & d(x_1, x_3) \\ & 0 & d(x_2, x_3) \\ & & 0 \end{bmatrix}$$

Cette matrice de dissimilarité forme une matrice symétrique de taille $N \times N$, où N représente le nombre d'individus du jeu de données, avec des 0 sur la diagonale car la distance entre un individu et lui-même est nulle.

1.2.2 La méthode de regroupement

Une fois la matrice de dissimilarité calculée, les individus peuvent être regroupés en fonction de leur ressemblance. Les individus les plus proches sont regroupés ensemble pour former un groupe d'individus, selon la distance indiquée dans la matrice de dissimilarité .

Pour la comparaison entre deux individus, la matrice de dissimilarité qui indique la distance entre chacun des individus est utilisée. Toutefois, pour calculer la distance entre deux groupes ou entre un groupe et un individu, il faut utiliser une méthode de regroupement. C'est ce second paramètre qui est requis pour la classification ascendante hiérarchique : il faut donc sélectionner une méthode de regroupement.

La méthode de regroupement permet de déterminer la proximité entre deux *clusters* notés C_1 et C_2. Ces *clusters* peuvent représenter un ou plusieurs individus. Cette étape va permettre de fusionner les deux *clusters* pour construire le dendrogramme. Chacune des méthodes suivantes (liste non-exhaustive) produira un dendrogramme différent :

La distance minimale
La distance minimale se détermine en calculant l'ensemble des distances possibles, $d(x_1, x_2)$, entre les observations respectives des deux groupes considérés et consiste à retenir la plus petite. Cette distance minimale est :

$$d(C_1, C_2) = \min_{x_1 \in C_1, x_2 \in C_2} d(x_1, x_2)$$

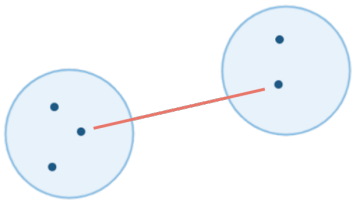

FIGURE 3.3 – Représentation de la distance minimale

Cette distance minimale est calculée entre chaque groupe existant, et la plus petite distance parmi ces distances minimales sera le critère menant à la fusion des deux groupes.

La distance maximale
La distance maximale se détermine en calculant l'ensemble des distances possibles, $d(x_1, x_2)$, entre les observations respectives des deux groupes considérés et consiste à retenir la plus grande. Cette distance maximale est :

$$d(C_1, C_2) = \underset{x_1 \in C_1, x_2 \in C_2}{max} d(x_1, x_2)$$

Cette distance maximale est calculée entre chaque groupe existant, et la plus petite distance

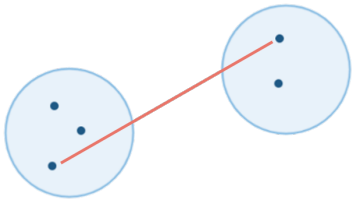

FIGURE 3.4 – Représentation de la distance maximale

parmi ces distances maximales sera le critère menant à la fusion des deux groupes.

La distance moyenne
La distance moyenne se détermine en calculant toutes les distances possibles, $d(x_1, x_2)$, entre les observations respectives des deux groupes étudiés et en déterminant la moyenne de ces distances. Cette distance moyenne est :

$$d(C_1, C_2) = \underset{x_1 \in C_1, x_2 \in C_2}{moyenne} d(x_1, x_2)$$

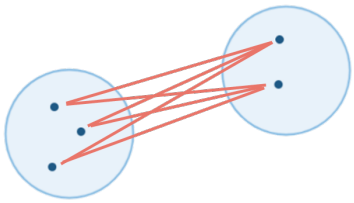

FIGURE 3.5 – Représentation de la distance moyenne

Cette distance moyenne est calculée entre chaque groupe existant, et la plus petite distance parmi ces distances moyennes sera le critère menant à la fusion des deux groupes.

La distance des barycentres
La distance des barycentres se détermine en calculant les barycentres des deux groupes à partir de la moyenne des coordonnées des observations respectives de ces *clusters*.

Le barycentre consiste à représenter un ensemble d'individus par un unique point. Le barycentre entre deux points identiquement pondérés se trouve au milieu du segment. Pour la généralisation à un nombre k de points, $k \in [1, n]$, le barycentre est le point G tel que :

$$\sum_{k=1}^{n} a_k \overrightarrow{GA_k} = \overrightarrow{0} \tag{3.3}$$

Avec :

— $n = |C_k|$ le nombre de points appartenant au *cluster* C_k

— A_k les points du *cluster* C_k
— a_k la pondération du point A_k
— G le barycentre du système $\{(A_k, a_k)_{(k\in[1,n])}\}$ de coordonnées : $x_{k,G} = \frac{\sum_{k=1}^{n} a_k x_{k,A_k}}{\sum_{k=1}^{n} a_k}$

Ainsi, en notant G_1 et G_2 les barycentres respectifs des clusters C_1 et C_2, la distance des barycentres est :

$$d(C_1, C_2) = d(G_1, G_2)$$

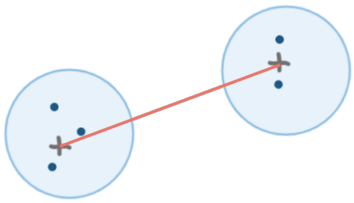

FIGURE 3.6 – Représentation de la distance des barycentres

Cette distance des barycentres est calculée entre chaque groupe existant, et la plus petite distance parmi ces distances de barycentres sera le critère menant à la fusion des deux groupes.

La distance de Ward

La distance de Ward tend à construire des groupes de la manière la plus distincte et cohérente possible. En effet, elle consiste à réunir les deux *clusters* dont le regroupement minimisera l'inertie intra-classe, soit l'inertie dans chaque *cluster*. Cette distance s'obtient donc en définissant la notion d'inertie qui est une mesure de dispersion dans un groupe d'individus. Plus l'inertie intra-classe est faible, plus les individus au sein de ce groupe seront similaires. A l'inverse, une forte inertie intra-classe indique que les individus sont dispersés au sein du *cluster*.

Mathématiquement, l'inertie d'un ensemble d'individus $\Gamma = \{x_i, i = 1, ..., N\}$, avec Γ appelé le nuage de points, correspond à la somme pondérée des carrés des distances de ces individus au centre de gravité du nuage. Donc si G désigne le centre de gravité de Γ alors l'inertie totale du nuage de points pour des individus pondérés par un poids w_i est égale à :

$$Inertie(\Gamma) = \sum_{i=1}^{N} w_i d(x_i, G)^2 \tag{3.4}$$

La distance de Ward entre deux clusters est le cas particulier où le nuage de points Γ inclut deux *clusters* et se détermine donc en calculant la distance au carré entre les barycentres de ces deux *clusters* avec leurs pondérations respectives :

$$d_{Ward}(C_1, C_2) = \frac{w_1 w_2}{w_1 + w_2} d(G_1, G_2)^2 \tag{3.5}$$

Avec :

— C_1 et C_2 les deux *clusters*
— G_1 et G_2 leurs barycentres respectifs
— w_1 et w_2 les poids respectifs des *clusters*

Par défaut, tous les individus ont le même poids $\frac{1}{N}$ et le poids w_i du *cluster* i est défini par :

$$w_i = \frac{\text{nombre d'individus du } \textit{cluster } i}{\text{nombre total d'individus}}$$

Cette distance de Ward est calculée entre chaque groupe existant, et la plus petite distance parmi ces distances de Ward sera le critère menant à la fusion des deux groupes.

1.3 Algorithme de la CAH

L'algorithme de la classification ascendante hiérarchique peut être résumé par la suite d'instructions suivante :

Entrée : chaque individu est considéré seul dans son groupe
* Matrice de distance D
* Choix d'une distance de regroupement

Etape itérative : jusqu'à agréger chaque individu dans un *cluster*
* Regroupement des deux éléments les plus proches selon la matrice de distance D
* Calcul de la nouvelle distance par le critère de regroupement entre ces deux éléments nouvellement fusionnés
* Mise à jour de la matrice D de distance

Sortie : L'algorithme converge vers une partition en un nombre prédéfini de *clusters*

2 Le partitionnement en K-moyenne

Principe du partitionnement en K-moyenne
La méthode du partitionnement en K moyenne, plus couramment connue sous le nom *K-means* en anglais, est une seconde méthode d'apprentissage non-supervisé utilisée en analyse de données. Elle a été conçue par M. McQueen en 1967. Son principe est de séparer les individus en K groupes distincts.

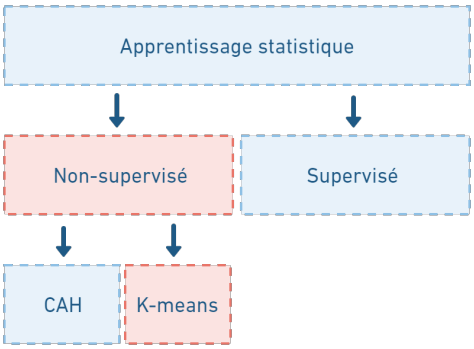

FIGURE 3.7 – Apprentissage statistique non-supervisé : *K-means*

Contrairement à la méthode CAH, la méthode du partitionnement en K-moyenne est radicalement différente dans la manière d'assembler les individus. En effet, elle procède non pas par une approche ascendante mais par une approche descendante. Elle divise les individus progressivement jusqu'à obtenir K groupes homogènes.

Paramètres du partitionnement K-moyenne
Le partitionnement K-moyenne se définit à partir de deux paramètres :

1. Un nombre K de *clusters*
2. Une méthode de division appliquée à la base de données (uniquement à variables quantitatives)

1. Le nombre de *clusters*

Le paramètre essentiel de la méthode du partitionnement en K-moyenne est le nombre K de *clusters* souhaité en résultat de l'algorithme et doit être renseigné par l'utilisateur. Ce nombre indique combien de centroïdes sont placés aléatoirement dans le plan lors de l'initialisation.

2. La méthode de division

Une fois le nombre de *clusters* renseigné, il faut mesurer la distance ou le degré de similarité des individus par rapport à chacun des centroïdes. Pour ce faire, l'algorithme *K-means* se base principalement sur la distance euclidienne ou la distance Manhattan (*définies ci-avant :* 1.2.1). C'est pourquoi, l'algorithme *K-means* a besoin d'un jeu de données avec des variables quantitatives. Toutefois, il existe plusieurs solutions permettant à cet algorithme de prendre en compte des variables caractéristiques qualitatives. Une de ces alternatives est d'utiliser l'algorithme des *K-médoïdes* qui prend en paramètre d'entrée la matrice des distances construite de la même manière que celle utilisée pour la CAH.

Les figures ci-dessous illustrent que les situations avant et après l'utilisation de la méthode de regroupement :

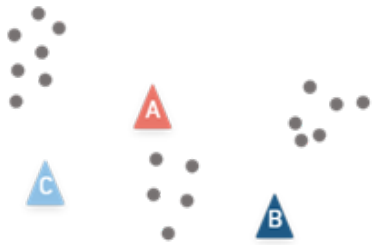

FIGURE 3.8 – Répartition des individus avant le partitionnement en K-moyenne

FIGURE 3.9 – Répartition des individus après le partitionnement en K-moyenne

Description du passage entre ces deux figures : le point le plus proche de l'un des centroïdes (A, B ou C), selon la distance euclidienne ou Manhattan, sera ajouté au *cluster* correspondant, puis la position du nouveau centroïde sera recalculée en prenant la moyenne des coordonnées de tous les individus qui lui sont assignés. Cette étape est répétée jusqu'à ce que tous les individus appartiennent à l'un des K *clusters*.

Algorithme du partitionnement K-moyenne

L'algorithme du partitionnement en K-moyenne peut être résumé par la suite d'instructions suivante :

Entrée : tous les individus sont considérés en un seul groupe
* Le nombre K de *clusters*
* Le jeu de données

Etape itérative : jusqu'à stabilité des nouveaux centroïdes
* Attribution de l'individu au plus proche centroïde : pour chaque individu calcul de la distance entre l'individu et les centroïdes
* Recalcul des centroïdes : pour chaque nouvel individu ajouté à un *cluster*, le nouveau centroïde du groupe est calculé

Sortie : Les centroïdes finaux représentent les résultats du *K-means*. Chaque individu appartient à un *cluster* spécifique

Chapitre 4

L'apprentissage supervisé

L'entraînement d'un modèle d'apprentissage supervisé se réalise à l'aide de données étiquetées c'est-à-dire sur les variables caractéristiques X et sur la variable cible Y de l'individu. L'algorithme de ce modèle vise à classifier de manière optimale de nouveaux individus en groupes qui capturent au mieux la relation entre les entrées X et les sorties observées Y.

L'algorithme est d'abord entraîné sur l'ensemble de données étiquetées. Il est ensuite utilisé pour prédire la classe de l'étiquette de nouvelles données non étiquetées, connaissant alors uniquement les variables caractéristiques X.

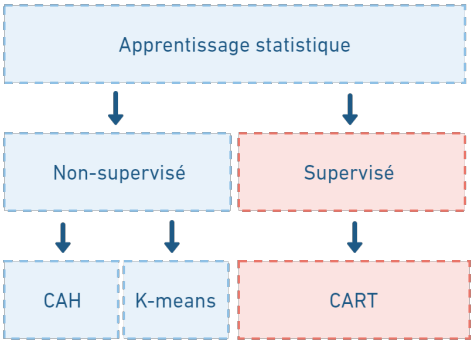

FIGURE 4.1 – Apprentissage statistique supervisé : CART

1 L'algorithme CART

1.1 Présentation de l'algorithme CART

L'algorithme CART (*Classification And Regression Tree*), a été introduit par Messieurs Breiman, Friedman, Olshen et Stone en 1984 [6]. Cet algorithme se décline de deux façons selon le type de la variable cible Y à prédire : l'arbre de décision CART visant à prédire des variables cibles Y catégorielles et l'arbre de régression CART approprié pour la prédiction des variables cibles Y non catégorielles, soit des variables continues. Pour la construction de l'arbre, ces deux versions de l'algorithme CART utilisent des critères différents afin d'optimiser de manière adaptée le regroupement d'individus en fonction de la variable cible Y, discrète ou continue. Les résultats issus de l'algorithme CART se représentent graphiquement par un arbre de classification ou de régression.

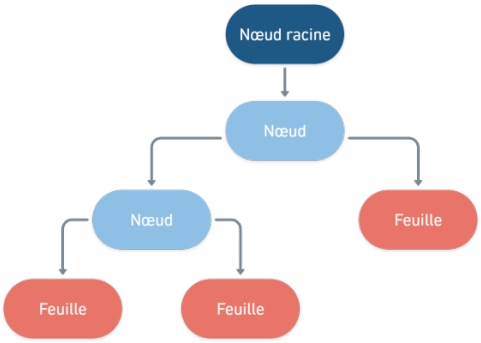

FIGURE 4.2 – Schéma d'un arbre CART

Un arbre est un classifieur, noté $T : X \to Y$ qui s'exprime sous la forme d'une succession de combinaisons de modalités des variables explicatives X contenues dans les données d'apprentissage.

L'arbre issu de l'algorithme d'apprentissage supervisé se compose en son premier niveau par un nœud racine suivi de deux branches sortantes créant deux nœuds fils : on parle d'arbre binaire. L'ensemble de ces nœuds effectuent des évaluations sur des sous-ensembles homogènes selon un critère de division appliqué aux caractéristiques des variables X disponibles. Les nœuds dits feuilles sont les derniers nœuds de l'arbre.

L'algorithme CART peut construire deux types d'arbres selon la variable cible Y :

L'arbre de décision CART
L'algorithme de décision CART est utilisé lorsque la variable cible est une variable catégorielle ou discrète. Son objectif est de construire un arbre qui divise l'ensemble des données en sous-ensembles homogènes en termes de classes Y. Parmi l'ensemble des variables caractéristiques X à disposition, l'algorithme sélectionne celles qui apportent le plus d'informations. Ces variables vont servir à diviser les données dans le but de maximiser l'homogénéité (équivalent à minimiser l'hétérogénéité) des classes dans chaque nœud de l'arbre.

L'arbre de régression CART
L'algorithme de régression CART est utilisé lorsque la variable cible est une variable continue. Son objectif est légèrement différent des arbres à classification car il s'agit de construire un arbre qui divise l'ensemble des données en sous-ensembles homogènes en termes de valeurs numériques prédites Y. Parmi l'ensemble des variables caractéristiques X à disposition, l'algorithme sélectionne également celles qui apportent le plus d'informations. Ces variables vont servir à diviser les données dans le but de minimiser l'erreur de régression dans chaque nœud de l'arbre.

1.2 L'arbre CART

Le processus de l'algorithme CART fournissant un arbre est le résultat de deux étapes :

1. **La phase de construction de l'arbre** qui consiste à ajuster une fonction d'hétérogénéité sur la base d'apprentissage par minimisation de l'erreur. L'erreur est définie par l'attribution d'un individu dans un nœud dans lequel sa variable réponse Y n'est pas du même bord que la variable cible Y que représente ce nœud ;
2. **La phase d'élagage de l'arbre** qui consiste à trouver l'arbre optimal en termes de nombre de nœuds terminaux (feuilles)

1.2.1 La construction de l'arbre

La construction de l'arbre par la méthode CART se fait progressivement. Le point de départ se situe au sommet, au nœud de la racine. A chaque nœud, l'algorithme vient séparer la population en plusieurs sous-populations selon un critère appliqué aux variables explicatives X. Cette séparation des données est discriminante, elle permet de séparer au mieux les deux sous-populations parmi les données d'apprentissage, selon une seule variable. Autrement dit, en chaque nœud, la séparation des données est relative à une seule des variables X et vient diviser la population en deux. D'un coté, un nombre fini de modalités de la dite variable est regroupé. De l'autre coté, le reste des modalités de la dite variable sont regroupées.

La fonction d'hétérogénéité
L'algorithme CART construit ainsi l'arbre de manière récursive en divisant l'ensemble des données initiales en sous-ensembles plus petits. A chaque étape, il sélectionne la variable pour diviser les données qui optimise un critère spécifique : la fonction d'hétérogénéité. Lorsque la fonction d'hétérogénéité est élevée, cela signifie que les valeurs de la variable cible Y au sein d'un même nœud sont différentes les unes des autres, ce qui indique une dispersion des individus nœud. En revanche, lorsque la fonction d'hétérogénéité est faible, cela signifie que les valeurs de Y au sein du nœud sont davantage similaires.

Pour effectuer la division, l'algorithme va chercher à maximiser l'écart entre l'hétérogénéité du nœud h et la somme des hétérogénéités de chacun des nœuds fils h_g et h_d. La finalité étant de rendre les hétérogénéités des deux sous-populations formées par les nœuds fils H_{h_g} et H_{h_d} les plus faibles possible. La division optimale notée d^* est donc celle qui résout :

$$d^* = \underset{j\in[1,p]}{max}\{\underset{\text{division de } X_j}{max}(H_h - H_{h_g} - H_{h_d})\} \tag{4.1}$$

La fonction d'hétérogénéité H : h $\rightarrow H(h)$ diffère selon la nature de la variable cible Y, et sera décrite par la suite.

Le critère de division peut également être appelé l'erreur R_h telle que :

$$R_h = H_h - H_{h_g} - H_{h_d} \tag{4.2}$$

Ce qui permet de réécrire la division optimale d^* par :

$$d^* = \underset{j\in[1,p]}{max}\{\underset{\text{division de } X_j}{max} R_h\} \tag{4.3}$$

$\rightarrow$ *Fontion d'hétérogénéité : variable cible Y discrète*

Lorsque Y est une variable discrète, la fonction d'hétérogénéité la plus couramment utilisée pour la classification est celle de l'indice de Gini. En notant Y^l la classe l de la variable réponse Y (pour un nombre total m de classes), de proportions respectives p_h^l, l'hétérogénéité du nœud h définie par l'indice de Gini est :

$$H_{Gini}(h) = \sum_{l=1}^{m} p_h^l(1-p_h^l) = \sum_{l=1}^{m} p_h^l - (p_h^l)^2 = 1 - \sum_{l=1}^{m}(p_h^l)^2 \tag{4.4}$$

Cet indice de Gini est ainsi calculé pour les nœuds h, h_g et h_d. Un indice de Gini faible indique que le œud est pur, c'est-à-dire qu'il contient principalement des données d'une même classe.

→ *Fontion d'hétérogénéité : variable cible Y continue*

Dans le cas d'une variable Y continue, l'hétérogénéité H du nœud est définie par sa variance :

$$H(h) = \frac{1}{|h|} \sum_{y_i \in h} (y_i - \overline{y_h})^2 \tag{4.5}$$

Où $\overline{y_h}$ représente la moyenne de la variable Y des individus présents dans le nœud h. Appliquée aux deux nœuds fils, la somme des fonctions d'hétérogénéité H_{h_g} et H_{h_d} est définie par :

$$H_{h_g} + H_{h_d} = p_{h_g} \sum_{y_i \in y_g} (y_i - \overline{y_{h_g}})^2 + p_{h_d} \sum_{y_i \in y_d} (y_i - \overline{y_{h_d}})^2 \tag{4.6}$$

Avec :
- $p_{h_g} = \frac{|h_g|}{|h|}$ la proportion d'individus dans le nouveau nœud gauche avec $|h_g|$ le cardinal du nœud gauche et $|h|$ celui du nœud h
- $p_{h_d} = \frac{|h_d|}{|h|}$ la proportion d'individus dans le nouveau nœud droit avec $|h_d|$ le cardinal du nœud droit et $|h|$ celui du nœud h

Arrêt du critère de division
L'algorithme arrête de diviser les feuilles une fois qu'une seule observation est présente dans la feuille ou lorsque les individus de la feuille ont les mêmes valeurs caractéristiques X. Ainsi, il est construit l'arbre dit "maximal" qui sera élagué lors de l'étape 2.

1.2.2 Élagage de l'arbre

Une fois l'arbre maximal développé, la méthode CART se poursuit en une seconde phase. Elle consiste à remonter l'arbre en partant des feuilles et à supprimer les nœuds dont la division n'améliore pas significativement l'arbre.

La construction de l'arbre T_{max} aboutit à un arbre complètement développé où chaque nœud terminal est pur. Ce nœud prend l'appellation d'une feuille. A chaque feuille, une valeur est affectée. Si Y est quantitative, il s'agit de sa moyenne pour les individus présents dans la feuille. Sinon, il s'agit de la modalité la plus représentée du nœud.

Dans cette seconde phase d'élagage, il n'y a pas de différence faite entre la classification ou la régression. Cette phase est donc identique quelle que soit la variable cible Y.

Construction des sous-arbres
Les notations suivantes vont permettre de présenter cette seconde partie de l'élagage de l'arbre :
- T_{max} est l'arbre maximal issu de l'étape 1
- T est un arbre
- $|T|$ est le nombre de feuilles de l'arbre T
- R_T est l'erreur associée à l'arbre T représentant la somme en chaque noeud de toutes les divisions réalisées par le critère d'hétérogénéité. Elle s'écrit : $R_T = \sum_{t \in T} R_t$ (avec R_t *définie ci-avant* 4.2)

A partir de l'arbre maximal T_{max}, l'algorithme d'élagage va construire une séquence de sous-arbres emboîtés de la forme : $T_1 \subset ... \subset T_{K-1} \subset T_K = T_{max}$ ùpour $j \in [1, K]$. La construction de cette sous-suite d'arbres se réalise à l'aide d'un critère d'élagage définit par la nouvelle fonction d'erreur $R_\alpha(T)$ introduisant le paramètre α :

$$R_\alpha(T) = R(T) + \alpha|T| \tag{4.7}$$

Avec :
- $R(T)$ le terme de "coût"
- $\alpha|T|$ le terme "d'optimisation" qui traduit le critère de pénalisation par une contrainte appliquée sur le nombre de feuilles de l'arbre
- α un réel ≥ 0, qui est le paramètre de compromis entre la pénalisation de la complexité de l'arbre et la qualité de l'ajustement aux données (pour $\alpha = 0$, l'arbre obtenu correspond à l'arbre maximal T_{max})

Ainsi, la construction de la sous-suite d'arbres $T_1 \subset ... \subset T_{K-1} \subset T_K = T_{max}$ revient à construire une sous-suite d'arbres $T_{\alpha_1} \subset ... \subset T_{\alpha_{K-1}} \subset T_{\alpha_K}$ par la fonction d'erreur $R_\alpha(T)$. Cette fonction se base sur le choix de la valeur α. A la différence du critère de division classique $R(T)$ utilisé lors de la phase de construction de l'arbre, le nouveau critère de division défini par $R_\alpha(T)$ arrête la division de l'arbre lorsque l'écart d'hétérogénéité devient plus petit que ce seuil α. En effet, pour diviser un nœud h issu d'un arbre comportant K feuilles en deux nœuds fils comportant un arbre plus profond avec $K + 1$ feuilles, le critère $R_\alpha(T)$ s'exprime :

$$R_\alpha(T_{K+1}) = R(T_{K+1}) + \alpha|T_{K+1}| = R(T_{K+1}) + \alpha(K + 1) \tag{4.8}$$

De même,

$$R_\alpha(T_K) = R(T_K) + \alpha|T_K| = R(T_K) + \alpha K \tag{4.9}$$

Ce qui implique par la soustraction de 4.9 par 4.8, de toujours avoir la relation suivante :

$$R_\alpha(T_K) - R_\alpha(T_{K+1}) > 0 \Leftrightarrow R(T_K) - R(T_{K+1}) > \alpha \tag{4.10}$$

Ceci démontre que le processus de division de l'arbre cesse dès lors que la division d'un nœud h en deux nœuds fils produit un écart d'hétérogénéité inférieur au seuil α. C'est pourquoi, lorsque $\alpha = 0$, le processus continue de se développer jusqu'à ce que l'arbre obtenu soit l'arbre maximal T_{max}.

Cette suite de sous-arbres peut être représentée par la fonction f qui à α associe la valeur $R_\alpha(T)$ soit f : $\alpha \to R_\alpha(T)$. Cette fonction admettra un minimum global car elle sera en forme de "U". Ce minimum sera donc la solution du meilleur arbre optimal.

Toutefois, les arbres CART sont considérés comme instables car de petites fluctuations dans l'échantillon d'apprentissage peuvent entraîner des variations importantes dans la structure de l'arbre et donc dans les prédictions du modèle. Cette instabilité peut être problématique car elle peut conduire à des résultats très différents pour des échantillons légèrement différents, ce qui rend le modèle peu fiable et difficile à interpréter.

Pour atténuer cette instabilité, d'autres approches peuvent être utilisées, comme celles présentées ci-après : l'algorithme *Random Forest* et l'algorithme *XGBoost*.

2 L'algorithme Random Forest

L'algorithme *Random Forest* a été formellement proposé par M. Breiman en 2001 [1]. Cet algorithme a pour objectif d'agréger un ensemble d'arbres utilisé pour la classification ou la régression, par la méthode du *bagging*, contraction de *bootstrap aggregating*. Le *bagging* est une méthode d'agrégation parallèle des arbres de la forêt aléatoire. L'algorithme *Random Forest* opère selon un double *bagging*. L'un consiste à échantillonner aléatoirement les observations pour construire chaque arbre. L'autre consiste à échantillonner aléatoirement les variables explicatives pour chaque arbre de décision, ce qui aide à réduire la corrélation entre les arbres et à améliorer la robustesse du modèle. Cet algorithme est donc issu de deux grandes étapes :

1. ***Le tree bagging*** : la première étape consiste à créer un nombre q d'échantillons aléatoires à partir de la base d'apprentissage initiale sur lesquels seront appliqués l'algorithme. Ces échantillons sont construits par la technique de ré-échantillonage du *bagging*, permettant d'obtenir des tirages avec remise de N individus. Chaque individu a la probabilité $\frac{1}{N}$ d'être tiré. Ces q échantillons d'arbres obtenus représentent la forêt aléatoire.

2. ***Le feature sampling*** : cette étape consiste à réaliser un second tirage aléatoire sur les variables explicatives en chacun des nœuds des q arbres, pour fournir au nœud un nombre de m variables caractéristiques sur lequel la division optimale est recherchée. Ce nombre m correspond à un tirage aléatoire : classiquement $m = \sqrt{p}$ pour la classification et $m = \frac{p}{3}$ pour la régression, parmi les p variables explicatives initialement présentes dans la base d'apprentissage.

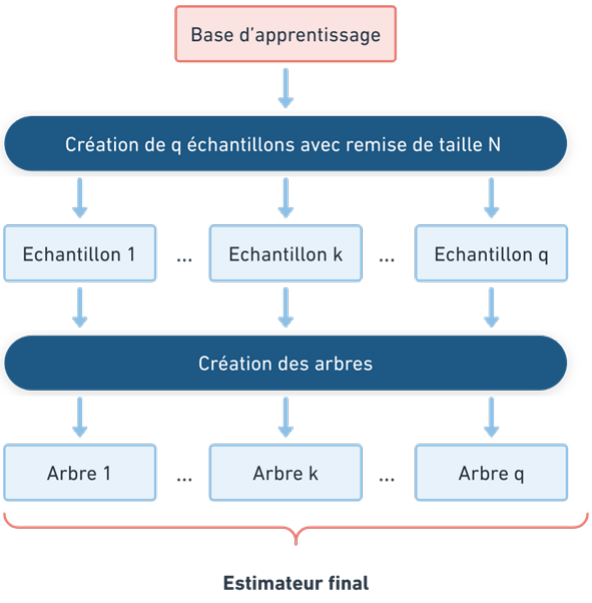

FIGURE 4.3 – Algorithme *Random Forest* (simplifié sur 3 arbres)

Finalement, pour une forêt aléatoire, l'estimateur final de la variable cible Y est obtenu par :
— un vote majoritaire sur les q arbres dans le cas d'une classification ;
— la moyenne des estimations issues des q arbres dans le cas d'une régression.

Le *Random Forest* fonctionne ainsi sur plusieurs estimateurs qui, en les assemblant, donnent une vision globale sur la prédiction de la variable cible. Ces prédictions sont davantage robustes quand le nombre de sous-échantillons q est grand.

3 L'algorithme XGBoost

L'algorithme *XGBoost*, de son nom *eXtreme Gradient Boosting*, a été développé par M. Chen et M. Guestrin en 2016 [2]. A contrario de l'algorithme *Random Forest* qui construit plusieurs arbres en parallèle, l'algorithme *XGBoost* construit des arbres en série, utilisant la méthode d'agrégation du *bootsting*. Le principe du *boosting* est d'améliorer le modèle de l'arbre initial au fur et à mesure en donnant plus de poids aux individus difficiles à prédire. Les poids w_i fournis à chaque nouveau modèle permettent de capter les individus qui ont mal été prédits par les classifieurs précédents. Le modèle s'améliore donc tout en apprenant de ses erreurs, où chaque arbre généré a accès à l'erreur de son prédécesseur.

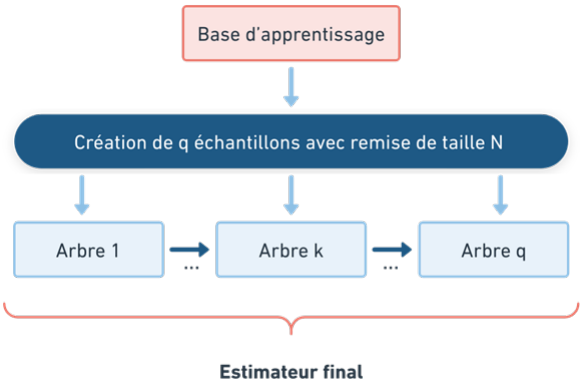

FIGURE 4.4 – Algorithme XGBoost (simplifié sur 3 arbres)

Finalement, par le *bootsting*, l'estimateur final de la variable cible Y est obtenu par :
- un vote majoritaire sur les q arbres dans le cas d'une classification ;
- la moyenne des estimations issues des q arbres, pondérée par les poids w_i obtenus grâce aux i modèles prédécesseurs dans le cas d'une régression.

Troisième partie

Une adaptation à l'approche actuelle de la classification ascendante hiérarchique

Cette partie présente la théorie des algorithmes innovants d'apprentissage statistique utilisés dans la suite de ce mémoire.

Le coeur de ce mémoire est la présentation d'une application actuarielle d'un algorithme, décrit dans cette partie, qui vise à pallier une contrainte inhérente à la CAH lorsqu'elle est appliquée à l'étude de la transformation des devis en contrats.

Actuellement, la méthode de la classification ascendante hiérarchique regroupe les individus qui se ressemblent dans un même *cluster*. Toutefois, ces *clusters* se ressemblent uniquement d'un point de vue des variables explicatives X portées par les individus et les résultats obtenus n'intègrent pas la composante de la variable à expliquer. Cette approche classique implique donc de regarder "à la main" les *clusters* qui correspondraient à des bons ou mauvais risques.

Or, dans le contexte du mémoire, la base de données présentée à la partie suivante offre la connaissance quant à la transformation ou non du devis en contrat. Il peut donc être intéressant d'utiliser cette information car celle-ci est connue.

Les approches mises en œuvre dans ce mémoire issues de la CAH classique, appelées **CAH1** et **CAH2**, proposent justement de prendre en compte les informations issues de la variable Y. Ceci permet de regrouper les individus par groupes homogènes en caractéristiques et en risque.

En associant la CAH classique à la connaissance de la transformation du devis en contrat par les individus de la base de données, une approche de *clustering* plus complète et adaptée aux données actuelles pourra être réalisée. Cette adaptation de l'approche de la classification ascendante hiérarchique peut s'illustrer comme suit :

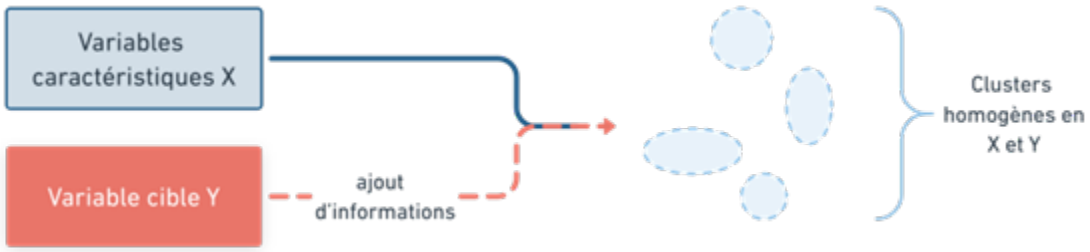

FIGURE 4.5 – Illustration de la nouvelle approche

Les résultats à venir de ces approches CAH1 et CAH2, permettront de conclure sur l'efficacité de ces méthodes à distinguer des profils types plus susceptibles de transformer leur devis en contrat, comparativement à l'approche classique de la CAH.

Chapitre 5

CAH1 : ajout d'une seconde matrice de dissimilarité

Cette première approche innovante, offre une alternative à la méthode classique CAH. Elle peut être vue comme un compromis entre deux CAH obtenues à l'aide de deux sources d'informations :

— d'une part les variables caractéristiques X ;
— d'autre part la variable cible Y de l'individu.

L'appellation de compromis permet de "*caractériser qu'une action de concession réciproque entre les deux sources de données (variables X et variable Y) est nécessaire dans l'implémention de cette approche*". [Cette approche a été étudiée par un groupe de chercheurs en statistiques en 2020 dans un contexte d'application socio-économique [8]. L'idée issue de cet article est de regrouper des individus qui sont proches géographiquement entre eux en prenant en compte deux sources d'informations : les caractéristiques des individus et leurs proximités géographiques. Dans le cadre de ce mémoire, la deuxième source d'information est adaptée différemment.]

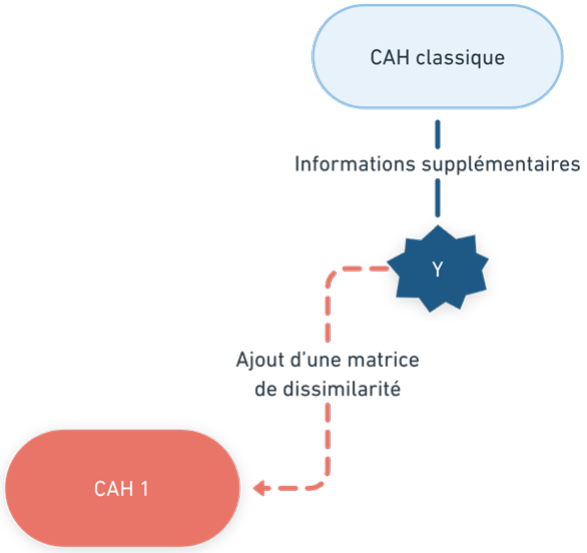

FIGURE 5.1 – Illustration simplifiée d'approche CAH1

1 Présentation des paramètres de la CAH1

Deux nouveaux paramètres sont requis pour implémenter cette approche de *clustering* :

1. La seconde matrice de dissimilarité relative à la variable Y
2. Le paramètre, dit de "mélange", α

1.1 Ajout d'une matrice de voisinage

L'amélioration de cet algorithme consiste à réaliser une CAH classique en utilisant deux matrices de dissimilarité $D1$ et $D2$. Initialement, une seule matrice est utilisée, correspondant à la matrice des distances calculée entre chaque individu. Cette matrice $D1$ se calcule exactement de la même manière que celle de la CAH classique (*définie ci-avant :* 1.2.1). La matrice $D1$ ressemble donc, à titre d'exemple, à la matrice ci-dessous :

$$\begin{bmatrix} 0 & d_1(x_1,x_2) & d_1(x_1,x_3) \\ & 0 & d_1(x_2,x_3) \\ & & 0 \end{bmatrix}$$

Toutefois, cette matrice de distances est uniquement calculée à partir des données caractéristiques X. C'est pourquoi, l'ajout d'une seconde matrice $D2$, dite matrice de voisinage, va venir apporter une information supplémentaire sur la variable cible Y. Pour construire la matrice $D2$, il faut comparer les modalités de la variable Y entre tous les individus. Si deux individus ont la même modalité, leur distance sera nulle. A l'inverse, si deux individus n'ont pas la même modalité, leur distance sera non nulle, fixée à 1.

Cette matrice $D2$ sera donc construite comme suit pour chaque paire d'individus (x_i,x_j) :

— $d_2(x_i, x_j) = 0$ si les deux individus ont la même modalité Y
— $d_2(x_i, x_j) = 1$ si les deux individus n'ont pas la même modalité Y

La matrice $D2$ ressemble donc, à titre d'exemple, à la matrice ci-dessous :

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

1.2 Le paramètre α

Les matrices $D1$ et $D2$ sont reliées par un paramètre de mélange α compris entre 0 et 1 et permettent de construire la matrice D_α de la manière suivante :

$$D_\alpha = \alpha D1 + (1-\alpha)D2$$

L'intérêt du paramètre α est de contrôler la part d'information apporter par chacune de ces deux matrices. Si $\alpha = 0$, alors la CAH1 sera seulement basée sur la matrice $D2$ et à l'inverse, si $\alpha = 1$, alors la CAH1 sera seulement basée sur la matrice $D1$. Dès lors, le principal enjeu de cette méthode réside dans la détermination optimale du paramètre α. L'idée de ce paramètre est d'augmenter l'intérêt de la variable Y sans pour autant ignorer les informations issues des résultats des variables caractéristiques X.

Le choix du paramètre α représente l'enjeu principal dans l'implémention de cette méthode. Il sera détaillé lors de l'application de la méthode aux données dans la *partie* V.

2 Mesure d'agrégation entre deux groupes

D'après la méthode classique de la CAH, il existe plusieurs manières de calculer le regroupement des individus. Celle retenue dans cette approche est la distance de Ward (*définie ci-avant :*

1.2.2). Pour appliquer la méthode de Ward à deux matrices de dissimilarités D_1 et D_2, il faut tout d'abord définir l'inertie mixée notée I_α d'un *cluster* C_k^α :

$$I_\alpha(C_k^\alpha) = \alpha \sum_{i \in C_k^\alpha} \sum_{j \in C_k^\alpha} \frac{w_i w_j}{2\mu_k} d_1^2(x_i, x_j) + (1-\alpha) \sum_{i \in C_k^\alpha} \sum_{j \in C_k^\alpha} \frac{w_i w_j}{2\mu_k} d_2^2(x_i, x_j) \tag{5.1}$$

Avec :
— w_i,w_j les poids respectifs des individus x_i et x_j
— $\mu_k = \sum_{i \in C_k^\alpha} w_i$, le poids du *cluster* C_k^α
— $d_1(x_i, x_j)$ la distance entre deux individus x_i et x_j selon la matrice D_1
— $d_2(x_i, x_j)$ la distance entre deux individus x_i et x_j selon la matrice D_2

Il en découle que pour une partition de K *clusters* notée $P_K^\alpha = (C_1^\alpha, ..., C_K^\alpha)$, l'inertie totale mixée W_α est égale à la somme des toutes les inerties mixées de ses *clusters* :

$$W_\alpha(P_K^\alpha) = \sum_{k=1}^{K} I_\alpha(C_k^\alpha)$$

La distance de Ward dans le cas de cette nouvelle approche entre deux *clusters* de barycentres respectifs G_1^α et G_2^α, correspond à :

$$d_\alpha(C_1^\alpha, C_2^\alpha) = \alpha \frac{\mu_i \mu_j}{\mu_i + \mu_j} d_1^2(G_1^\alpha, G_2^\alpha) + (1-\alpha) \frac{\mu_i \mu_j}{\mu_i + \mu_j} d_2^2(G_1^\alpha, G_2^\alpha) \tag{5.2}$$

L'idée de la méthode de Ward est d'agréger deux *clusters* C_1^α et C_2^α de telle manière à minimiser l'inertie mixée lorsque deux *clusters* sont réunis. Ainsi, pour obtenir une nouvelle partition optimale P_K^α à K *clusters* à partir de la partition P_{K+1}^α à $K+1$ *clusters*, il faut résoudre le roblème suivant :

$$\min_{C_1^\alpha, C_2^\alpha \in P_{K+1}^\alpha} I_\alpha(C_1^\alpha \cup C_2^\alpha) - I_\alpha(C_1^\alpha) - I_\alpha(C_2^\alpha)$$

Dans le cas particulier où les *clusters* ne sont composés que d'un seul individu respectivement x_i et x_j, la distance de Ward se simplifie comme suit :

$$d_\alpha(x_i, x_j) = \alpha \frac{w_i w_j}{w_i + w_j} d_1^2(x_i, x_j) + (1-\alpha) \frac{w_i w_j}{w_i + w_j} d_2^2(x_i, x_j) \tag{5.3}$$

3 Algorithme de la CAH1

L'algorithme de la classification ascendante par de cette première approche hiérarchique peut être résumé par la suite d'instructions suivantes :

Entrée : chaque individu est considéré seul dans son groupe
* Matrice de distance $D1$
* *Matrice de distance* $D2$
* *Choix du paramètre* α

Etape itérative : jusqu'à agréger chaque individu dans un *cluster*
* Regroupement des deux éléments les plus proches selon la matrice de distance D_α (résultant des matrices $D1$, $D2$ et de α)
* Calcul de la nouvelle distance par le critère de regroupement entre ces deux éléments nouvellement fusionnés
* Mise à jour de la matrice D_α de distance

Sortie : L'algorithme converge vers une partition en un nombre prédéfini de *cluster*

Chapitre 6

CAH2 : pondération par l'importance des variables caractéristiques

Cette seconde approche, en alternative à la méthode classique CAH, consiste à pondérer les variables caractéristiques X par leur importance perçue dans le choix de la classification des individus selon la variable Y. Cela permet de prendre en compte l'importance relative des différentes variables X et d'ajuster leur contribution dans le modèle. Toutes les variables ne sont pas nécessairement importantes dans l'analyse.

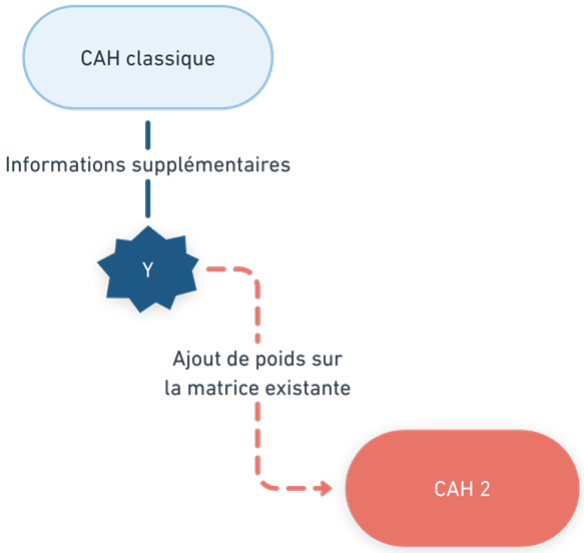

FIGURE 6.1 – Illustration simplifiée de l'approche CAH2

1 Présentation des paramètres de la CAH2

Initialement, dans la CAH, la matrice de distance construite entre tous les individus deux à deux ne prend pas en compte la différence d'importance qui pourrait exister pour chacune des variables caractéristiques X. Or, il peut être intéressant de repérer l'importance des variables utilisées. Ceci permet de reconnaître celles qui jouent un rôle essentiel dans l'apport d'informations, concernant le regroupement d'individus conformément à la variable cible Y.

1.1 Importance des variables par la méthode de permutations

L'ajout de l'importance des variables va apparaître pendant l'étape de la construction de la matrice de distance entre tous les individus. Dans la CAH classique, le poids de chaque variable est identiquement égale à $\frac{1}{p}$, avec p le nombre de variables caractéristiques. Ici, il s'agit d'attribuer un poids représentatif de l'importance des variables. Ce poids va être déterminé par la méthode de la mesure de l'importance des variables, appelée PFI (*Permutation Feature Importance*), introduite par M. Fisher en 2018 [7]. Cette méthode a l'avantage d'être indépendante du modèle considéré. Elle pourra aussi bien être implémentée sur un modèle *Random Forest* que sur un modèle *XGBoost*.

L'idée de cette méthode se base sur la théorie que si une variable est dite importante alors la modification des valeurs de cette variable aura un fort impact sur le modèle. Cette méthode consiste donc à permuter toutes les valeurs de la variable et à regarder la sensibilité du modèle par rapport à cette permutation. Si le modèle est sensible à cette permutation, cela indique que la variable joue un rôle important. En revanche, si la variable, après modification de ses valeurs n'a pas d'impact significatif sur les prédictions du modèle, elle ne sera pas considérée comme importante.

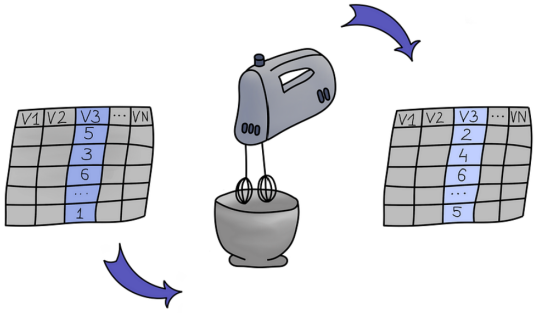

FIGURE 6.2 – Illustration de la permutation de la colonne *coloriée V3* (*Source : R-Bloggers*

Description mathématique de la *Permutation Feature Importance*
Soit un jeu de données à N individus et p variables explicatives, formant la matrice des variables explicatives, notée $(x_j^{(i)})$, avec $1 \leq i \leq N$ et $1 \leq j \leq p$, et y le vecteur des valeurs de la variable cible.

Soit f la fonction associée au modèle étudié et L la fonction de perte, également nommée fonction d'erreur. C'est une mesure pour évaluer la qualité de la prédiction d'un modèle par rapport aux valeurs réelles des données. La fonction de perte compare donc les prédictions du modèle aux valeurs réelles de la variable cible Y en attribuant un score qui quantifie cet écart. Son objectif est d'être minimale pour que les valeurs prédites soient les plus proches des valeurs réelles. Pour les modèles de régression, la fonction de perte L est la RMSE (*Root Mean Square Deviation* et pour les modèles de classification, la fonction de perte L est la mesure de performance "1-AUC" (*Area Under Curve*) ; détaillée au chapitre 4.2.

L'algorithme du calcul de l'importance des variables par la méthode PFI est le suivant :

- L'erreur initial $L_0 = L(y, f(x))$ est calculée
- Pour $j \in [1; p]$:
  - Une permutation aléatoire σ est choisie de $1, ..., N$ dans $1, ..., N$ ce qui entraîne une nouvelle matrice de variables explicatives, notée $(\tilde{x}_j^{(i)})_{\substack{1 \leq i \leq N \\ 1 \leq j \leq p}}$ et telle que :

$$\forall i \in 1, ..., N, \forall k \in 1, ..., p, \tilde{x}_k^{(i)} = \begin{cases} x_k^{(i)} si\ k \neq j \\ x_j^{(\sigma(i))} si\ k = j \end{cases}$$

Cette permutation permet d'attribuer aléatoirement aux individus, des nouvelles valeurs issues de la variable j étudiée, et de laisser toutes les autres variables intactes
    * L'erreur liée à cette permutation $L_{j,\sigma_j} = L(y, f(x))$ est calculée
    * L'importance de cette variable X_j est calculée par le rapport $V_{j,\sigma_j}(X_j) = \frac{L_{j,\sigma_j}}{L_0}$
— L'algorithme fourni en sortie l'importance $V_j(X_j)$ de chacune des p variables explicatives, triée par ordre croissant, pour plusieurs permutations effectuées

Ainsi, chacune des variables caractéristiques présente dans le jeu de données étudié aura un poids, relatif à son importance, noté

$$w_j = \frac{V_j(X_j)}{\sum_{j=1}^{p} V_j(X_j)}$$

1.2 La distance pondérée

Un fois le poids des variables déterminé, il est possible de calculer une distance de Gower pondérée entre les individus, contrairement à la version classique de la CAH dans laquelle la distance de Gower n'est pas pondérée. Ainsi, l'indice de Similarité de Gower, notée S_g, entre deux individus tout en tenant compte d'une pondération w_j, $j \in [1, p]$, pour chacune des variables caractéristiques vaut :

$$S_g(x_1, x_2) = \frac{1}{p} \sum_{j=1}^{p} w_j s_{12,j} \tag{6.1}$$

En rappelant que $s_{12,j}$ représente la similarité partielle entre les individus x_1 et x_2 concernant la variable X_j, définie en détail au paragraphe 1.2.1.

Ce qui aboutit à une nouvelle distance de Gower D_g :

$$D_g(x_1, x_2) = 1 - S_g(x_1, x_2)$$

2 Algorithme de la CAH2

L'algorithme de la classification ascendante hiérarchique dans de cette seconde approche peut être résumé par la suite d'instructions suivante :

Entrée : chaque individu est considéré seul dans son groupe
    * *Matrice de distance D pondérée par l'importance des variables caractéristiques X*
    * Choix d'une distance de regroupement

Etape itérative : jusqu'à agréger chaque individu dans un *cluster*
    * Regroupement des deux éléments les plus proches selon la matrice de distance D
    * Calcul de la nouvelle distance par le critère de regroupement entre ces deux éléments nouvellement fusionnés
    * Mise à jour de la matrice D de distance

Sortie : L'algorithme converge vers une partition en un nombre prédéfini de *cluster*

Quatrième partie

Présentation de la base de données

Chapitre 7

Construction de la base de données

Les deux approches de clustering *qui intègrent la variable cible Y peuvent trouver leur application sur une base de données traitant la transformation de devis en contrat. En prenant en compte la variable cible, ces méthodes peuvent fournir des regroupements de données qui non seulement capturent les similarités entre les devis, mais également tiennent compte du résultat final de la transformation ou non en contrat.*

L'analyse des résultats obtenus par ces approches sur une base de devis pourrait permettre d'identifier des populations ou des profils types précis. Cela permettrait au comparateur de cibler ses démarches commerciales pour augmenter le nombre de devis sur certaines populations qui ont des taux de transformation plus élevés par exemple, ou encore d'adapter les produits proposés afin d'augmenter le taux de transformation des profils pour lesquels le nombre de devis simulés est important mais le taux de transformation trop faible.

À présent que toutes les méthodes sont définies, il est nécessaire de procéder à la création et au nettoyage de la base de données. Cette phase revêt une importance primordiale au sein de l'analyse de données. Son objectif est de nettoyer les données brutes en vue de les préparer à une utilisation optimale dans les algorithmes d'apprentissage statistique.

Le schéma ci-dessous synthétise les grandes étapes du retraitement des données. Ces étapes sont appliquées à la base de données brute afin d'obtenir une base retraitée et exploitable pour la suite.

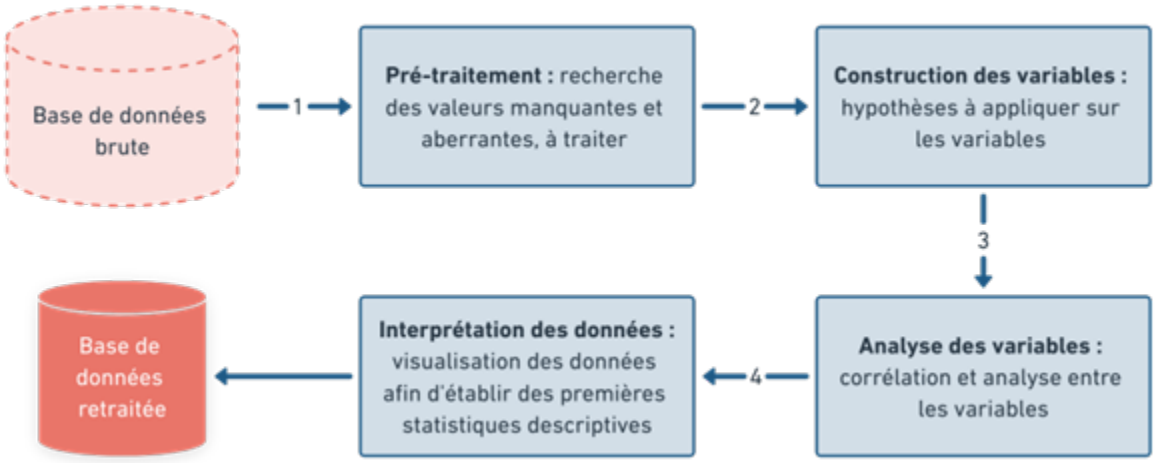

FIGURE 7.1 – Les grandes étapes du retraitement des données

Les données présentées ci-après s'inscrivent dans un cadre commercial et se résument en l'analyse du taux de transformation de devis en contrats.

1 Présentation des données

Origine des données
La base de données utilisée contient des données réelles issues d'un comparateur d'assurance en ligne. Chacune des lignes représente un devis. Dans la pratique, il est possible qu'une même personne puisse simuler plusieurs devis sur le comparateur. C'est pourquoi le terme utilisé pour représenter les lignes est un "devis" et non pas un "individu".

Filtrage de la base de données
Les données sélectionnées concernent :

- un historique de valeurs comprises entre janvier 2020 et juin 2023
- les crédits immobiliers amortissables liés à l'achat de résidences principales, de résidences secondaires ou d'investissements locatifs, d'une durée stricte de 7 ans, 10 ans, 15 ans, 20 ans ou 25 ans,
- l'ensemble des régions de France à l'exception des DROM-COM (Départements et Régions d'Outre-Mer et Collectivités d'Outre-Mer) et de la Corse

.

A la suite de ce filtrage, les devis deviennent comparables entre eux car ils concernent uniquement les crédits immobiliers, et la base comporte :

- 203 974 devis
- Moins de 1% de valeurs manquantes (dans toute la base de données filtrée)

2 Hypothèses réalisées

Les devis concernent des nouveaux prêts immobiliers (souscription d'une assurance emprunteur) ou des prêts en cours (changement d'assurance emprunteur). Certains de ces devis peuvent facilement être simulés en ligne par des personnes pouvant ne pas être informées des conditions actuelles du marché. Afin de travailler avec des données cohérentes, des hypothèses sont appliquées :

Âge lors de la réalisation du devis

- L'âge maximal est fixé à 64 ans, en raison de la réforme des retraites.
- L'âge minimal est fixé à 20 ans pour les non-cadres, à 23 ans pour les cadres et à 22 ans pour les autres catégories socio-professionnelles. Les hypothèses se basent sur les résultats d'une enquête réalisée par l'INSEE sur l'âge moyen du premier emploi en France : "*Les jeunes trouvent leur premier emploi significatif à 22 ans et 5 mois en moyenne*". [1]

Taux du crédit immobilier[2]

- Le taux maximal d'emprunt pour un crédit immobilier est fixé à : 4.80%[3], correspondant au taux maximal relevé pour une durée de 25 ans, juin 2023.
- Le taux minimal d'emprunt pour un crédit immobilier est fixé à : 0.20%[4], correspondant au taux minimal relevé pour une durée de 7 ans, janvier 2020.

Il convient de noter que ces hypothèses reflètent des années marquées par des taux très bas, en 2020, et une augmentation des taux à partir de 2022 et se poursuivant en 2023.

1. En début de carrière, les mobilités permettent d'améliorer la qualité de l'emploi : INSEE, analyse Rhône-Alpes, novembre 2014
2. Période d'observation du taux maximal et minimal : de janvier 2020 à juin 2023
3. Taux maximal durée 25 ans juin 2023 : Empruntis, historique des taux d'intérêt du crédit immobilier, dernière mise à jour à date du 01.08.2023
4. Taux minimal durée 7 ans janvier 2020 : Empruntis, historique des taux d'intérêt du crédit immobilier, dernière mise à jour à date du 01.08.2023

La durée du crédit immobilier

— La durée maximale du crédit immobilier est fixée à 25 ans, selon les recommandations du Haut conseil de stabilité financière de janvier 2021.[5]

Montant de crédit immobilier

— Le montant minimal du crédit immobilier est fixé à 50 000 €, selon un courtier d'assurance : "*En règle générale, le montant minimum d'un crédit immobilier s'établit à 50 000 €*"[6]

— Le montant maximal du crédit immobilier est fixé arbitrairement à 1 000 000 €, de telle sorte à exclure les montants extrêmes.

A la suite de ces hypothèses, la base de données comporte 175 142 devis.

3 Présentation des variables

Les données comportent 15 variables explicatives et une variable cible notée Y, représentant le taux de transformation des devis en contrats. Un devis qui ne s'est pas transformé en contrat a une étiquette de 0, à l'inverse d'un devis transformé en contrat qui a une étiquette de 1.

Les variables caractéristiques peuvent être présentées sous trois grandes catégories :

1. Les variables caractéristiques propres au devis

Les devis sont étudiés au travers des années 2020 à 2023, des mois, des jours de la semaine et des horaires. Ces variables sont issues des données enregistrées par le comparateur d'assurance lors de la réalisation du devis en ligne. De plus, le comparateur enregistre ces mêmes informations lorsqu'une modification est apportée au devis enregistré. Une variable binaire indiquant si le devis a été modifié ou non a donc été créée.

2. Les variables caractéristiques propres à la personne réalisant le devis

Concernant les caractéristiques de la personne réalisant le devis, celles disponibles sont : l'âge au moment de la réalisation du devis, la civilité, si cette personne est fumeur, son département actuel et si elle souhaite souscrire à la garantie MNO (par une présence de mal de dos et/ou de suivi psychologique).

3. Les variables caractéristiques propres au crédit immobilier

Les autres variables concernent le crédit en lui-même caractérisé par : son montant, son taux d'emprunt et sa durée d'emprunt.

En raison du très faible nombre de valeurs manquantes, peu de corrections sont apportées et le traitement des valeurs aberrantes est réalisé de manière à respecter les hypothèses imposées.

5. Recommandations R-HCSF-2021-1 relative à l'octroi de crédits immobiliers résidentiels en France : Haut conseil de stabilité financière, janvier 2021

6. Montant minimal : Réassurez-moi

L'ensemble de ces données sont répertoriées dans le tableau suivant où l'utilisateur désigne la personne ayant réalisé le devis sur le comparateur en ligne :

Nom de la variable	**Description de la variable**	**Valeurs prises**
annee_devis	L'année de la création du devis	2020 à 2023
mois_devis	Le mois de la création du devis	Janvier à Décembre
jours_devis	Le jour de la création du devis	Lundi à Dimanche
heure_devis	L'heure de la création du devis	0h00 à 23h59
modification_devis	Indique si le devis a été modifié au moins une fois	Oui / Non
age_devis	L'âge de l'utilisateur à la date du devis	20 à 64 ans
psy	Indique si l'utilisateur souscrit à la garantie MNO	Oui / Non
civilite	Sexe de l'utilisateur	Homme / Femme
fumeur	Indique si l'utilisateur est un fumeur	Oui / Non
csp	Catégorie socio-professionnelle de l'utilisateur	Cadres / Non-Cadres / Autres (retraités, fonctionnaires, libéraux...)
region	Région actuelle de l'utilisateur	Les régions françaises portant sur la Nouvelle Organisation Territoriale de la République (NOTRe)
duree_pret	La durée du prêt	7, 10, 15, 20 et 25 ans
montant	Le montant du prêt	50 000 € à 1 000 000 €
taux	Le taux du prêt	0.20% à 4.80%
Y	Transformation du devis en contrat	0 / 1

FIGURE 7.2 – Présentation des variables

4 Corrélation entre les variables caractéristiques

L'étude de la corrélation des variables caractéristiques permet d'identifier leurs relations linéaires et indique si elles sont indépendantes les unes des autres, ou au contraire, si elles apportent de l'information similaire au regard de la variable à expliquer. Dans ce cas, il est judicieux de supprimer les variables fortement corrélées entre elles pour éviter d'affecter les performances et l'interprétation des futurs modèles.

Corrélation des variables quantitatives
Pour l'étude des corrélations entre deux variables quantitatives, X_1 et X_2, le coefficient de corrélation utilisé est celui de Pearson, r_s, défini par :

$$r_s(X_1, X_2) = \frac{cov(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}$$

Avec $cov(X_1, X_2)$ désignant la covariance entre les variables caractéristiques X_1, X_2, et σ_{X_1}, σ_{X_2} leur écart-type respectif.

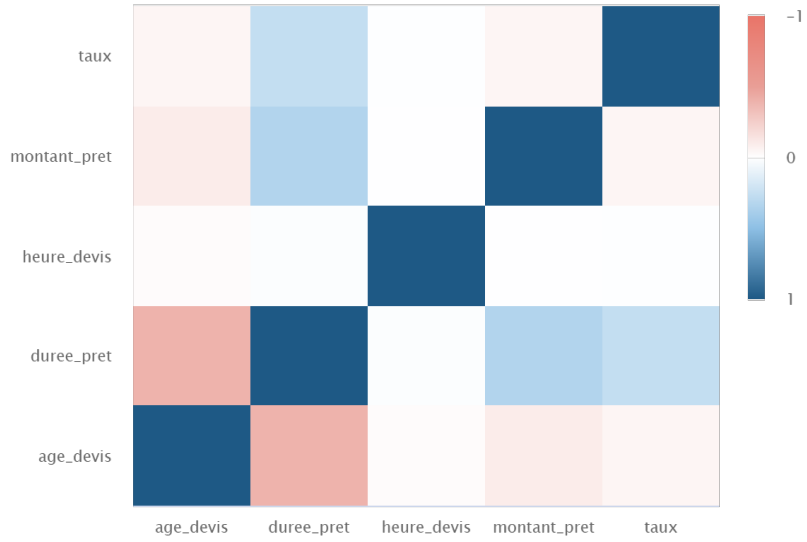

FIGURE 7.3 – Matrice de corrélations entre variables quantitatives

La corrélation de Pearson est comprise entre -1 et 1. Sur cette première matrice de corrélation, il est observable que les variables quantitatives sont peu corrélées entre elles.

Il est naturel que la durée du prêt ressorte corrélée négativement, dans une mesure acceptable, avec l'âge du devis, car les jeunes personnes peuvent souscrire à des prêts d'une plus longue durée en début de carrière, en comparaison à des personnes âgées.

Une faible corrélation positive se remarque également entre le taux, le montant et la durée du prêt. Un taux plus élevé est appliqué à des durées d'emprunt plus longues et un montant plus important nécessite une durée plus longue de remboursement.

Toutes ces variables caractéristiques sont gardées dans la suite du mémoire.

Corrélation des variables qualitatives

Pour l'étude des corrélations entre variables qualitatives, le coefficient de corrélation utilisé est celui du test de V *Cramer*, noté V, se définissant à partir du test du χ^2 :

$$V = \sqrt{\frac{\chi^2}{N \min(p_1 - 1; p_2 - 1)}}$$

Avec p_1, p_2 le nombre de modalités des variables respectives X_1, X_2, et N le nombre total d'observations. Le test de V Cramer à la particularité de rester stable lorsque la taille de l'échantillon est augmentée de manière proportionnelle entre les différentes modalités des variables.

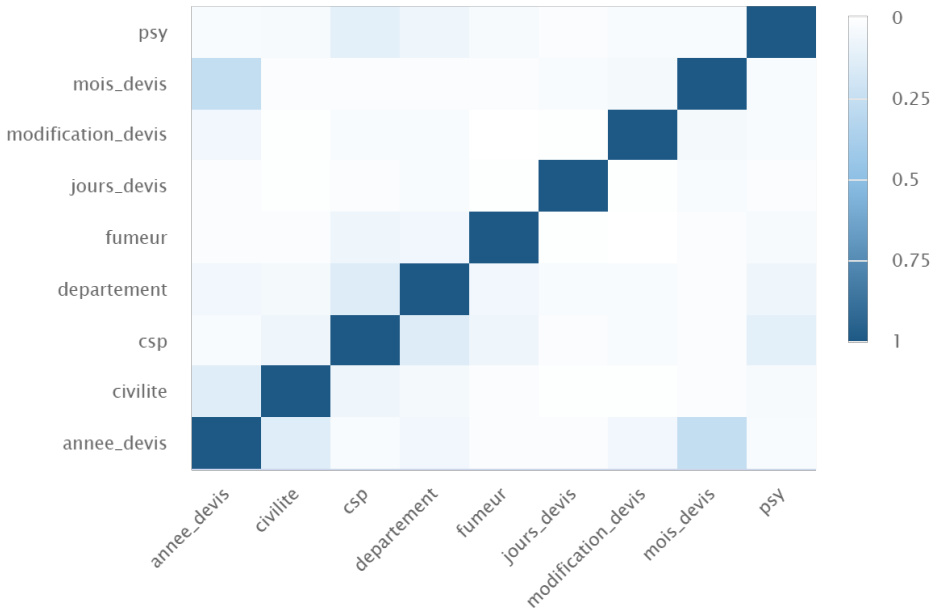

FIGURE 7.4 – Matrice de corrélations entre variables qualitatives

La corrélation de V *Cramer* est comprise entre 0 et 1. Sur cette seconde matrice de corrélations, il est observable que les variables qualitatives ne sont pas corrélées entre elles.

A la suite de cette étude des corrélations, aucune variable n'est écartée des données.

Chapitre 8

Statistiques descriptives des données

Après le retraitement des données, des statistiques préliminaires sont analysées pour mettre en évidence les premières relations entre les variables caractéristiques, obtenues à partir de la simulation des devis en ligne, et la variable cible Y, résultant de la fusion des bases de données devis et contrats. Sur l'ensemble du portefeuille de devis, le taux moyen de transformation de devis en contrats est de 10,2%.

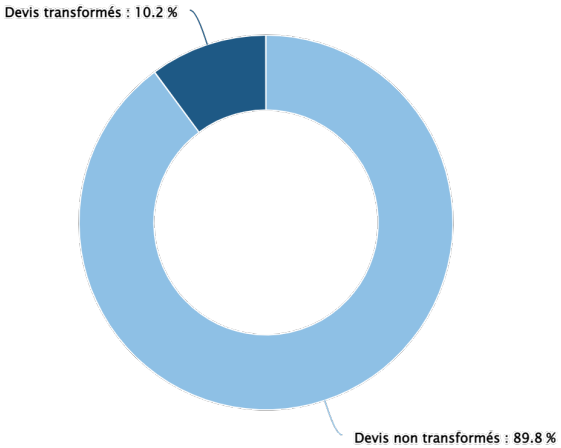

FIGURE 8.1 – Répartition de la variable cible Y du portefeuille

La corrélation entre les variables caractéristiques et la variable cible aide à la visualisation des variables les plus pertinentes et influentes. Selon les modèles utilisés dans la *partie V*, seules certaines variables seront sélectionnées pour obtenir de meilleurs résultats de prédiction. L'étude de ces corrélations est présentée ci-dessous.

1 Le taux de transformation selon les variables propres au devis

Dans un premier temps, la corrélation est étudiée pour les variables propres au devis. Les représentations graphiques ci-dessous représentent le taux de transformation sur l'échelle des ordonnées de gauche et le nombre de devis sur celle de droite, selon les différentes modalités prises par les variables. Le taux moyen du portefeuille est représenté par la ligne grise en pointillée.

Les interprétations de ces graphiques sont détaillées ci-dessous.

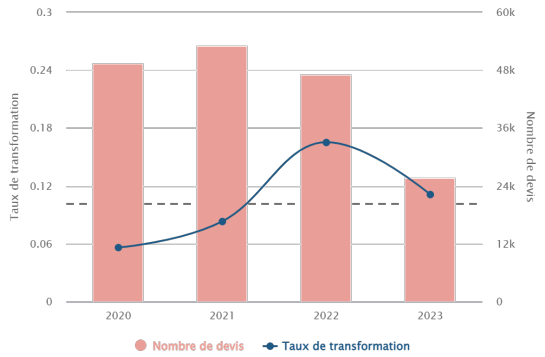

FIGURE 8.2 – Taux de transformation et nombre de devis par année

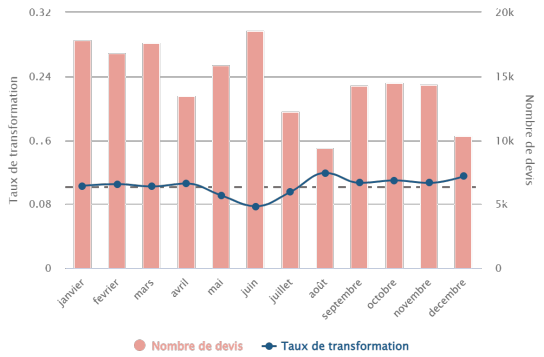

FIGURE 8.3 – Taux de transformation et nombre de devis par mois

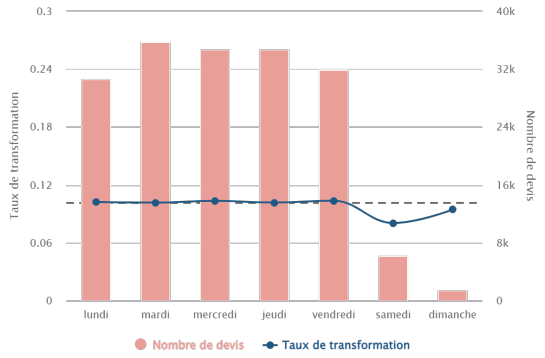

FIGURE 8.4 – Taux de transformation et nombre de devis par jour

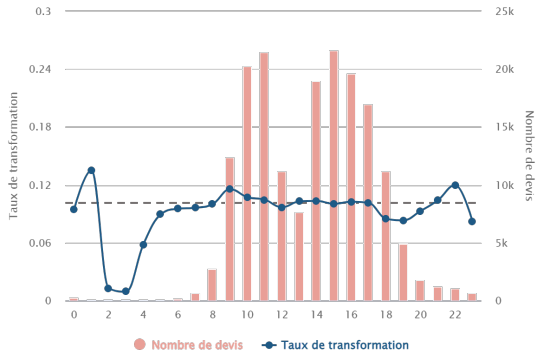

FIGURE 8.5 – Taux de transformation et nombre de devis par heure

Le mois de la réalisation du devis reflète une tendance : en milieu d'année, de mai à juillet, le taux de transformation est en dessous du taux de transformation des autres mois. Il remonte au mois d'août marqué par un nombre plus faible de devis, et ce également pour le mois de décembre. Une intention particulière pourrait éventuellement être accordée à ces mois d'août et de décembre par leur plus forte chance de voir le devis être souscrit en contrat.

Le taux de transformation ne semble pas être impacté par les jours de la semaine.

Également, ce taux ne semble pas être impacté par les heures, bien que d'importantes variations soient visibles sur la plage horaire de la nuit, dues à la variabilité du taux porté par un nombre quasi-nul de devis.

La comparaison des données pour l'historique 2020 à 2023 se révèle délicate en raison de multiples facteurs ayant impacté les taux d'intérêt. Les taux bas en 2020, le contexte économique marqué par la crise de la Covid-19 et la soudaine remontée des taux en 2022 ont engendré des dynamiques volatiles dans les taux d'emprunt et dans les comportements de consommation chez les ménages.

Les années 2020 et 2021 affectées par la crise Covid-19 marquent un plus faible taux de transformation pour un nombre quasiment stable de devis. Il est a noter que le nombre de devis de l'année 2023 est plus faible car les données sont arrêtées à fin juin. Une interprétation particulière de l'année 2022 est détaillée ci-dessous, permettant de trouver une explication à cette forte variation du taux de transformation. L'année 2022 a été bouleversée par la crise en Ukraine, les réglementations du taux d'usure et le début de la Loi Lemoine.

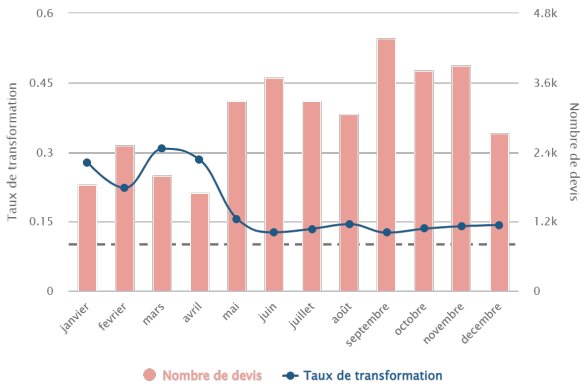

FIGURE 8.6 – Taux de transformation et nombre de devis en 2022

D'après le graphique ci-dessus 8.6, les mois de l'année 2022 sont régis par une forte variation du taux de transformation et du nombre de devis. Jusqu'au mois d'avril inclus, le taux moyen de transformation atteint 27% contre 13% à partir de mai, soit deux fois plus faible. Cette évolution semble être le reflet de la forte montée du taux d'intérêt. Pendant cette période, le taux d'usure n'a pas suivi la même évolution que le taux d'intérêt, ce qui à considérablement réduit l'écart entre ces deux taux, en confère la figure ci-après 8.7. Les taux d'usure indiqués en janvier et en mai 2022 sont respectivement de 1.10 et 1.50 ; tandis que les taux d'intérêts sont respectivement de 2.40 et de 2.41.

"Le relèvement du taux d'usure était réclamé par les professionnels de l'immobilier dans un contexte de remontée rapide des taux : certains emprunteurs, une fois ajoutés les frais de dossier et l'assurance au taux nominal, voyaient en effet leur taux annuel effectif global (TAEG) dépasser la barre fatidique. Ce qui annulait de fait leur crédit immobilier.", indique la Banque de France.[1]

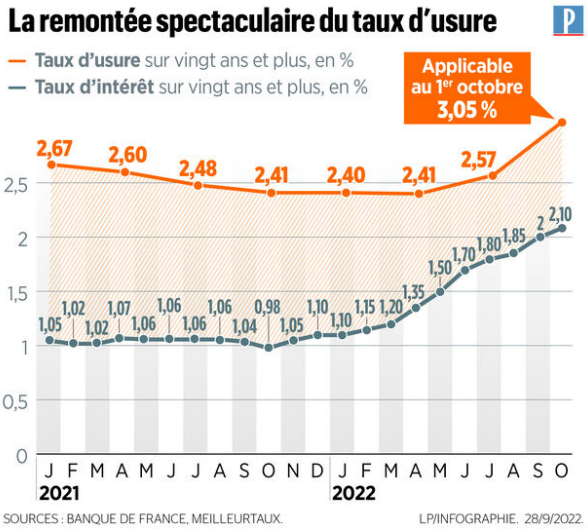

FIGURE 8.7 – Évolution du taux d'usure et du taux d'intérêt en 2022 et 2021 (*Source : Le Parisien, septembre 2022*)

Une fois la remontée du taux d'usure au dernier trimestre 2022, les ménages ont éventuellement pu redevenir actifs en simulant par exemple des devis sur des comparateurs en ligne. Cet effet semblerait expliquer l'augmentation du nombre de devis sur la deuxième période de l'année 2022 de la *figure* 8.6. L'entrée en vigueur de la loi Lemoine a également pu participer à la remontée du nombre de devis par la possibilité de changer d'assurance emprunteur.

1. Crédits immobiliers, le taux d'usure va repasser au-dessus de 3% au 1er octobre : Source 1, Le Parisien, septembre 2022

Enfin, un devis modifié est un devis qui a bien plus de chance de se transformer en contrat, bien que ces devis représentent moins de 20% du portefeuille :

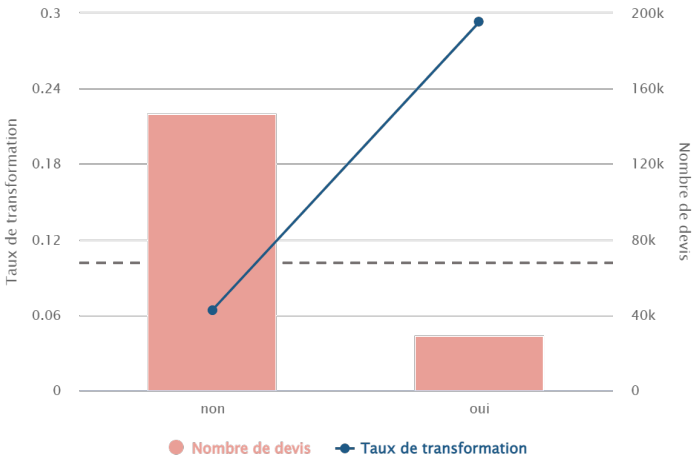

FIGURE 8.8 – Taux de transformation et nombre de devis selon la présence d'une modification du devis

2 Le taux de transformation selon les variables propres à la personne réalisant le devis

Dans un deuxième temps, la corrélation est étudiée pour les variables propres à l'utilisateur.

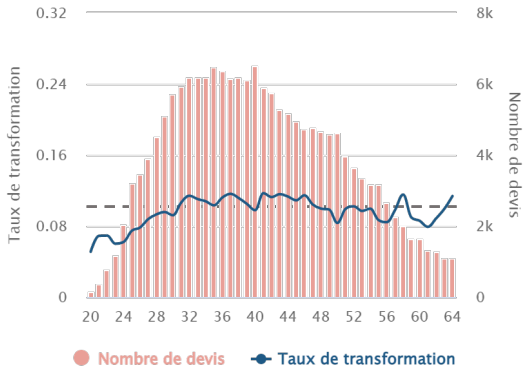

FIGURE 8.9 – Taux de transformation et nombre de devis par âge lors du devis

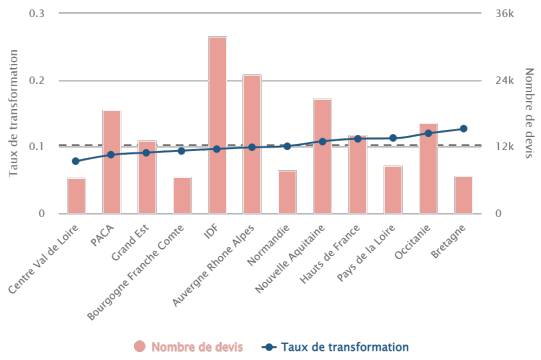

FIGURE 8.10 – Taux de transformation et nombre de devis par région

Les personnes dont l'âge est compris entre 30 et 45 ans sont les plus susceptibles de transformer leur devis en contrat. Cette tendance peut s'expliquer probablement par leur maturité financière et leur stabilité professionnelle, qui les rendent plus à même de prendre des engagements immobiliers.

Au travers des régions de France, le taux de transformation minimal est égal à 7,8% pour la région Centre-Val-de-Loire et maximal égal à 12,8% pour la région Bretagne.

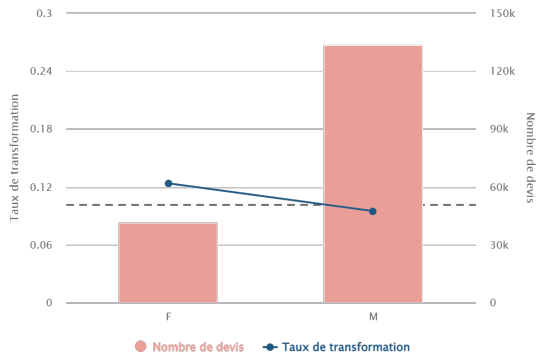

FIGURE 8.11 – Taux de transformation et nombre de devis par civilité

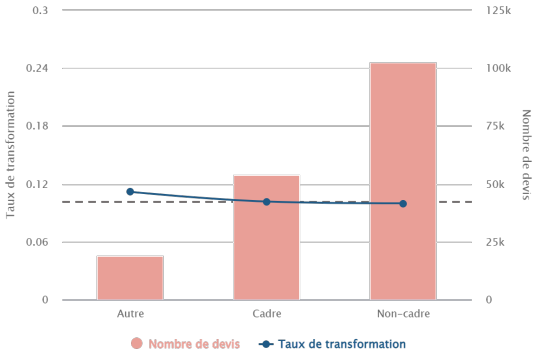

FIGURE 8.12 – Taux de transformation et nombre de devis par catégorie socio-professionnelle

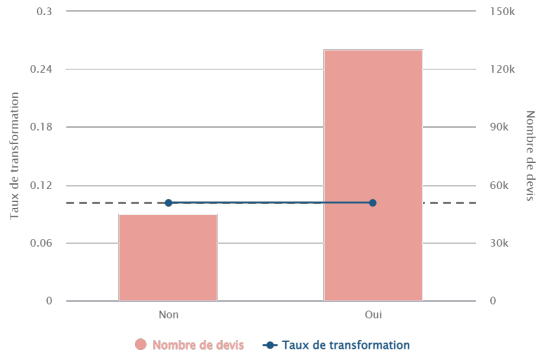

FIGURE 8.13 – Taux de transformation et nombre de devis selon la prise de garantie MNO

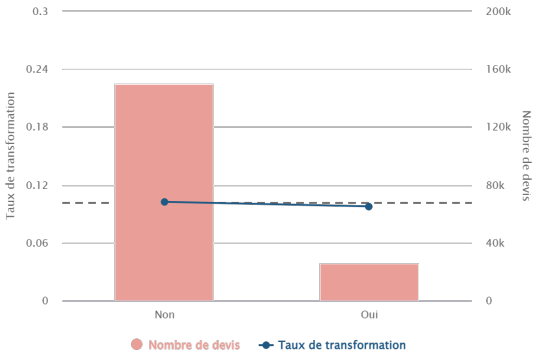

FIGURE 8.14 – Taux de transformation et nombre de devis selon la distinction Fumeur / Non fumeur

Au regard des graphiques ci-dessus, les modalités des variables caractéristiques de l'utilisateur comportent peu de disparités entre elles.

Toutefois, une tendance plus élevée à transformer son contrat est visible pour la modalité féminine.

Pour les autres variables : catégorie socio-professionnel, indicateur de santé et la distinction Fumeur/Non fumeur, le taux moyen de transformation est stable selon les modalités. Ces variables n'affichent pas d'impact significatif sur la réussite de la transformation du devis en contrat.

A priori, les résultats des statistiques descriptives de l'utilisateur ont peu d'influences, mais l'application futures des modèles confirmeront ou non ce constat par une étude plus poussée des données.

Ce constat pourrait éventuellement être à l'origine d'hypothèses suivantes :

— Une collecte limitée d'informations. Il est possible que le comparateur en ligne ne collecte qu'un nombre limité d'informations sur les utilisateurs et les variables disponibles dans ce portefeuille ne captureraient pas les facteurs les plus déterminants qui influencent la transformation des devis en contrats.

— Des facteurs externes. Il est également possible que des facteurs externes, tels que la conjoncture économique, la concurrence sur le marché ou la sensibilité aux prix, aient une plus grande incidence sur la décision de l'utilisateur à souscrire ou non un contrat, plutôt que ses caractéristiques individuelles.

3 Le taux de transformation selon les variables propres au crédit immobilier

Dans un dernier temps, la corrélation est étudiée pour les variables propres au crédit immobilier.

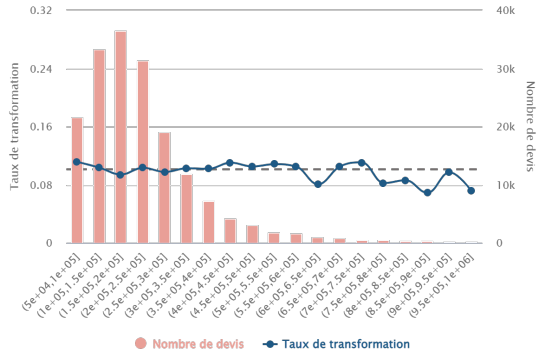

FIGURE 8.15 – Taux de transformation et nombre de devis selon le montant du crédit

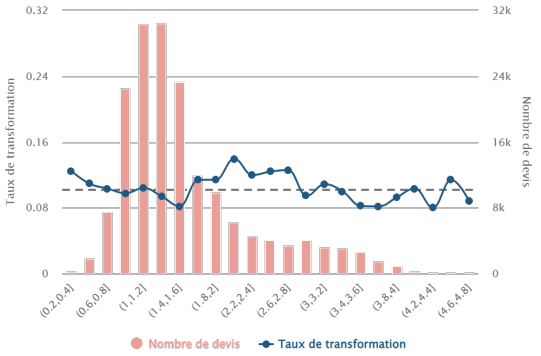

FIGURE 8.16 – Taux de transformation et nombre de devis selon le taux du crédit

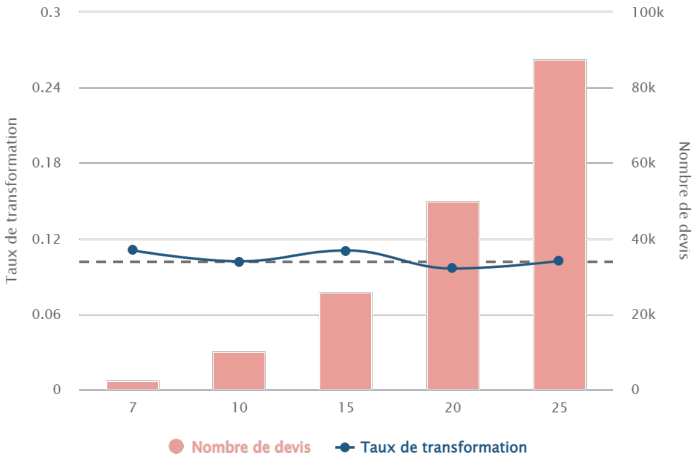

FIGURE 8.17 – Taux de transformation et nombre de devis selon la durée du crédit

Dans les graphiques ci-dessus, le montant du devis est représenté par tranche de 50 000 € et le taux de crédit par tranche de 0.20%. Le choix de ces tranches est ici à titre indicatif pour représenter les données. Le montant et le taux du crédit ne révèlent pas de variations significatives en ce qui concerne la transformation du devis en contrat. Il est donc possible que les utilisateurs du comparateur en ligne présentent des caractéristiques similaires en termes de besoins ou de préférences d'achats immobiliers.

Toutefois, les graphiques mettent en évidence une tendance d'affaiblissement pour des montants élevés, à la différence des taux situés entre 1,6% et 2,8% qui affichent les meilleures transformations.

Les durées des crédits ne reflètent quant à elles pas de tendance distinctive.

→ Finalement, certaines modalités des variables présentées seront amenées à être regroupées dans la partie suivante pour faciliter l'implémentation des modèles sur R.

Cinquième partie

Application des méthodes d'apprentissage statistique à la sélection de devis

Chapitre 9

Préparation spécifique des données et initialisation des paramètres

Dans ce chapitre, les données sont préparées et les modèles précédemment décrits sont calibrés pour déterminer les profils types des prospects les plus ou moins susceptibles de transformer leur devis en contrat.

La classification ascendante hiérarchique nécessite de calculer la distance entre tous les individus. Ce calcul se réalise par la fonction $daisy()$ du *package Cluster* sur R, et renvoie la matrice des distances. En raison du nombre important de devis étudiés, le stockage en mémoire de cette fonction devient rapidement saturée, ne pouvant traiter approximativement que $30\,000$ données en entrée.

Pour tenir compte de cette contrainte, ce chapitre présente tout d'abord l'agrégation de certaines variables caractéristiques.

Ce chapitre présente également plusieurs plusieurs applications possibles des modèles selon différents traitements de la données d'entrée, appelés dans la suite scénarios. Ils sont au nombre de trois.

Enfin, ce chapitre présente l'initialisation des paramètres des deux nouvelles approches de la classification ascendante hiérarchique : la CAH1 et la CAH2. Une fois les données prêtes à l'emploi et les paramètres calculés, le chapitre suivant présentara l'implémentation de ces algorithmes.

Avant de poursuivre, il est introduit :

— la base d'apprentissage, composé à 70% de la base initiale, comportant $122\,599$ devis
— la base de test, composé à 30% de la base initiale, comportant $52\,543$ devis

1 Agrégation de la base de données par regroupement des modalités

Dans l'objectif de traiter l'ensemble des devis, une alternative est proposée. Elle évite de devoir tirer aléatoirement 30 000 devis parmi les 122 599 présents dans la base d'apprentissage. L'idée consiste à agréger les devis qui présentent des caractéristiques identiques, moyennant la variable cible. Cette agrégation diminue le nombre de devis.

Toutefois, au regard des statistiques descriptives précédentes, il existe un très grand nombre de modalités dans l'ensemble des données, notamment induit par la présence de variables continues. L'agrégation effectuée sur un tel jeu de données ne réduit que très peu le nombre de devis. L'enjeu consiste à regrouper de façon pertinente certaines modalités entre elles, permettant d'offrir une agrégation plus efficace tout en essayant de ne pas biaiser les résultats de l'analyse.

Le suivant regroupement, accompagné de l'agrégation des données, va permettre de capturer l'ensemble des devis présents dans la base de données. Il s'effectue par trois approches indépendantes et en fonction des variables :

1. Regroupement des variables catégorielles : $mois\_devis$, $jours\_devis$, $heure\_devis$ et $region$, en nouvelles modalités basées sur les premières statistiques descriptives
2. Catégorisation des variables continues, age_devis et $montant\_devis$, à l'aide d'un algorithme CART
3. Transformation de la variable $taux\_devis$ basée sur l'historique de taux d'un comparateur autre que celui de la base de données

1.1 Regroupement des modalités d'après les statistiques descriptives

Le premier regroupement est appliqué sur les variables catégorielles, principalement celles comportant un grand nombre de modalités. Il s'effectue de telle manière à garder la fidélité du taux de transformation initialement présenté par les statistiques descriptives du chapitre 8.

La variable mois_devis

D'après les statistiques de la figure 8.3, la variable $mois\_devis$ est regroupée selon trois nouvelles modalités :

— Quadrimestre 1 : de janvier à avril
— Quadrimestre 2 : de mai à août
— Quadrimestre 3 : de septembre à décembre

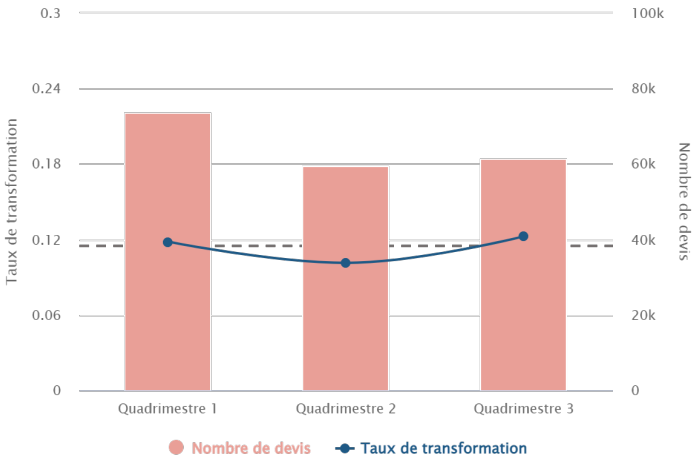

FIGURE 9.1 – Statistiques de la variable $mois\_devis$ après regroupement

La variable jours_devis

D'après les statistiques de la figure 8.4, la variable $jours\_devis$ est regroupée selon deux nouvelles modalités :

— Semaine : de lundi à vendredi
— Week-end : samedi et dimanche

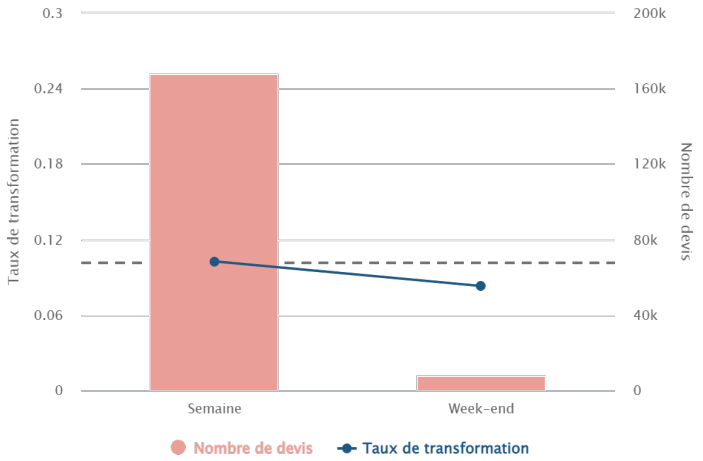

FIGURE 9.2 – Statistiques de la variable $jours\_devis$ après regroupement

La variable heure_devis

D'après les statistiques de la figure 8.5, la variable $heure\_devis$ est regroupée selon deux nouvelles modalités :

— Horaire travail : de 8h00 à 11h59 et de 14h00 à 17h59
Cette plage horaire comprend les horaires classiques de travail en France, tout en permettant de respecter la fidélité du taux de transformation
— Horaire libre : de 12h00 à 13h59 et de 18h00 à 7h59
Cette plage horaire comprend les horaires "de temps libre" où les utilisateurs travaillent généralement pas

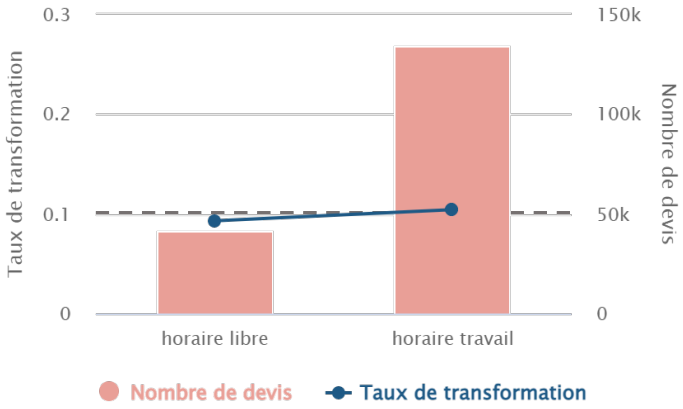

FIGURE 9.3 – Statistiques de la variable $heure\_devis$ après regroupement

La variable région

D'après les statistiques de la figure 8.10, la variable $region$ est naturellement regroupée selon trois nouvelles modalités, reflétant un taux de transformation suffisamment distinct :

— Région 1 : Bretagne, Pays-de-la-Loire, Occitanie, Hauts-de-France et Nouvelle-Aquitaine
— Région 2 : Ile-de-France, Auvergne-Rhône-Alpes et Normandie
— Région 3 : Provences-Alpes-Côtes-d'Azur, Centre-Val-de-Loire, Grand-Est et Bourgogne-Franche-Comte

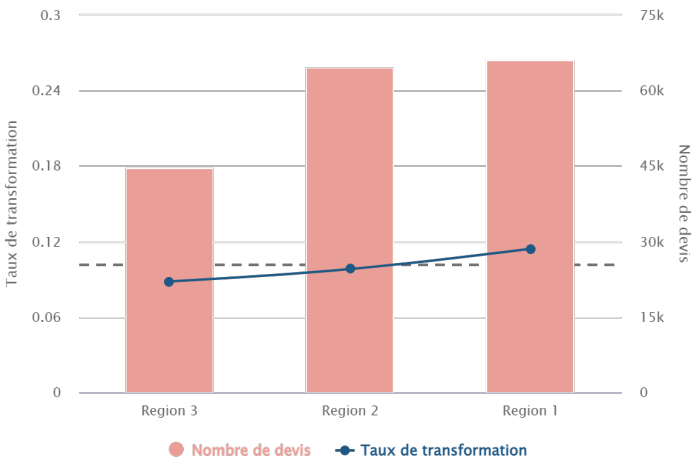

FIGURE 9.4 – Statistiques de la variable *region* après regroupement

1.2 Catégorisation des variables continues par arbre CART

Le deuxième type de regroupement s'effectue par une catégorisation des variables continues à l'aide d'un arbre CART : la variable cible Y est uniquement prédite par la variable caractéristique à catégoriser. Cette approche consiste à découper la variable continue en intervalles distincts. Ces intervalles représentent les catégories de la variable continue, en identifiant les valeurs qui présentent des similitudes en termes de taux de transformation.

Les arbres sont obtenus par la fonction $rpart()$ sur R par le *package RPART*, en précisant une profondeur maximale à 4 niveaux pour obtenir un nombre restreint de regroupements. Les arbres sont représentés ci-dessous :

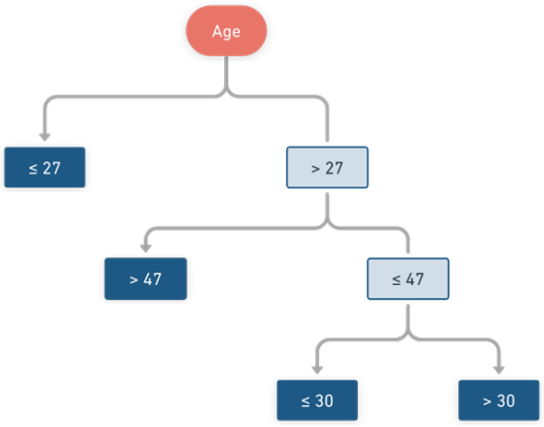

FIGURE 9.5 – Représentation de l'arbre obtenu pour la variable *age_devis*

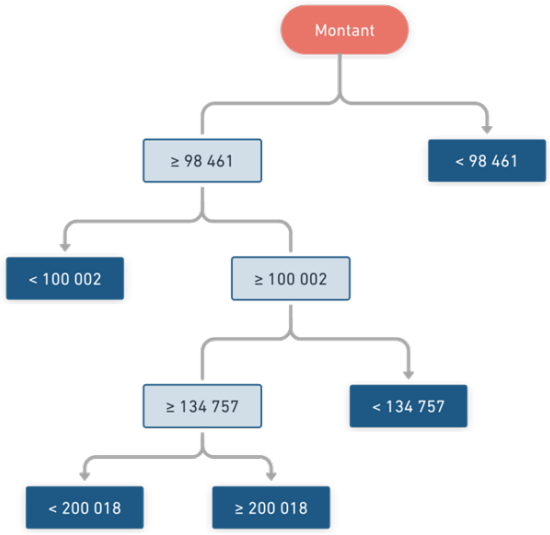

FIGURE 9.6 – Représentation de l'arbre obtenu pour la variable *montant_devis*

La variable age_devis

D'après les résultats du premier arbre CART 9.5, la variable *age_devis* est répartie en quatre intervalles comme suit :

- de 20 à 27 ans
- de 28 à 30 ans
- de 31 à 47 ans
- de 48 à 64 ans

La variable montant_devis

D'après les résultats du second arbre CART 9.6, la variable *montant_devis* est répartie en cinq intervalles comme suit :

— Montant 5 : de 50 000 à 98 460 €
— Montant 4 : de 98 461 à 100 001 €
— Montant 3 : de 100 002 à 134 756 €
— Montant 2 : de 134 757 à 200 017 €
— Montant 1 : de 200 018 à 1 000 000 €

Les taux de transformation selon ces nouveaux regroupements sont représentés par les graphiques suivants :

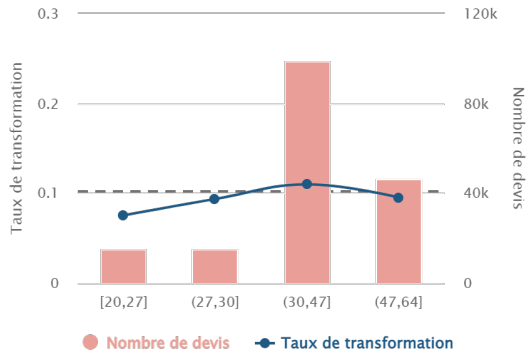

FIGURE 9.7 – Statistiques de la variable *age_devis* après regroupement

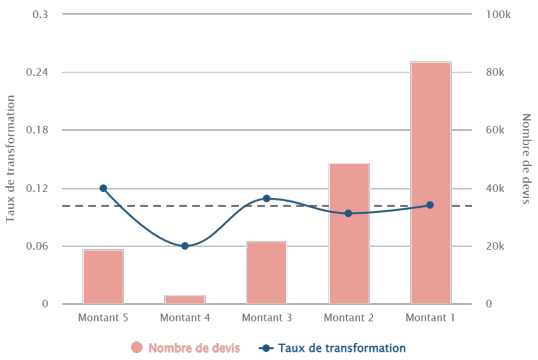

FIGURE 9.8 – Statistiques de la variable *montant_devis* après regroupement

1.3 Transformation du taux de l'emprunt par indicateur du marché

Le dernier regroupement concerne la variable *taux*. Cette variable est transformée en trois classes de taux : "faible", "moyen" et "élevé" qui tiennent compte de la date des devis et des conditions financières à l'aide des données historiques de taux précisées par un comparateur d'assurances emprunteur. L'historique de ces taux fixes (à l'échelle nationale), est disponible pour les années 2020 à 2023 et pour des durées d'emprunt de 7, 10, 15, 20 et 25 ans.[1]

A titre d'information, voici la présentation des taux du comparateur pour une durée d'emprutn de 25 ans :

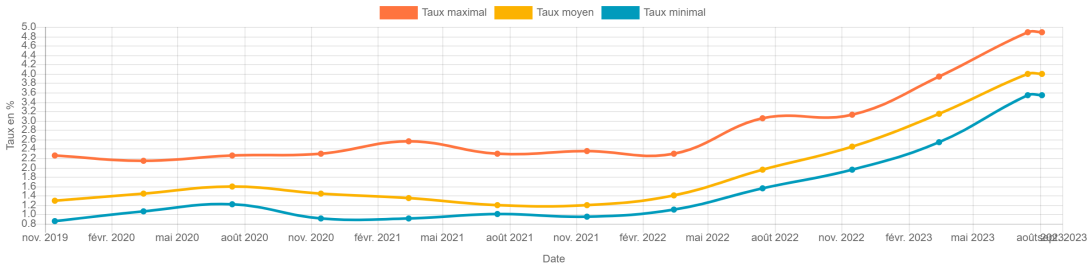

FIGURE 9.9 – Taux minimal, moyen et maximal du comparateur (25 ans)

1. Historique des taux (2020 à 2023) : Empruntis, historique des taux d'intérêt du crédit immobilier, dernière mise à jour à date du 01.08.2023

Les taux immobiliers sont déterminés en fonction du taux directeur, fixé par la Banque Centrale Européenne (BCE) et des Obligations Assimilables du Trésor (OAT) sur 10 ans. D'autres éléments sont également pris en compte, ce qui explique la variabilité de ces taux : la durée du prêt et les objectifs commerciaux de la banque, selon la concurrence et le contexte économique et financier.

Avant de procéder au découpage de la variable $taux$, le comportement des taux de la base de données du mémoire est comparé à celui des taux du comparateur, nommés par la suite taux de "référence". Pour chaque trimestre, de 2020 à 2023, les graphiques ci-dessous présentent les taux moyens du portefeuille et ceux du comparateur :

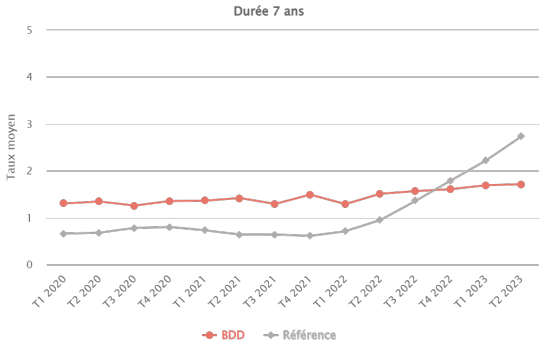

FIGURE 9.10 – Comparaison des taux moyens d'emprunts (7 ans)

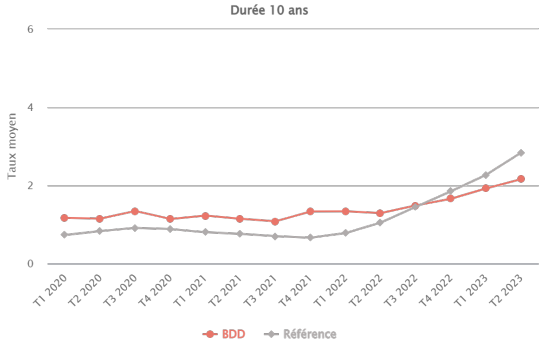

FIGURE 9.11 – Comparaison des taux moyens d'emprunts (10 ans)

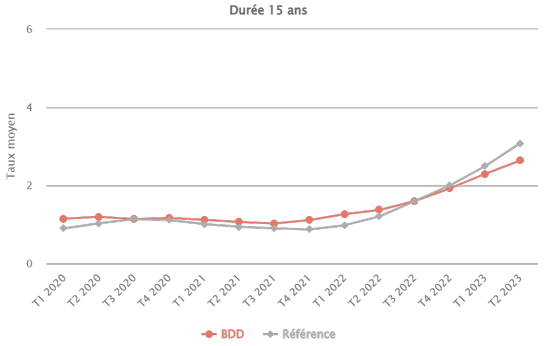

FIGURE 9.12 – Comparaison des taux moyens d'emprunts (15 ans)

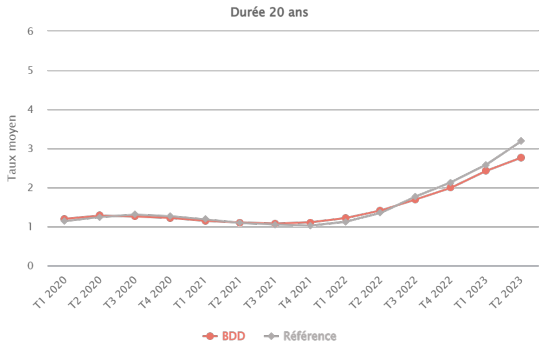

FIGURE 9.13 – Comparaison des taux moyens d'emprunts (20 ans)

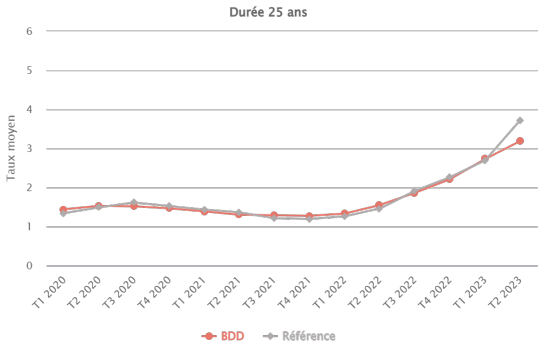

FIGURE 9.14 – Comparaison des taux moyens d'emprunts (25 ans)

Les courbes permettent de constater que pour les emprunts sur 7 et 10 ans, les taux moyens de la base de données sont supérieurs à ceux du comparateur. Cependant la tendance s'inverse à partir du quatrième trimestre 2022. Néanmoins, pour les durées d'emprunt de 15, 20 et 25 ans, les courbes sont très proches jusqu'au premier trimestre 2023.

Pour chacun des trimestres de l'historique des taux du comparateur (du premier trimestre 2020 au deuxième trimestre 2023), un découpage peut être effectué comme suit :

1. Création de deux intervalles : I_{inf} = [taux minimal ; taux moyen] et I_{sup} = [taux moyen ; taux maximal]
2. Découpage respectif de I_{inf} et I_{sup} en 3 sous-intervalles équidistants
3. Formation de trois classes : les taux faibles regroupent les deux intervalles inférieurs, les taux moyens regroupent les deux intervalles centraux, et les taux élevés regroupent les deux intervalles supérieurs

Ce processus peut s'illustrer par le schéma suivant :

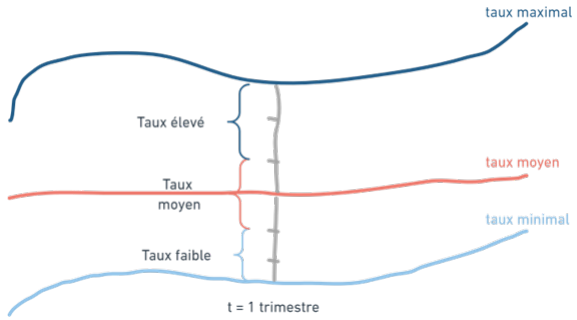

FIGURE 9.15 – Principe du découpage des taux pour création d'intervalles

La variable taux

D'après ce calibrage, les valeurs de la variable *taux* du portefeuille sont affiliées à leurs intervalles correspondant, par une jointure selon le trimestre, l'année et la durée du prêt. Cette variable est ainsi regroupée en trois modalités :

— Taux faible
— Taux moyen
— Taux élevé

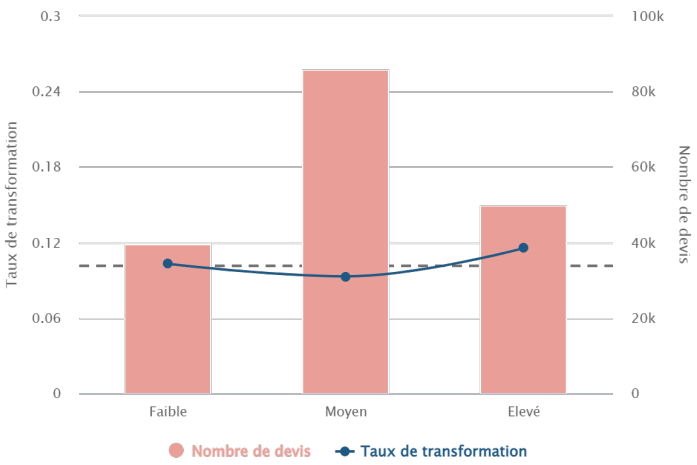

FIGURE 9.16 – Statistiques de la variable *taux* après regroupement

→ Une fois ces regroupements effectués et après agrégation de la base de données, la base d'apprentissage comportent 56 151 devis, contre 122 599 avant regroupement. Par la construction des scénarios, ce nombre de devis sera encore réduit.

2 Représentation des données en dimension réduite

Avant de construire les scénarios, il peut être intéressant de visualiser les devis dans un plan afin de voir si le regroupement a eu une impact sur leurs proximités.

La représentation des données dans un espace réduit à deux dimensions oblige l'application d'une technique de réduction de dimension. Celle choisie est la *t-Distributed Stochastic Neighbor Embedding* (t-SNE), introduit par M. Van Der Maaten et M. Hinton en 2008 [9]. Cette technique a la particularité d'être non linéaire. Cela signifie qu'elle peut permettre de séparer les données qui ne peuvent pas être séparées par une ligne droite. Cette technique est non déterministe : elle consiste à créer une distribution de probabilité qui représente les similarités entre observations voisines dans l'espace initial, de grande dimension (appelé l'espace des données), et une distribution de probabilités dans l'espace de dimension réduite (appelé espace d'observation). Cet algorithme se découpe en 3 grandes étapes, synthétisées en annexes (2).

Voici la représentation t-SNE des devis appliquée sur la base de données avant le précédent regroupement :

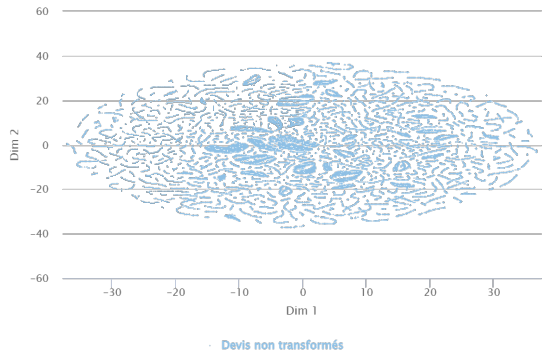

FIGURE 9.17 – Représentation t-SNE des devis non transformés

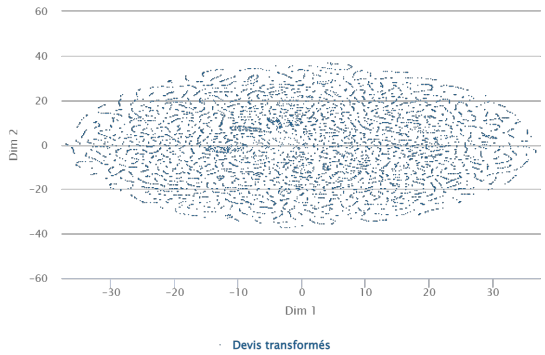

FIGURE 9.18 – Représentation t-SNE des devis transformés

Cette représentation dans un espace en deux dimensions, avec d'un coté les devis non transformés en contrat 9.17, et de l'autre ceux transformés en contrat 9.18, met en évidence la répartition homogène dans tout l'espace des deux modalités de la variable Y. Il n'y a pas de regroupement visible qui soit plus fort pour l'une des deux modalités.

Les graphiques suivants donnent la représentation t-SNE des devis appliquée à la base de données après le précédent regroupement des variables :

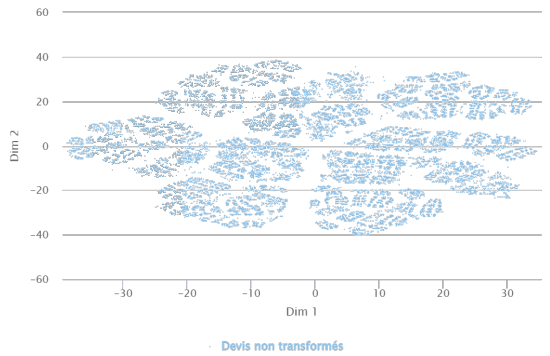

FIGURE 9.19 – Représentation t-SNE des devis non transformés après regroupement

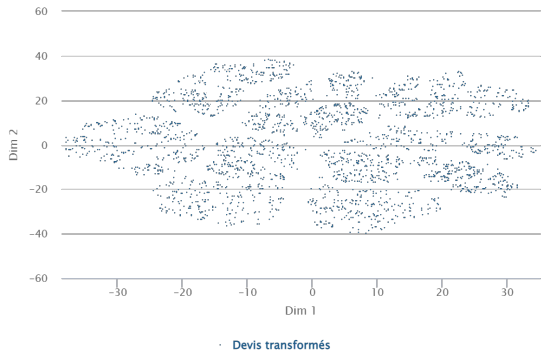

FIGURE 9.20 – Représentation t-SNE des devis transformés après regroupement

A la différence de la précédente représentation, celle-ci montre des regroupements distincts, dus à l'effet du regroupement des modalités des variables numériques, dorénavant traitées en modalités catégorielles.

$\rightarrow$ Ainsi, cette représentation permet de visualiser la localité des devis à partir de leurs caractéristiques X. La répartition des modalités de la variable cible Y est placée aléatoirement dans tout l'espace et ne présente pas de caractère particulier. De plus, le regroupement a globalement conservé la structure de la base de données brutes.

3 Construction des scénarios

Le regroupement des données a permis de diminuer efficacement le nombre de devis de la base d'apprentissage sans pour autant la détériorée. Toutefois, ce nombre est encore au dessus de la limite induite par la contrainte mémoire de la fonction $daisy()$. La construction des scénarios suivants va permettre d'atteindre le nombre de devis souhaité.

3.1 Le scénario 1 : l'initial

Le premier scénario correspond au scénario initial. Pour permettre l'utilisation de la fonction $daisy()$, toutes les variables ne sont pas sélectionnées. Par conséquent, le scénario 1 sélectionne l'ensemble des variables caractéristiques présentées dans la partie IV en excluant $annee$ et $modification\_devis$.

Le choix d'exclure ces deux variables permet de se concentrer uniquement sur les caractéristiques intrinsèques du devis, indépendamment du contexte temporel (pris en compte par ailleurs dans la construction de la variable taux) ou des modifications ultérieures. En effet, l'objectif du *clustering* est de pouvoir classer chaque nouveau devis. Or, l'année de la réalisation dépend des conditions passées et ne correspond pas à une caractéristique d'un nouveau devis. Également, le fait de savoir si un utilisateur à modifié son devis ou pas, ne donne pas d'information supplémentaire : lors de la réalisation du devis, il est directement attribué a la modalité "non modifié".

Après agrégation, la base d'entraînement du scénario 1 comporte $29\ 151$ devis.

3.2 Le scénario 2 : le rééquilibrage des données

Le nombre de devis transformés en contrats représente seulement 10.2% de l'ensemble des devis, ce qui engendre donc un fort déséquilibre ; c'est pourquoi ce scénario est étudié.

A noter qu'un rééquilibrage de la variable cible est principalement utilisé dans le contexte de la classification, où l'objectif est de prédire des étiquettes de classe. Dans le contexte du *clustering*, l'objectif est de regrouper les données en sous-groupes similaires. Or, les nouvelles approches prennent en compte la variable cible et il peut être intéressant de voir l'impact qu'aurait un rééquilibrage sur l'apprentissage des données.

En général, le rééquilibrage des données s'effectue par la combinaison d'une approche de sur-échantillonnage (*oversampling*) de la classe minoritaire et d'une approche de sous-échantillonnage de la classe majoritaire (*undersampling*). Ces deux méthodes sont expliquées en annexes (3).

Ainsi, le deuxième scénario est similaire au scénario 1 à l'exception du taux de transformation de devis en contrat calibré à 30% par la combinaison d'un *SMOTE-NC* et d'un tirage aléatoire.

Après agrégation, la base d'entraînement du scénario 2 comporte $28\ 298$ devis.

3.3 Le scénario 3 : la sélection des variables

La construction du dernier scénario permet d'étudier la sensibilité des modèles suite à la modification des variables caractéristiques prises en compte.

Au regard des statistiques descriptives, la variable $modification\_devis$ joue un rôle important quant à l'indication sur la transformation du devis en contrat. Elle apporte la plus grande disparité du taux de transformation entre ses modalités, parmi l'ensemble des variables disponibles. Par ailleurs, la moyenne du portefeuille de la durée de modification, durée entre la date de la réalisation du devis et la date de la modification du devis est de $2,75$ mois. Cette durée peut signifier que l'utilisateur modifie assez rapidement son devis à la suite de sa création. L'inclusion la variable $modification\_devis$ peut donc s'avérer pertinente à inclure dans le modèle et est sélectionnée au détriment de la variable $fumeur$, qui apparaît comme la moins importante d'après la méthode PFI (ci-après 9.25), et ce toujours dans le but de pouvoir garder un nombre de devis proche de $30\,000$. (A noter que la variable $annee$ n'est également pas prise en compte.)

Après agrégation, la base d'entraînement du scénario 3 comporte $30\,133$ devis.

4 Initialisation des paramètres

L'application des deux nouvelles approches, CAH1 et CAH2, nécessite l'initialisation de paramètres spécifiques, non requis pour l'approche classique de la CAH0.

4.1 CAH1 : détermination du paramètre α

Le paramètre α joue le rôle de poids dans l'importance attribuée aux matrices de dissimilarités D_1 et D_2, formant la matrice de dissimilarité $D_\alpha = \alpha D_1 + (1-\alpha)D_2$, utilisée pour le modèle de la CAH1. Pour un α proche de 1, la matrice D_α se rapproche de celle utilisée pour la CAH0.

La détermination de ce paramètre repose sur la corrélation cophénétique, proposée par Mr. Sokal et Mr. Rohlf en 1962 [3]. Elle évalue la fidélité du dendrogramme par rapport aux distances initiales des devis. Autrement dit, la corrélation cophénétique notée $\mathcal{C}$, (étudiée plus en détail au chapitre 1), résulte :

1. des distances initiales, avant la classification, renseignées par les matrices des distances D_1 et D_2 et calculées à l'aide la distance de Gower ,
2. des distances cophénétiques, issues de la matrice cophénétique D_α^{Coph} à la sortie de la classification. Ces distances correspondent à celles de la réunion de deux devis qui rejoignent un même groupe, indiquées par le dendrogramme.

Trouver la valeur optimale de α revient à trouver un compromis entre les informations contenues dans la matrice D_1 et D_2. La détermination du paramètre α est donc le résultat de l'optimisation de la fonction suivante, notée $\mathcal{C}_\alpha$, qui "équilibre" le poids entre D_1 et D_2 :

$$\mathcal{C}_\alpha = |\mathcal{C}(D_\alpha^{\text{Coph}}, D_1) - \mathcal{C}(D_\alpha^{\text{Coph}}, D_2)| \tag{9.1}$$

Avec D_α^{Coph}, la matrice cophénétique obtenue à partir du dendrogramme issue de la CAH1 pour une valeur fixée de α. Le critère $\mathcal{C}_\alpha$ représente donc la différence en valeur absolue entre deux corrélations. L'une mesure la fidélité entre les distances du dendrogramme (indiquée dans la matrice D_α^{Coph}) et les distances initiales des devis indiquée dans la matrice D_1. L'autre mesure de même la corrélation entre D_α^{Coph} et la matrice D_2.

La valeur de α s'obtient par : $\hat{\alpha} = \operatorname{argmin} \mathcal{C}_\alpha$.

Le processus d'optimisation du paramètre α peut être illustré par le schéma ci-dessous :

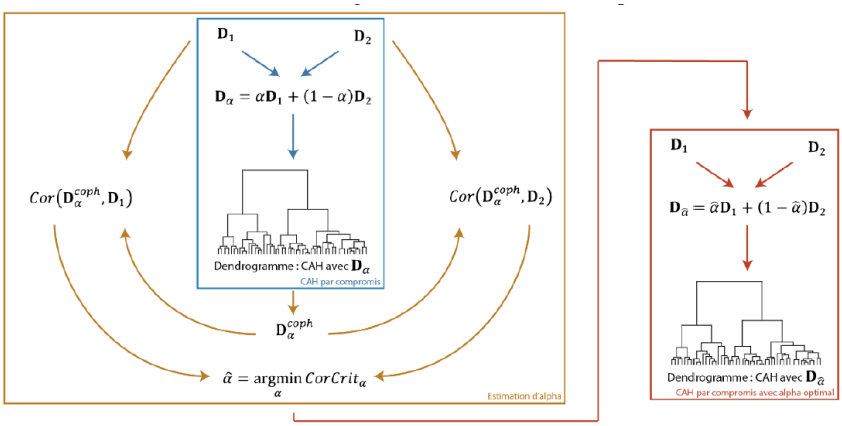

FIGURE 9.21 – Illustration du processus d'optimisation du paramètre α

Application aux données

Ce paramètre est calculé pour chacun des scénarios via la fonction $hclustcompro\_select\_alpha()$ du *package cluster* sur R. Cette fonction prend en paramètres les deux matrices de dissimilarités D_1 et D_2.

	Scénario 1	Scénario 2	Scénario 3
Valeur α	0.85	0.88	0.86

FIGURE 9.22 – Valeurs retenues pour le paramètre α

Les valeurs retenues correspondent à la moyenne de la valeur α obtenue sur 5 échantillons d'un tirage aléatoire issu de 10 *bootstrapping* sur 100 devis.

4.2 CAH 2 : détermination du poids des variables

La détermination des poids attribués aux variables est réalisée par la méthode PFI, précédemment introduite en détail : 1.1. Cette méthode est implémentée sur le logiciel R par le *package DALEX*. La méthode PFI est appliqué à l'ensemble des données de la base d'apprentissage, exceptée la variable $annee\_devis$. Elle est calculée à l'issue des algorithmes *Random Forest* et *XGBoost*. Afin de retenir les importances issues du meilleur modèle, il faut comparer leur prédiction sur la base de test.

Sélection du meilleur modèle

Dans le jeu de données étudié, la variable Y est binaire. Il s'agit ainsi d'un problème de classification dont les métriques de référence sont :

- La courbe ROC (*Receiver Operating Characteristic*), qui représente le taux de vrais positifs (le modèle qui prédit correctement un devis transformé en contrat), en fonction du taux de faux positifs (le modèle qui prédit à tort un devis transformé en contrat alors qu'il est en réalité non transformé), pour différents seuils de classification (valeur critique utilisée pour séparer les classes de prédiction positives des classes négatives) ;
- L'AUC-ROC (*Area Under the Curve*), qui représente la surface sous la courbe ROC. Plus l'AUC est élevée, meilleure est la performance du modèle pour distinguer les devis transformés des devis non transformés.

Les courbes ROC obtenues pour les deux modèles sont les suivantes :

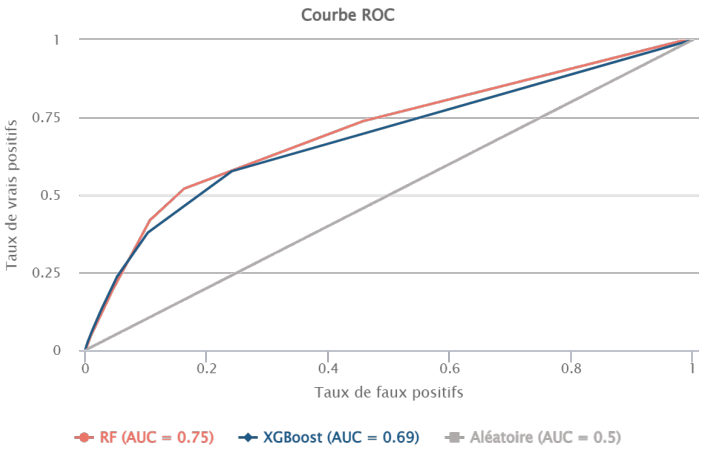

FIGURE 9.23 – Courbe ROC des modèles *Random Forest* et *XGBoost* après PFI

Il est visible que la courbe ROC du modèle *Random Forest* à tendance à être légèrement au-dessus ou confondue à celle du modèle *XGBoost* pour l'ensemble des seuils de classification. L'AUC du modèle *Random Forest* est égale à 0.75 et est supérieure à l'AUC du modèle *XGBoost* égale à 0.69. Pour le modèle *Random Forest*, la probabilité que deux devis provenant de deux classes distinctes est juste dans 3 cas sur 4.

Toutefois, l'AUC seule ne permet pas de juger la pertinence du modèle en raison du caractère "déséquilibré" de la base d'apprentissage. Les métriques suivantes (comprises entre 0 et 1) permettent également d'évaluer la pertinence des modèles :

- Le rappel évalue la capacité du modèle à capturer les devis transformés.
- La précision évalue la capacité du modèle à éviter de prédire à tord des devis transformés.
- La F1 Score évalue la performance globale du modèle en tenant compte à la fois de sa capacité à effectuer des prédictions précises (précision) et de sa capacité à capturer un maximum de devis transformés (rappel).

Ces métriques seront formellement détaillées lors du chapitre dédié 2 sur la prédiction finale des modèles.

Le tableau suivant présente le résultat de ces métriques (et de l'AUC) pour les deux modèles étudiés :

	XGBoost	RF
Rappel	0.30	**0.48**
Précision	0.31	0.29
F1-Score	0.30	**0.36**
AUC	0.69	0.75

FIGURE 9.24 – Comparaison des métriques pour sélection du modèle

Les métriques obtenues sont relativement du même ordre de grandeur. Néanmoins le rappel et le F1-Score sont respectivement 1.6 et 1.2 fois plus grand pour l'algorithme *Random Forest* que pour l'algorithme *XGBoost*. Par conséquent, l'algorithme *Random Forest* a tendance à mieux reconnaître les devis qui se transforment en contrat. C'est la raison pour laquelle, au vu de ces

métriques et de l'AUC, les importances des variables issues de l'algorithme *Random Forest* sont retenues.

Importances des variables sur l'algorithme *Random Forest*
Les importances (en %) obtenues sont indiquées sur la figure 9.25. D'après ce graphique, les importances des variables sont sensiblement différentes. Les variables $modification\_devis$ et age_devis sont les plus contributives, représentant chacune une part d'importance supérieure à 20%. Les variables $fumeur$, psy et csp apparaissent parmi les variables qui contribuent le moins au modèle. Ces résultats sont cohérents avec les statistiques descriptives effectuées lors du traitement des données. Il avait été conclu que les variables caractéristiques propres à la personne réalisant le devis apportaient peu d'information sur le taux de transformation, ce qui dans l'ensemble se reflète dans ces résultats.

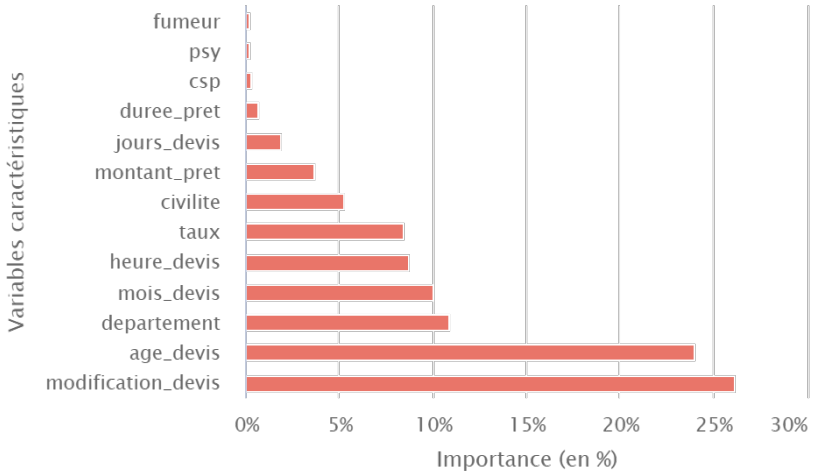

FIGURE 9.25 – Importance des variables selon *Random Forest* par la PFI

Limites de la PFI
La PFI est une méthode utile pour repérer les variables qui ont le plus d'influence sur les prédictions du modèle. Cependant, il faut noter que la PFI peut être sensible aux nouvelles permutations en créant de nouvelles associations de données non conformes à la réalité du marché de l'assurance emprunteur. Par exemple, l'association d'un homme âgé de 60 ans prenant un crédit immobilier d'une durée de 25 ans à un taux très faible, de 1.10%.

$\rightarrow$ Tous les paramètres initiaux sont désormais déterminés. Les algorithmes de classification ascendante hiérarchique, CAH0, CAH1 et CAH2 sont appliqués par la fonction $hclust()$ du package *fastcluster*. La prochaine étape consiste à déterminer un nombre de *cluster* optimal pour poursuivre l'analyse de ces différentes approches.

Chapitre 10

Implémentation des méthodes

Ce chapitre présente l'analyse des clusters *suite à l'application des modèles sur les 3 scénarios.*

Pour rappel les méthodes CAH0, CAH1 et CAH2 correspondent à :
— CAH0 : approche classique ;
— CAH1 : approche par l'ajout d'une matrice de voisinage portée sur la variable cible Y ;
— CAH2 : approche par pondération selon l'importance des variables caractéristiques X ;

L'implémentation de ces méthodes sur R se différencie par l'application des coefficients suivants aux matrices de distances :

	CAH0	**CAH1**	**CAH2**
Matrice de dissimilarité $D1$	1	α	w_j
Matrice de voisinage $D2$	0	$(1-\alpha)$	0

FIGURE 10.1 – Coefficients appliqués aux matrices de distances

Ce chapitre étudie le nombre optimal de *clusters* à prendre en compte pour faciliter l'analyse des résultats. Les informations sur la qualité de ces regroupements aideront, lors du chapitre suivant, à identifier le ou les profils les plus susceptibles de transformer leur devis en contrat. Une attention particulière sera également accordée aux profils les moins sujet susceptibles de transformer. En effet, cette information est importante pour avoir une bonne compréhension du marché et éventuellement adapter son offre pour attirer ces profils.

1 Corrélation cophénétiques

Avant d'étudier le nombre optimal de *clusters*, cette étape préliminaire basée sur la corrélation cophénétique (déjà introduite pour l'optimisation du paramètre α : 4.1) contribue à comparer la qualité des différentes approches de *clustering*.

La corrélation cophénétique se construit à l'aide :
— des distances réelles entre les devis obtenues à partir de la matrice de dissimilarité
— des distances cophénétiques qui sont les distances issues du dendrogramme. La distance cophénétique entre deux devis se définit comme la distance issue du dendrogramme lorsque deux devis se rencontrent pour la première fois. A titre d'informmation, les dendrogrammes du scénario 1 sont illustrés en annexes (4).

En posant les notations suivantes :
— $d(x_i, x_j)$ la distance entre deux devis x_i et x_j de la matrice de dissimilarité

— $c(x_i, x_j)$ la distance cophénétique entre ces deux devis x_i et x_j issus du dendrogramme
— $\overline{d}$ la moyenne des distances de dissimilarités pour l'ensemble des devis
— $\overline{c}$ la moyenne des distances cophénétiques pour l'ensemble des devis

Le coefficient de corrélation cophénétique, noté $\mathcal{C}$, se définit comme suit :

$$\mathcal{C} = \frac{\sum_{i<j}(d(x_i, x_j)) - \overline{d})(c(x_i, x_j) - \overline{c})}{\sqrt{(\sum_{i<j}(d(x_i, x_j)) - \overline{d})^2)(\sum_{i<j}(c(x_i, x_j) - \overline{c})^2)}}$$

Il mesure à quel point les distances cophénétiques conservent la structure de similarité des données initiales. Plus la corrélation cophénétique est élevée (mesure à maximiser), plus les structures de regroupement issues du dendrogramme sont similaires aux distances initiales.

En comparant les corrélations cophénétiques entre différentes approches de *clustering*, celle qui produit des regroupements les plus cohérents avec les données d'origine peut être identifiée. Les résultats de la corrélation cophénétique pour les trois scénarios, calculée avec la fonction $cophenetic()$ du *package stats* sur R sont présentés dans le tableau suivant :

	Scénario 1	Scénario 2	Scénario 3
CAH 0	0.30	0.29	0.31
CAH 1	0.50	0.37	0.51
CAH 2	0.75	0.75	0.77

FIGURE 10.2 – Corrélations cophénétiques

Analyse des différences entre scénarios
Pour les CAH0 et CAH2, les corrélations restent stables par l'application des trois scénarios. A l'inverse, la CAH1 montre un affaiblissement de sa corrélation pour le scénario 2. La CAH1 semble donc sensible au rééquilibrage de la variable cible. Ce rééquilibrage a certainement augmenté la séparation entre les étiquettes de la variable Y, conduisant a une forte modification des distances entre les devis transformés et ceux non transformés.

Analyse des différences entre CAH
A l'exception de la CAH1 du scénario 2, la corrélation de la CAH1 et la CAH2 est respectivement 1.6 et 2.5 fois supérieure à celle de la CAH0. Cela signifie qu'il y a un intérêt à ajouter une information supplémentaire sur Y car elle permet de structurer les données de telle manière à rendre l'algorithme fidèle à sa représentation. Par conséquent, la matrice des distances fournie en entrée offre une meilleure fidélité des nouveaux algorithmes, et ce en particulier pour la CAH2 qui offre des meilleurs résultats que la CAH1.

En conclusion, cette étape préliminaire établit une base pour la sélection d'une méthode de *clustering* : les nouvelles approches de *clustering* produisent des regroupements plus cohérents avec les données d'origine, en confère leur corrélation plus élevée.

2 La sélection du nombre optimal de *clusters*

La représentation des devis en un nombre optimal de *clusters* simplifie la structure des données en identifiant des groupes homogènes. La recherche de ce nombre permet de contrôler la formation excessive de petits groupes, éventuellement difficiles à interpréter. Cette recherche permet également d'éviter la création insuffisante de grands groupes qui pourraient masquer des structures importantes.

Cette recherche s'établit au travers des indicateurs de :

— qualité interne, se calculant uniquement à partir des informations caractéristiques des données, sans avoir recours à la connaissance a priori de la variable Y
— qualité externe, s'appuyant sur une connaissance a priori de la variable Y

2.1 Mesure de qualité interne

Les indicateurs de qualité interne fournissent une évaluation intrinsèque de la cohérence des *clusters*. L'indice des sauts d'inertie du dendrogramme puis l'indice de Silhouette sont présentés dans cette partie.

2.1.1 Indice des sauts d'inertie du dendrogramme

Définition

L'inertie mesure la similarité des données au sein de chaque *cluster*. Les sauts d'inertie du dendrogramme représentent les changements dans l'inertie des *clusters* lors du regroupement des données. À chaque étape de fusion, l'inertie globale diminue car les observations deviennent de plus en plus similaires. Cette diminution se représente par des sauts d'inertie plus ou moins importants.

Application aux données

L'inertie est calculée par la commande $arbre\$height$ suite à la création de l'arbre par la fonction $hclust()$ du *package cluster*. Afin de comparer les différents sauts d'inertie entre les approches, ces mesures sont ramenées en pourcentage parmi les 10 plus grandes valeurs. Les indices de sauts d'inertie obtenus sont représentés ci-dessous :

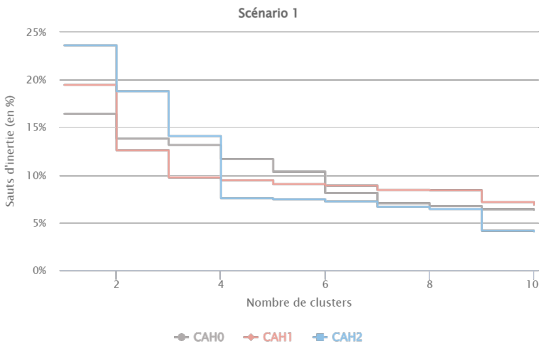

FIGURE 10.3 – Indice de sauts d'inertie pour le scénario 1

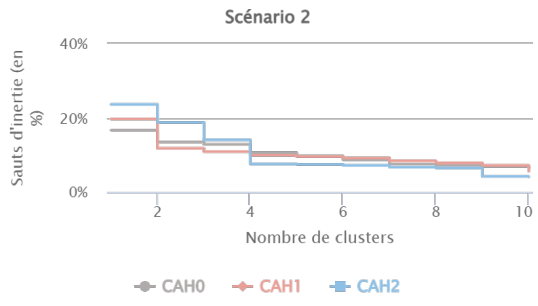

FIGURE 10.4 – Indice de sauts d'inertie pour le scénario 2

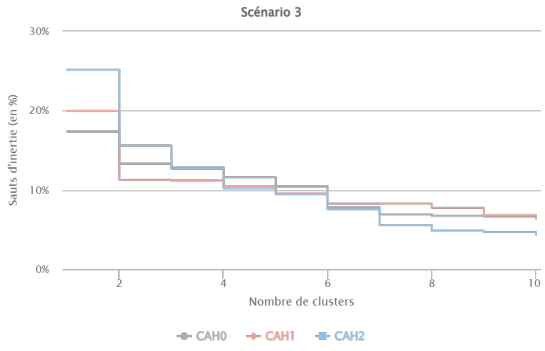

FIGURE 10.5 – Indice de sauts d'inertie pour le scénario 3

D'après les indicateurs du scénario 1 (10.3), l'inertie diminue brusquement de 6% lors du passage de trois à quatre *clusters* pour la CAH2 et se stabilise ensuite. Cela signifie que ce regroupement permet de capturer les variations significatives présentes dans les données, sans devoir continuer la subdivision en nouveaux groupes : c'est la méthode du "coude". Elle consiste à identifier le point de la courbe où la réduction de l'inertie ralentit. La CAH1 affiche quant à elle un unique saut net de 8% pour le passage de un à deux *clusters*. Toutefois, réduire l'analyse à deux *clusters* ne semble pas intéressant car cela pourrait entraîner une perte importante d'informations et peut ne pas refléter la diversité des données réelles.

D'après les indicateurs du scénario 2 (10.4), les sauts d'inertie sont assez faibles à partir du regroupement en quatre *clusters*.

D'après les indicateurs du scénario 3 (10.5), un unique saut net est remarqué sur l'ensemble des CAH lors du passage de un à deux *clusters*. Les autres sauts sont peu marqués et aucun coude n'est facilement observable pour la CAH2 car l'inertie décroît continuellement, bien qu'un léger "coude" semble se former à partir du huitième *clusters*.

L'indicateur des sauts d'inertie montre que :

— la CAH0 offre peu de dispartié dans ses sauts à l'instar des deux autres approches, indiquant déjà une dispartié d'inertie qui sera visible lors de l'analyse des *clusters*.

Limites

Bien que l'ensemble des inerties totales soient proches de 10% pour un regroupement en quatre groupes, l'interprétation des sauts d'inertie peut être subjective et ce, spécialement lorsque les sauts ne sont pas assez distincts les uns des autres. Il est donc difficile de sélectionner celui qui va permettre de repérer le nombre de *clusters* optimal. Un autre indicateur de qualité interne est donc étudier.

2.1.2 Indice de Silhouette

Définition

L'indice de Silhouette [5] représente la différence entre la distance moyenne des points d'un même *cluster* et la distance moyenne des points des *clusters* voisins. Plus cette valeur est haute, meilleur est la classification des données.

Pour calculer cet indicateur, il faut d'abord calculer deux distances a_i et b_i :

— La distance a_i est la distance moyenne entre le devis i et son groupe C_i calculée en prenant la moyenne des distances entre le devis i et tous les devis $j \in C_i, i \neq j$ présents dans le groupe. En notant $|C_i|$ le cardinal du groupe C_i, elle vaut :

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i \, i \neq j} d(i, j)$$

— La distance b_i est la distance minimale entre le devis i et le centre de chaque autre groupe C_j. Elle vaut :

$$b(i) = \underset{i \neq j}{min} \frac{1}{|C_j|} \sum_{j \in C_j} d(i, j)$$

Si la différence $b(i) - a(i)$ est négative, alors la distance $a(i) > b(i)$ et le devis est mal classé car il est en moyenne plus proche du groupe voisin que du sien. À l'inverse, si cette différence est positive, alors le devis est bien classé car il est en moyenne plus proche de son groupe que du groupe voisin.

L'indicateur de Silhouette du devis i est ainsi défini par la formule suivante :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Pour obtenir l'indice du groupe, il faut prendre la moyenne des indices obtenus pour tous les devis dans chaque groupe. Un indicateur de Silhouette élevé signifie que le regroupement des données en un nombre K de *clusters* considéré est approprié. Par conséquent, l'indicateur de Silhouette maximal est retenu.

Application aux données
La fonction $silhouette()$ du *package cluster* sur R est utilisée, et présente les résultats suivants :

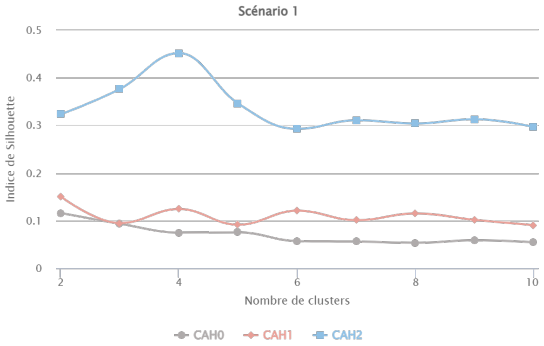

FIGURE 10.6 – Indice de Silhouette pour le scénario 1

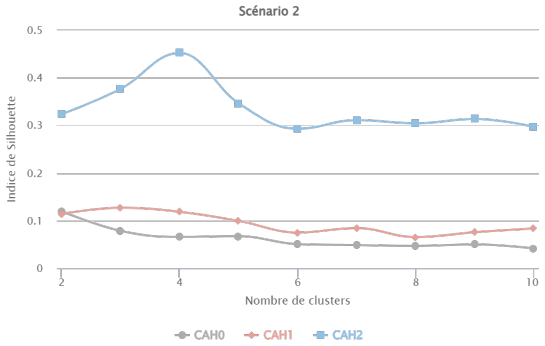

FIGURE 10.7 – Indice de Silhouette pour le scénario 2

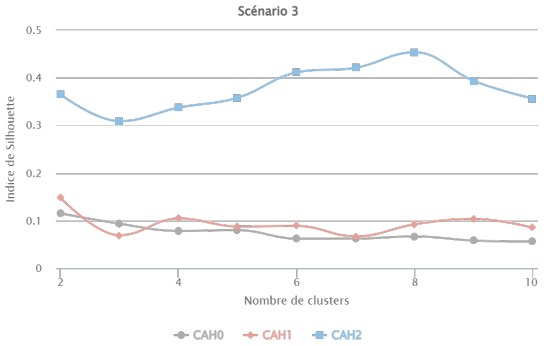

FIGURE 10.8 – Indice de Silhouette pour le scénario 3

Dans l'ensemble, l'indice de Silhouette représenté par les courbes de la CAH1 et CAH2 donne des valeurs supérieures à celles de la CAH0 pour un regroupement en quatre *clusters*. La forte supériorité de la CAH2 peut s'expliquer par l'ajout d'une pondération sur les variables qui vient impacter les distances entre les devis.

Pour le scénario 1 et le scénario 2, les résultats sont très semblables. L'indice n'est donc pas sensible au rééquilibrage des données.

Pour le scénario 3, il est observé que la méthode de regroupement CAH2 avec un nombre plus élevé de *clusters* présente de meilleurs résultats. Cette observation sera prise en considération dans le dernier chapitre, où les profils types seront examinés au travers d'un plus grand nombre de *clusters*.

L'indice de Silhouette montre que :
— la CAH2 est particulièrement sensible à cet indice suite à la modification des variables prises en compte et offre de meilleurs résultats que la CAH1.

Limites
L'indicateur de Silhouette obtenu reste relativement faible pour les CAH0 et CAH1 (proche de

0.10). Un indicateur de silhouette supérieur à 0.30 peut indiquer une classification raisonnable. Cet indicateur s'utilise lorsque les "proximités sont sur une échelle de ratio", c'est-à-dire que les distances utilisées pour mesurer la similarité entre les points sont basées sur une échelle continue et proportionnelle ; en confère la distance euclidienne. Or ici, la distance utilisée est celle de Gower, ce qui peut expliquer les faibles valeurs obtenues.

Il est donc intéressant d'obtenir une qualité des *clusters* par un autre type de mesure.

2.2 Mesure de qualité externe : l'indice de Rand

A contrario des indicateurs de qualité interne, les indicateurs de qualité externe utilisent les étiquettes connues de la variable Y.

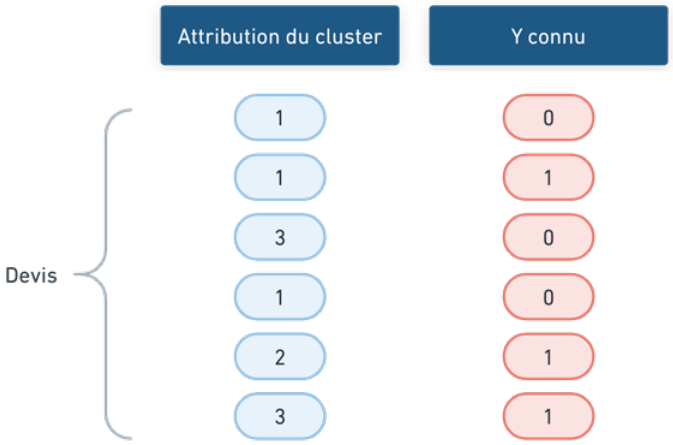

FIGURE 10.9 – Illustration des deux partitions à comparer pour 3 *clusters*

La mesure de qualité externe compare des paires issues :
— du regroupement effectué par l'algorithme CAH (représenté par le numéro du *cluster* attribué) ;
— de la partition de référence étant l'étiquette Y du devis.

L'indice de Rand est présenté ci-dessous. A titre d'information, un autre indice est présenté en annexe (5), offrant des résultats similaires.

Définition
L'indice de Rand évalue la concordance entre deux ensembles de partitions en attribuant un score. Ce score correspond au pourcentage global de paires en accord entre le numéro des *clusters* et les étiquettes Y. Il se calcule selon les quatre quantités suivantes :

— Les paires concordantes (a) sont les paires de devis regroupées ensemble à la fois dans la partition de référence et dans le regroupement effectué par l'algorithme
— Les paires discordantes (b) sont les paires de devis séparées à la fois dans la partition de référence et dans le regroupement effectué par l'algorithme
— Les paires regroupées dans la partition de référence mais séparées par l'algorithme (c)
— Les paires séparées dans la partition de référence mais regroupées par l'algorithme (d)

À partir de ces quantités, l'indice de Rand est calculé selon la formule suivante :

$$I_{Rand} = \frac{a+b}{a+b+c+d}$$

Il varie entre 0 et 1. Une valeur proche de 1 indique que les partitions sont très similaires, tandis qu'une valeur proche de 0 indique une dissimilarité élevée.

Application aux données
La fonction $comparing.Partitions(, type = "rand")$ du *package clusterSim* sur R est utilisée.

D'après les résultats ci-après, l'indice de Rand offre un score relativement élevé en comparaison de l'indice de Silhouette. Sur l'ensemble des résultats, l'indice est supérieur à 0.4 pour un regroupement en quatre *clusters*. Cet indice tend spécialement à s'affaiblir pour la CAH1 à partir du regroupement en cinq *clusters*.

A noter que les CAH0 et CAH2 offrent globalement les mêmes tendances et ne reflètent pas de variations spécifiques.

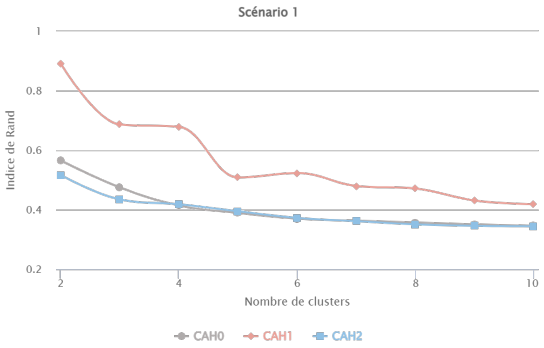

FIGURE 10.10 – Indice de Rand pour le scénario 1

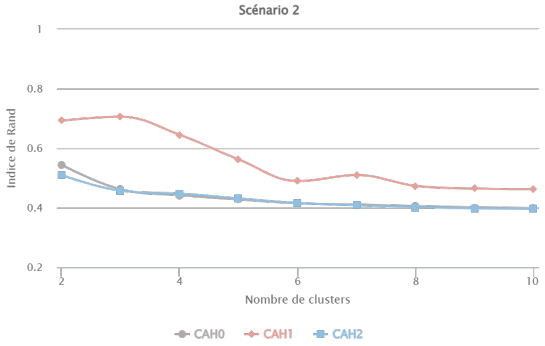

FIGURE 10.11 – Indice de Rand pour le scénario 2

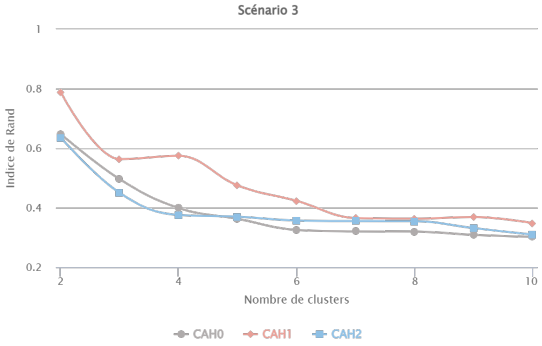

FIGURE 10.12 – Indice de Rand pour le scénario 3

L'indice de Rand permet de montrer que :
— la CAH1 arrive bien à classer les paires de devis pour un nombre de *clusters* inférieur à quatre. Ceci peut être dû une meilleure distinction par l'algorithme des paires bien classées (a) et (b) des paires mal classées (c) et (d).

Limites
Cet indice ne tient pas compte de la structure interne des *clusters*, c'est à dire la manière dont les devis sont regroupés en termes de compacités et de séparation. C'est la raison pour laquelle ses résultats sont différents de ceux des indicateurs de qualité interne.

→ Au vu de tous ces indicateurs et pour l'ensemble des scénarios, le regroupement en quatre *clusters* semble être un compromis approprié. Il est souligné que l'indice de Silhouette est davantage approprié pour la CAH2, et que l'indice de Rand l'est pour la CAH1.

3 Étude des *clusters*

Après avoir déterminé le nombre de groupes à prendre en considération, l'analyse de ces clusters *permet de confronter les résultats obtenus à partir des différentes approches.*

Dans chaque *cluster*, le taux de transformation de devis en contrats est calculé en rapportant le nombre de devis transformés sur le nombre total de devis du groupe. Cela permet de comparer le taux de transformation entre les différents *clusters* et d'identifier, en conséquence, les groupes avec des taux de transformation élevés ou faibles.

3.1 Représentation des *clusters* par leur barycentre et inertie

La représentation visuelle des *clusters* par leur barycentre et leur inertie permet d'avoir une vue d'ensemble des groupes formés et synthétise l'interprétation des résultats.

Le barycentre d'un *cluster* représente le centre de tous les devis le composant (définit ici : 1.2.2). Il est calculé selon la moyenne de tous les points issus des coordonnées de la réduction des données par t-SNE.

L'inertie d'un *cluster* représente la dispersion des devis en son intérieur, calculée comme la somme des distances au carré entre les devis et leur barycentre (définit ici : 3.4). Une faible inertie indique que les devis du *cluster* sont proches de leur barycentre. Dans le cas contraire, les devis sont éparpillés au sein du groupe.

3.2 Etude du scénario 1

La représentation des *clusters* pour le scénario 1 est donnée ci-dessous. Le centre de chaque *cluster* correspond au centre du cercle, le diamètre représente l'inertie et la couleur représente le taux de transformation de devis en contrats. Une intensité plus forte de couleur reflète un taux de transformation plus élevé.

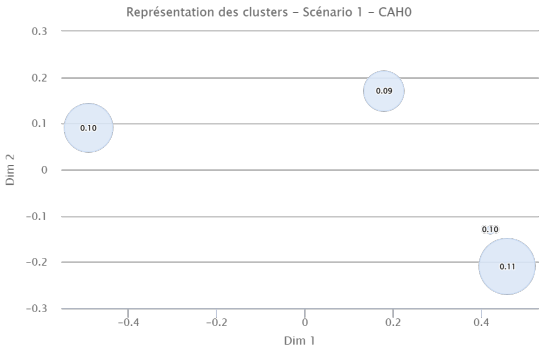

FIGURE 10.13 – Représentation des *clusters* CAH0 - Scénario 1

La représentation de la CAH0 montre un ensemble de couleur similaire pour tous les *clusters*, avec un taux de transformation moyen de 10%. De plus, deux *clusters* sont sensiblement proches entre eux.

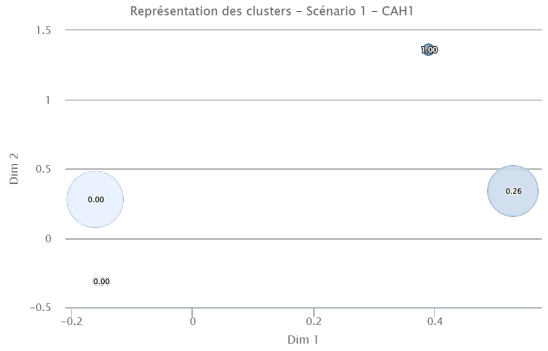

FIGURE 10.14 – Représentation des *clusters* CAH1 - Scénario 1

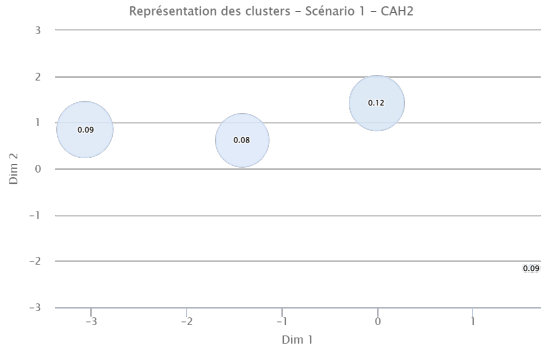

FIGURE 10.15 – Représentation des *clusters* CAH2 - Scénario 1

La représentation de la CAH1 montre quant à elle de fortes disparités sur le taux de transformation pour ses *clusters*. L'un est de 100%, un autre de 30% et les deux derniers sont de 0% (arrondi à la deuxième décimale). Deux *clusters* présentent une aire très faible, ce qui pourrait résulter d'un nombre très faible de devis ou d'une densité particulièrement élevée. Les graphiques de la section suivante fourniront davantage d'informations à ce sujet.

La représentation de la CAH2 ne révèle pas de réelle disparité entre les taux de transformation, cependant, les coordonnées suggèrent que les barycentres de ses *clusters* présentent un éloignement plus marqué, surpassant ceux de la CAH0 et de la CAH1.

Le scénario 1 permet de montrer que :
— la CAH1 démontre une disparité accrue des taux de transformation, tandis que ;
— la CAH2 se caractérise par une séparation plus marquée entre les *clusters*.

3.3 Etude du scénario 2

La représentation des *clusters* pour le scénario 2 est donnée ci-dessous :

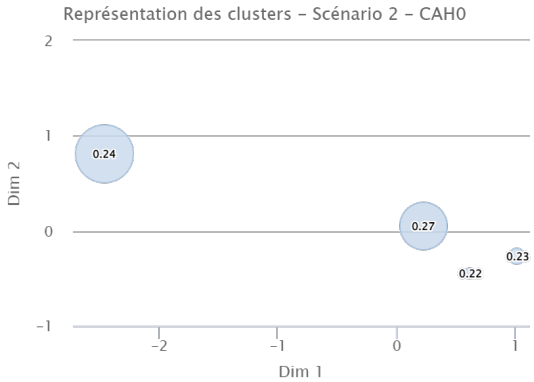

FIGURE 10.16 – Représentation des *clusters* CAH0 - Scénario 2

Les taux moyens issus du scénario 2 sont plus élevés. Ceci est dû au rééquilibrage des données. L'analyse des résultats de la CAH0 reste similaire à celle du scénario 1.

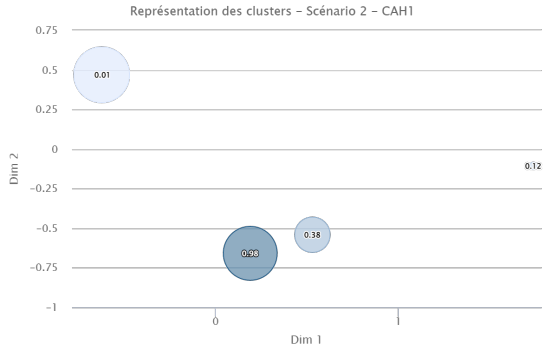

FIGURE 10.17 – Représentation des *clusters* CAH1 - Scénario 2

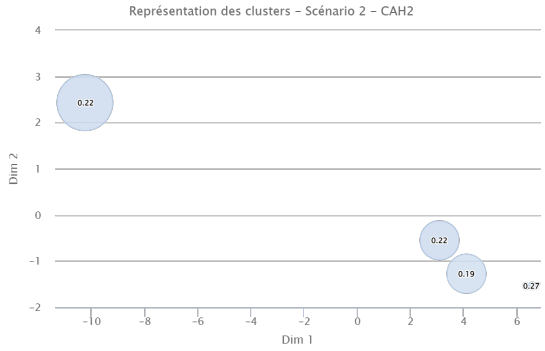

FIGURE 10.18 – Représentation des *clusters* CAH2 - Scénario 2

La CAH1 conserve une distinction plus nette entre les taux, bien que la proximité entre les *clusters* demeure similaire à celle de la CAH0.

La CAH2 conserve son avantage en termes de meilleure séparation des *clusters* dans l'espace, cependant, elle ne présente pas de différence significative sur la disparité entre les taux moyens.

Le scénario 2 permet de montrer que :

— la CAH1 est sensible au rééquilibrage des données car elle offre des inerties bien différentes de celles observées au scénario 1,

— la CAH2 a du mal a capturer une différence sur les taux moyens de transformation, en comparaison à la CAH0. Recalculer l'importance des variables sur cette base de données rééquilibrée pourrait fournir de meilleurs résultats.

3.4 Etude du scénario 3

La représentation des *clusters* pour le scénario 3 est donnée ci-dessous.

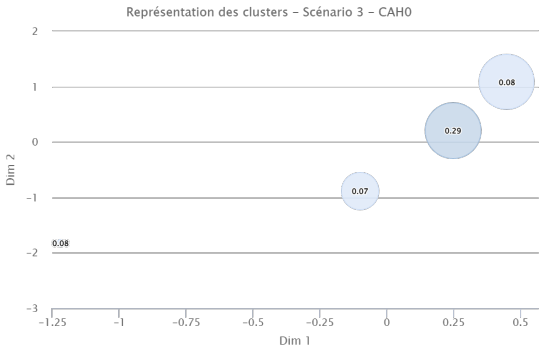

FIGURE 10.19 – Représentation des *clusters* CAH0 - Scénario 3

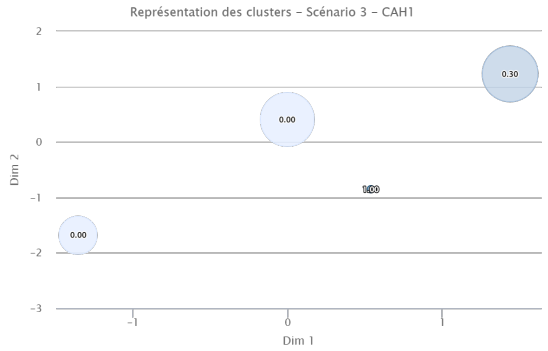

FIGURE 10.20 – Représentation des *clusters* CAH1 - Scénario 3

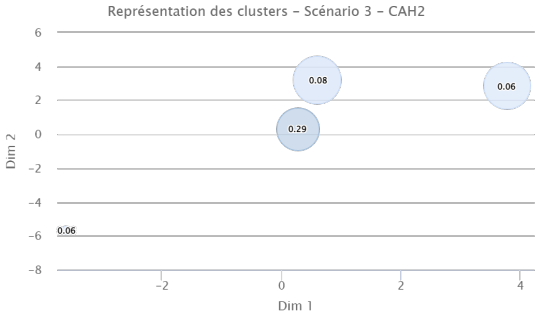

FIGURE 10.21 – Représentation des *clusters* CAH2 - Scénario 3

L'analyse des résultats de la CAH1 reste similaire à celle du premier scénario.

La CAH2 présente des taux de transformation comparables à ceux de la CAH0, tout en offrant une séparation améliorée des *clusters* dans l'espace. Contrairement au scénario 1, les taux de transformation sont significativement plus différenciés et distincts, soulignant ainsi l'intérêt d'ajouter la variable $modification\_devis$ au modèle.

Le scénario 3 permet de montrer que :
— la CAH1 reste peu sensible aux changements de variables caractéristiques
— la CAH2 est sensible au choix des variables caractéristiques

→ Afin d'avoir une idée plus concrète sur la capacité des nouvelles approches à différencier les groupes de devis, un autre graphique est présenté : celui de la densité pour chaque *cluster*. Ce graphique permet de comparer à la fois la disparité des taux de transformation et la séparation des devis dans l'espace.

3.5 Représentation des *clusters* par leur densité

Dans cette étude, la densité peut s'interpréter comme le nombre de devis par unité d'inertie. Elle représente, de manière similaire à un indice démographique, la concentration de devis par unité de surface. Une densité élevée, illustrée par des histogrammes de plus grande hauteur, signifie que les devis du *cluster* sont spatialement plus regroupés entre eux.

Les densités sont représentées pour les trois approches dans les histogrammes ci-dessous :

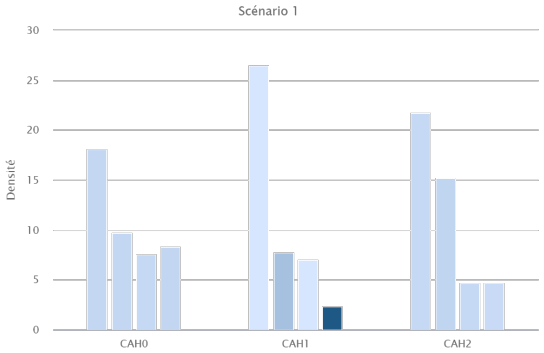

FIGURE 10.22 – Densité des *clusters* - Scénario 1

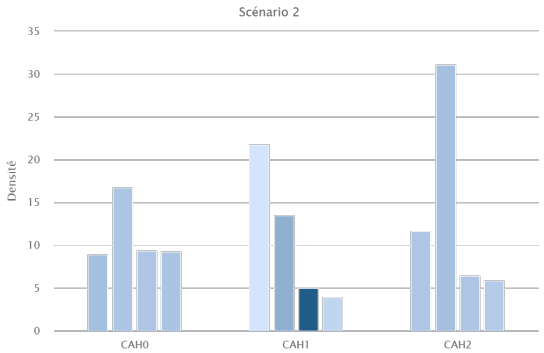

FIGURE 10.23 – Densité des *clusters* - Scénario 2

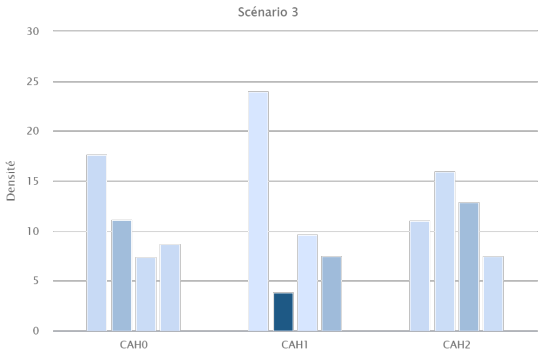

FIGURE 10.24 – Densité des *clusters* - Scénario 3

D'après les résultats du scénario 1, l'approche CAH1 offre une très faible densité pour son *cluster* au plus fort taux de transformation. Cela signifie que l'algorithme ne capte qu'un faible nombre de devis transformés pour sa surface. Cela confirme l'analyse précédentes des *clusters* où une inertie très faible a été obtenu. Sur le scénario 2, la CAH1 offre pour ce même *cluster* une densité double pour une inertie qui a quadruplée. Ainsi, le nombre de devis transformés est deux fois mieux capturés pour ce *cluster* suite au rééquilibrage.

La CAH2 affiche le meilleur compromis entre une densité plus élevée, égale à 12.5, un taux plus élevé parmi l'ensemble des *clusters* du scénario 3.

L'approche par densité permet de retenir que :

— la CAH1 offre ses meilleurs résultats pour le scénario 2 (en comparaison des scénarios 1 et 3)

— la CAH2 offre ses meilleurs résultats pour le scénario 3 (en comparaison des scénarios 1 et 2)

Chapitre 11

Prédiction des modèles

Le modèle entraîné sur la base d'apprentissage peut ensuite être évalué sur la base de validation. Cette dernière étape permet d'estimer la capacité des modèles à prédire la transformation de nouveaux devis en contrats.

Bien que la prédiction soit habituellement associée aux méthodes d'apprentissage supervisé, il existe des moyens de réaliser des prédictions à partir de l'apprentissage non-supervisé, à l'issu du *clustering*. Dans ce contexte, l'objectif n'est pas de prédire des étiquettes spécifiques, mais de déterminer à quel groupe ou *cluster* appartient un nouveau devis, en fonction de sa similarité avec les devis existants.

De ce fait, les prédictions réalisées sur les *clustering* sont comparées de manière indépendantes aux prédictions d'un algorithme de classification : *Random Forest*. Cette comparaison doit être réalisée avec prudence, car théoriquement, ces deux méthodes ne sont pas directement comparables en raison de leur nature différente.

L'algorithme utilisé dans ce mémoire permettant de prédire l'appartenance d'un devis à un *cluster* est d'abord présenté.

1 Prédiction par l'algorithme KNN

Pour chaque devis de la base de test, les modèles entraînés utilisent les variables caractéristiques pour générer une prédiction de la variable de sortie Y.

Il convient de souligner que les algorithmes de la classification ascendante hiérarchique ne permettent pas de "classer" de nouvelles données mais a pour objectif de "connecter" des devis entre eux, de telle sorte qu'il n'est pas directement possible d'attribuer de nouveaux devis à ce modèle.

Néanmoins, il est envisageable d'appliquer la classification ascendante hiérarchique dans un objectif de classer de nouveaux devis, en mettant en place un autre système de classification. Par exemple, l'algorithme du k-plus proches voisins (KNN) peut être utilisé pour déterminer l'attribution d'un nouveau devis à un *cluster*.

L'algorithme KNN, *k-nearest neighbors*, en anglais, est un algorithme d'apprentissage automatique simple à mettre en oeuvre. Cet algorithme utilise un paramètre noté k, qui représente le nombre de voisins les plus proches à prendre en compte. Dans ce cas bien précis, pour prédire le plus finement possible la classe du nouveau devis, un seul voisin est pris en compte. Ainsi, pour $k = 1$, l'algorithme KNN se résume par les étapes suivantes :

1. Pour chacune des observations de la base de validation, calcul de la distance avec toutes les observations de la base d'apprentissage
2. Sélection de la plus petite distance

3. Attribution à l'observation de la classe de ce plus proche voisin

Cet algorithme peut être implémenté par la fonction $knn()$ du *package class*, mais ne traite que les variables quantitatives. Pour s'adapter à ma présence de données qualitatives, une fonction a été créée. Elle permet de :

— Calculer de la distance de Gower entre un devis de la base de validation et tous les autres devis de la base d'apprentissage
— Récupérer le *cluster* du devis de la base d'apprentissage correspondant à la distance minimale parmi celles calculées
— Attribuer au devis à prédire le *cluster* correspondant

Ainsi, cette fonction fournie pour chaque devis de la base de validation un numéro de *cluster* prédit. Une illustration est donnée ci-dessous et indique la représentation t-SNE des devis de la base d'apprentissage et des devis de la base de test :[1]

FIGURE 11.1 – Représentation t-SNE des devis sur la base d'entraînement

FIGURE 11.2 – Représentation t-SNE des devis sur la base de test (après KNN)

D'après ces représentations, les devis de la base de test sont sensiblement regroupés de la même manière que ceux de la base d'entraînement. A noter que les regroupements ne sont pas identiques entre ces deux représentations car la base de test comporte des nouveaux devis en termes de caractéristiques.

L'algorithme KNN est appliqué sur l'ensemble des scénario à l'issu des trois approches de *clustering*. L'étape suivante consiste à confronter les prédictions de ces approches. Ainsi, les résultats obtenus à partir du *clustering* sont présentés en parallèle avec ceux issus d'une approche supervisée classique.

2 Comparaison avec une approche supervisée classique

Dans le processus de l'apprentissage statistique, la prédiction apporte, à l'aide d'indicateurs, des résultats sur la performance des modèles.

Métrique de classification binaire

La classification d'une variable binaire passe par la définition d'un seuil de classification approprié. Ce seuil permet au modèle de catégoriser les devis dans les deux classes, en fonction des probabilités prédites par les CAH.

Par exemple, en fixant le seuil à 0.5, toutes les probabilités prédites au-dessus de cette valeur sont considérées comme appartenant à la classe 1, tandis que celles en dessous de ce seuil sont classées dans la classe 0.

1. Par l'application de la CAH2 du scénario 2 : ce choix est argumenté par la forte corrélation cophénétique et la meilleure capacité de l'approche CAH2 à espacer les *clusters* entre eux.

En appliquant ce seuil, les prédictions peuvent être de quatre natures différentes par rapport à leur étiquette, avec les notations définies ci-dessous :

— Si le modèle prédit la valeur comme positive (1) et qu'elle est effectivement positive (1), alors il s'agit d'un vrai positif (VP)
— Si le modèle prédit la valeur comme négative (0) et qu'elle est effectivement négative (0), alors il s'agit d'un vrai négatif (VN)
— Si le modèle prédit la valeur positive (1) et qu'elle est au contraire négative (0), alors il s'agit d'un faux positif (FP)
— Si le modèle prédit la valeur négative (0) et qu'elle est au contraire positive (1), alors il s'agit d'un faux négatif (FN)

Voici un exemple pour un taux de classification fixé à 0.3% :

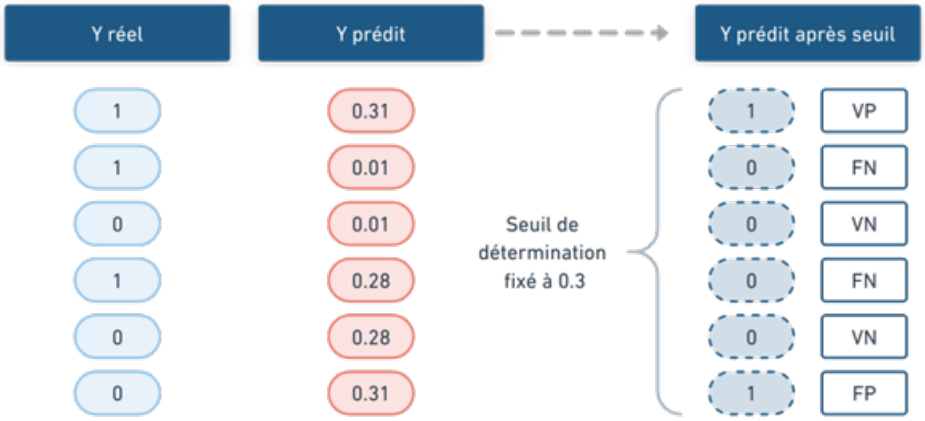

FIGURE 11.3 – Illustration du seuil de détermination pour classification binaire

Le seuil de classification joue un rôle essentiel car il détermine la frontière de décision entre les deux classes en fonction des probabilités prédites par le modèle.

Adaptation au *clustering*

Les algorithmes de CAH actuels prédisent un nombre limité de probabilités de transformation : une pour chaque *cluster*, correspondant à la moyenne des transformations du groupe.

Ainsi, d'après la recherche optimal en 4 *clusters*, il y a seulement 4 probabilités différentes. Ces probabilités peuvent être très similaires entre elles dans le cas où les résultats du *clustering* ne sont pas satisfaisants (en termes de taux moyen de transformation). Un seuil de probabilité élevé n'est donc pas approprié car il classerait toutes les données dans la classe 0. Le seuil de classification envisagé est la moyenne du taux de transformation de la base de données d'entraînement. Ce choix permet a priori de distinguer au mieux les taux des 4 *clusters*.

Matrice de confusion

La matrice de confusion est une synthèse des résultats de prédiction pour un problème de classification. Elle évalue la performance des modèles à prédire correctement les données en comparant les valeurs réelles de la variable cible avec les valeurs prédites par le modèle.

La matrice de confusion est représentée par un tableau à quatre valeurs pour une classification binaire. Elle met en évidence les prédictions correctes (VP et VN) ainsi que les erreurs de prédiction (FP et FN). Une illustration est présentée ci-dessous :

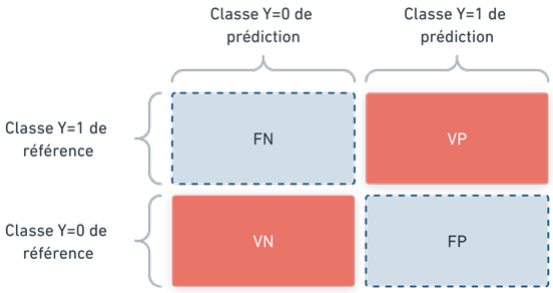

FIGURE 11.4 – Illustration d'une matrice de confusion

Les valeurs issues de la matrice de confusion servent à construire divers indicateurs permettant d'apprécier la qualité d'un modèle de classification. Certains sont présentés ci-dessous :

— **Le rappel** : mesure la proportion de cas positifs réellement identifiés parmi tous les cas réellement positifs. C'est un indicateur important dans les problèmes où la détection des cas positifs est crucial (fraude, domaine médical). Dans le contexte actuel, il permet de quantifier l'identification des devis susceptibles de se transformer en contrat.

$$\text{Rappel} = \frac{VP}{(VP + FN)}$$

— **La précision** : mesure la proportion de cas positifs réellement identifiés parmi toutes les prédictions positives. Elle mesure la capacité du modèle à éviter les erreurs de classification positives. Il s'assure que les devis positifs ne soient pas classés en devis négatifs.

$$\text{Précision} = \frac{VP}{(VP + FP)}$$

— **Le F1-Score** : est défini comme la moyenne harmonique du rappel et de la précision. Cet indicateur est particulièrement utile lorsque les faux positifs et les faux négatifs ont des conséquences importantes et doivent être équilibrés de manière appropriée. Il est souvent privilégié lorsque la détection des cas positifs est tout aussi importante que l'évitement des faux positifs.

$$\text{F1-Score} = 2\,\frac{(\text{rappel} * \text{précision})}{(\text{rappel} + \text{précision})}$$

Les meilleures prédictions sont celles qui classifient aussi bien les devis négatifs que les devis positifs. Il est donc intéressant de s'appuyer en particulier sur le F1-Score qui prend en compte le rappel et la précision. Un modèle est qualifié comme juste pour un F1-score élevé, unique métrique présentée en cette fin de mémoire.

Application aux données issues du *clustering*

Les résultats obtenus à partir du *clustering* sont comparés en parallèle des performances de l'algorithme *Random Forest*. Cet algorithme qui découle de l'apprentissage supervisé sert de référence pour renforcer la contextualisation des résultats de prédiction.

Les matrices de confusions de l'algorithme *Random Forest* sur les trois scénarios, obtenues avec la fonction $confusionMatrix()$ du *package caret* du logiciel R, sont présentées ci-dessous :

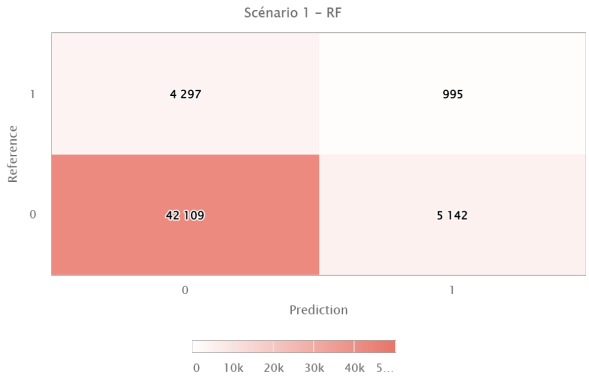

FIGURE 11.5 – Matrice de confusion *Random Forest* - Scénario 1

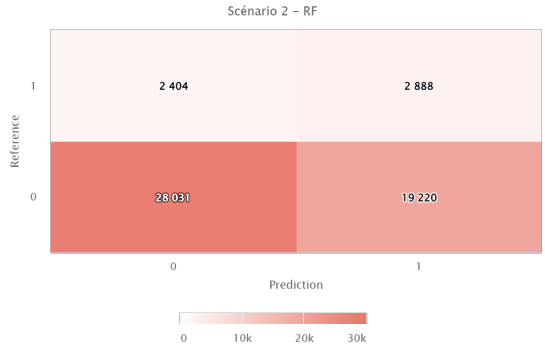

FIGURE 11.6 – Matrice de confusion *Random Forest* - Scénario 2

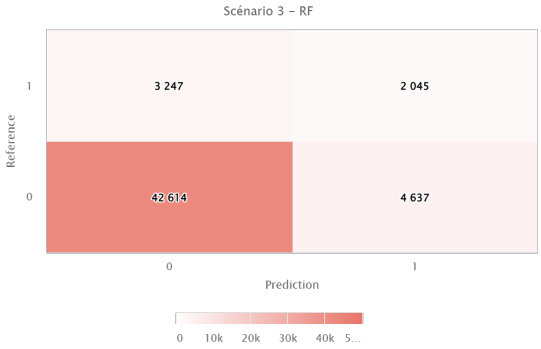

FIGURE 11.7 – Matrice de confusion *Random Forest* - Scénario 3

Les matrices de confusions des méthodes de *clustering* ne sont pas présentées. Les résultats du F1-Score pour les différentes méthodes et les différents scénarios sont présentés dans le tableau suivant :

F1 Score	CAH 0	CAH 1	CAH 2	**RF**
Scénario 1	0.11	0.17	0.17	**0.17**
Scénario 2	0.15	0.19	0.18	**0.21**
Scénario 3	0.11	0.17	0.12	**0.35**

FIGURE 11.8 – F1 Score des trois scénarios

Les résultats du *Random Forest* sont à comparer de manière indépendante de ceux des trois CAH. Ils sont présentés à titre indicatif et servent de référence :

— Le scénario 2 est un peu meilleur que le scénario 1 suite au rééquilibrage ;
— Le scénario 3 offre le F1-Score le plus élevé, signifiant que la variable $modificatin\_devis$ joue un rôle essentiel dans le modèle.

En ce qui concerne les résultats des *clustering*, ceux qui se révèlent les plus pertinents sont issus de la CAH1. Cette prééminence peut être attribuée à la disparité significative des ses taux moyens de transformation obtenus dans chaque *cluster*. Par conséquent, le seuil de classification parvient efficacement à discriminer des classes distinctes de taux.

Pour le scénario 2, il est notable que le F1-score s'améliore à la fois pour la CAH1 et la CAH2, et ce pour des raisons différentes. Contrairement à la CAH1 qui se distingue par des taux distincts, la performance de la CAH2 est attribuable à une densité particulièrement élevée dans l'un de ses *clusters*. Cette situation peut causée la réalisation d'un nombre important de devis classés en transformés.

Pour le scénario 3, le F1-Score demeure équivalent entre la CAH0 et la CAH2. Des observations analogues concernant les *clusters* et les densités avaient déjà été faites lors de l'étude précédente. Toutefois, la CAH2 offre un meilleur espacement de ses *clusters*.

Limites de cette adaptation de prédiction
L'adaptation de cette prédiction sur un nombre de quatre *clusters* restreint la nuance des prédictions car il y a moins d'options pour attribuer un nouveau devis à un *cluster*. Plutôt que de prédire individuellement chaque devis, cette adaptation se rapproche davantage d'une prédiction globale au niveau du *cluster*. De plus, il est important de noter que la répartition du nombre des devis entre ces *clusters* peut varier considérablement, ce qui peut avoir un impact sur la précision des prédictions via le classifieur KNN.

3 Sélection des approches adéquates aux scénarios

Sélection du scénario
Pour construire les profils caractéristiques des devis, il convient de sélectionner l'approche la plus adéquate. Étant donné que l'algorithme *Random Forest* démontre des performances supérieures dans les scénarios 2 et 3, des profils types sont étudiés sur ces deux scénarios.

Sélection de l'approche *clustering*
D'après les analyses précédentes, il a été conclu que la CAH1 était plus appropriée pour le scénario 2 au regard des densités (*figure* 10.23). Bien que sa corrélation cophénétique reste plus faible (0.37) que sur les scénarios 1 et 3 (respectivement 0.50 et 0.51), d'après l'étape de prédiction, la CAH1 affiche le F1-Score le plus élevé (0.19) parmi l'ensemble des résultats issus des approches de *clustering*.

Pour la CAH2, il est plus difficile de sélectionner le meilleur scénario. Cette approche offre une très bonne corrélation cophénétique sur les 3 scénarios (0.75 sur les scénarios 1 et 2, et 0.77 sur le scénario 3), mais les résultats en termes de distinction moyenne des taux de transformation sont peu satisfaisants. Néanmoins, il est remarqué, sur le graphique des densités du scénario 3, un taux légèrement plus élevé, reflété par une couleur plus foncée (*figure* 10.24). De plus, d'après l'indice de Silhouette, la CAH2 atteint son maximum pour une division en huit *clusters*, pour ce même scénario 3. Il peut être judicieux d'étudier cette approche pour un nombre de *clusters* autre que quatre.

Dérive de scénario : sélection du nombre de *clusters*
Une variation de scénario est donc envisagée concernant l'approche CAH2 du scénario 3. Cette décision qui s'appuie en particulier des résultats de l'indicateur de Silhouette (*figure* 10.8), vise à élargir la diversité des profils types.

Pour argumenter ce choix, un nouveau graphique est présenté : il représente le taux moyen de transformation pour l'ensemble des *clusters* obtenus selon le nombre de regroupements choisi. Par exemple, pour une division des données en quatre *clusters*, quatre points sont affichés et représentent les différents taux obtenus. Le taux moyen du portefeuille est affiché par la courbe en pointillés. Ce graphique est réalisé sur le scénario 3 pour la CAH2 d'une part et sur le scénario 2 pour la CAH1 d'autre part. Les graphiques obtenus sont les suivants :

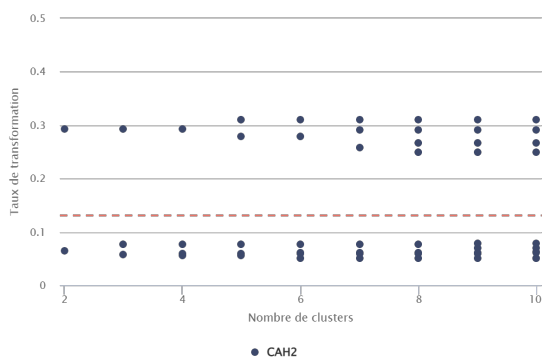

FIGURE 11.9 – Taux moyen de transformation par *clusters* - CAH2 Scénario 3

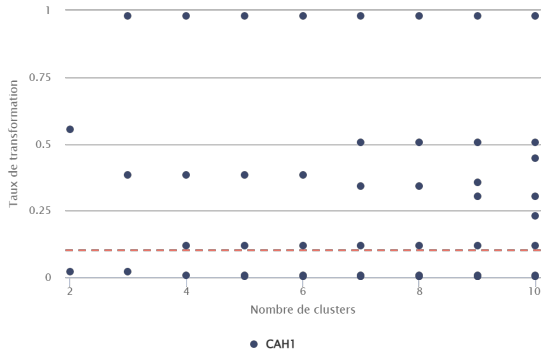

FIGURE 11.10 – Taux moyen de transformation par *clusters* - CAH1 Scénario 2

D'après le graphique 11.30, la CAH2 affiche une différence de taux plus importante entre ses *clusters* pour une division des données en huit regroupements, en comparaison d'un regroupement en quatre *clusters*. En effet, les taux minimal et maximal pour huit *clusters* sont respectivement de 0.05 et de 0.31, contre 0.06 et de 0.28 pour quatre *clusters*. Il est également visible que la répartition en huit *clusters* présente la formation de nouveaux groupes non identiques en taux (par l'apparition de nouveaux points reflétant des taux différents).

A l'inverse, d'après le graphique 11.31, le regroupement en un nombre de quatre ou huit *clusters* affiche des taux extrêmes identiques. Par conséquent, le choix de quatre profils types reste confirmé pour la CAH1 du scénario 2. De plus, l'indice de Rand (*figure* 10.11) présentait des résultats en cohérence avec cette analyse : à partir du quatrième regroupement, les valeurs de l'indice diminuent fortement. Cette diminution est ici reflétée par une distinction visuelle où l'augmentation du nombre de *clusters* ne vient pas augmenter l'apparition de nouveaux points. Par exemple, il y a bien quatre points affichés pour quatre *clusters* mais uniquement cinq points visibles pour huit *clusters* (à la différence respective de trois et sept points visibles dans le cas de la CAH2).

A noter que la présence de *clusters* avec des taux identiques n'est en soit pas un problème car cela peut refléter des profils types distincts présentant des taux de transformation similaires. Toutefois, le choix de conserver un regroupement en quatre *clusters* pour la CAH1 et de dériver en huit *clusters* pour la CAH2 est priorisé par les résultats des indices de Rand et de Silhouette.

Finalement, en combinant l'ensemble des informations obtenues sur la qualité des regroupements, il est retenu la construction de profils types sur :
— l'approche CAH1 du scénario 2 pour la construction de 4 profils type
— l'approche CAH2 du scénario 3 pour la construction de 8 profils type

4 Construction des profils types

Les profils types servent à distinguer et à catégoriser les différentes situations qui se présentent lors de la transition d'un devis vers un contrat.

La construction des profils types s'effectue en comparant le pourcentage des modalités de chaque *cluster* à celui de la base de données entière. Les pourcentages sont observables dans les tableaux, en annexe (6).

D'après ces tableaux, les pourcentages attribués aux modalités sont davantage distincts dans le cas de la CAH2, et ce pour les variables age_devis et $modification\_devis$. Cela est dû à leur forte importance et vient grandement impacter la construction des *clusters*.

A l'inverse, les pourcentages des modalités du tableau issu de la CAH1 apparaissent plus confus pour dégager des caractéristiques propres à chaque *cluster*. Cela peut être dû à la faible corrélation cophénétique de la CAH1 sur le scénario 2 : les données sont déformées pour réunir des devis ayant un taux de transformation semblable et ne font pas ressortir de tendance démarquée.

Ainsi, sur ce jeu de données, l'application de la CAH2 donne de meilleurs résultats.

Néanmoins, pour ces deux approches, les tableaux ont permis de faire ressortir les caractéristiques les plus intéressantes de chaque *cluster*. Cela mène à deux profils types différents :

1. ceux qui transforment leur devis en contrat (attribué par un taux moyen de transformation supérieur à celui du portefeuille[2])
2. ceux qui ne transforment pas leur devis en contrat (attribué par un taux moyen de transformation inférieur à celui du portefeuille).

Les profils les plus susceptibles de transformer leur devis en contrat

Le premier groupe de profils types présente ceux qui ont le plus de chance de faire évoluer leur devis en contrat. Ces profils peuvent fournir des renseignements pour optimiser le processus de contractualisation en identifiant les caractéristiques clés et permettre d'adapter les pratiques commerciales.

Cinq profils types sont présentés ici, deux autres figurent en annexe (7).

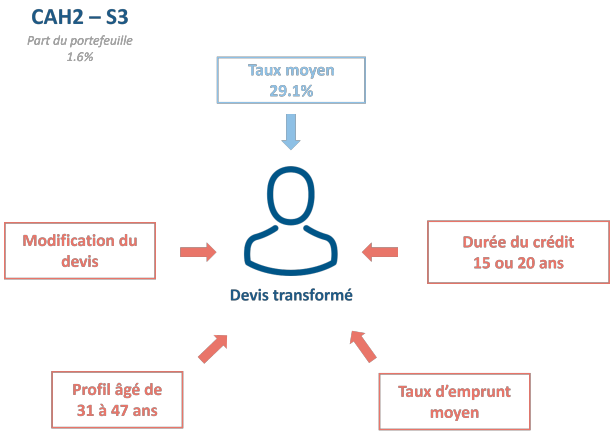

FIGURE 11.11 – Profil type d'un devis transformé - CAH2 Scénario 3 (*cluster 4*)

D'après ce profil type (11.11), il y a en moyenne 29.1% de devis transformés en contrats lorsque les caractéristiques suivantes sont regroupées : l'âge de la personne réalisant le devis est compris entre 31 et 47 ans, la durée du crédit est de 15 ou 20 ans, le taux d'emprunt est qualifié de moyen et le devis a fait l'objet d'une ou plusieurs modifications à la suite de sa création. Ce profil type a été obtenu par la méthode CAH2 du scénario 3 et représente 1,6% du portefeuille.

Les autres profils types obtenus sont les suivants :

2. Pour rappel, le taux moyen du portefeuille est de 10.2%

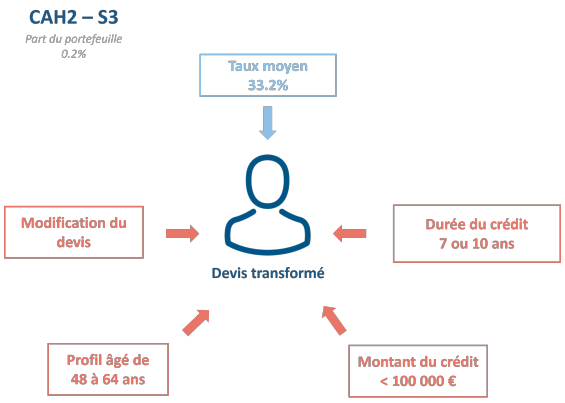

FIGURE 11.12 – Profil type d'un devis transformé - CAH2 Scénario 3 (*cluster 3*)

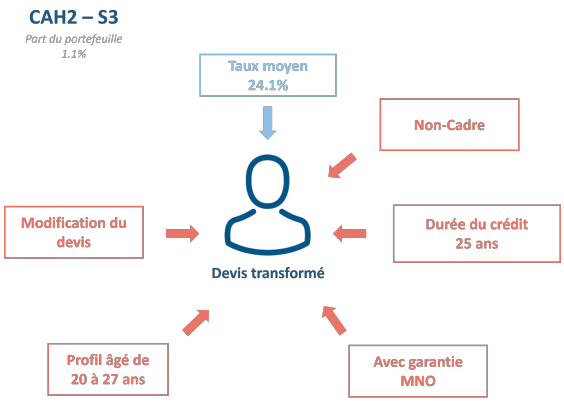

FIGURE 11.13 – Profil type d'un devis transformé - CAH2 Scénario 3 (*cluster 6*)

Les devis ayant été modifiés ont bien plus de chance d'être transformés en contrat et offrent des profils types à taux moyen de transformation au moins supérieur à 24%. Ceci reflète l'analyse descriptive initiale de la variable $modification\_devis$.

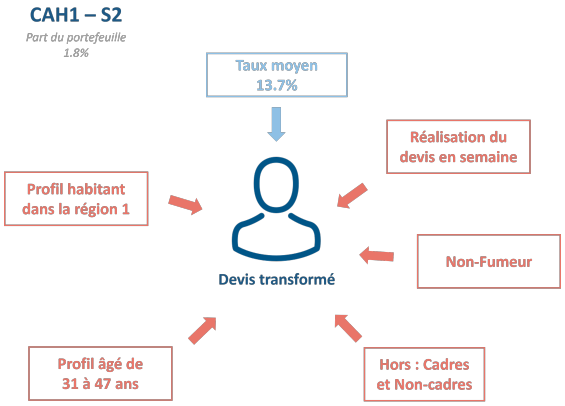

FIGURE 11.14 – Profil type d'un devis transformé - CAH1 Scénario 2 (*cluster 2*)

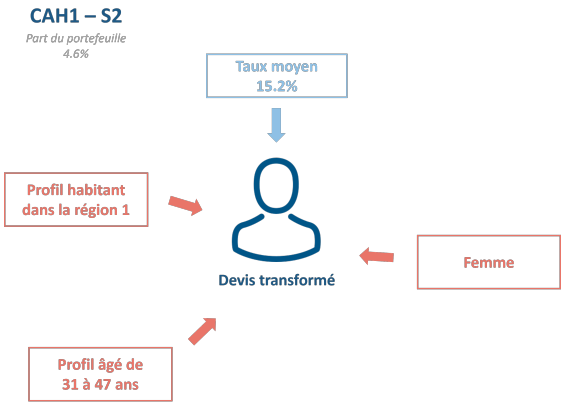

FIGURE 11.15 – Profil type d'un devis transformé - CAH1 Scénario 2 (*cluster 3*)

Les devis présentant les caractéristiques suivantes : "Région 1"[3], "31 à 47 ans", "femme" et "non fumeur", étaient déjà favorables à une meilleure transformation en contrats lors des analyses descriptives.

Les profils les moins susceptibles de transformer leur devis en contrat

Le second groupe de profils types présente les situations où la transformation en contrat s'opère avec une efficacité bien plus faible. Ces profils font ressortir des situations moins attractives mais peuvent néanmoins être opportunistes et servir de passerelle non seulement pour faire évoluer les produits d'assurance qui pointeraient vers ces profils mais également pour accentuer les démarches commerciales et trouver des solutions permettant d'améliorer leur transformation.

3. Pour rappel, la modalité "Région 1" regroupe les régions suivantes : Bretagne, Pays-de-la-Loire, Occitanie, Hauts-de-France et Nouvelle-Aquitaine

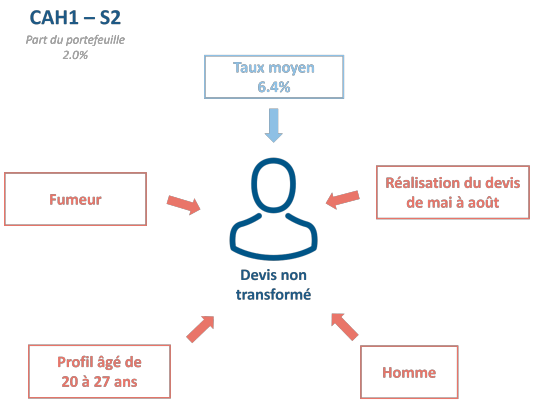

FIGURE 11.16 – Profil type d'un devis non transformé - CAH1 Scénario 2 (*cluster 1*)

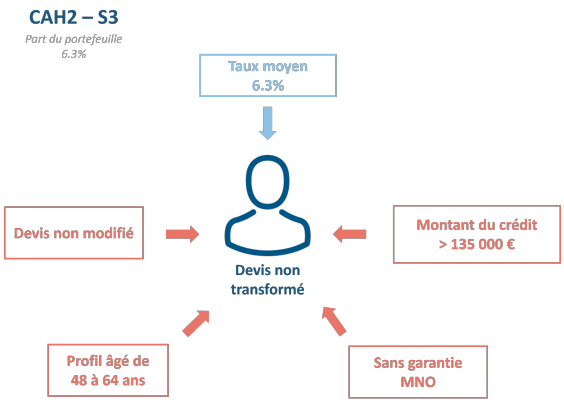

FIGURE 11.17 – Profil type d'un devis non transformé - CAH2 Scénario 3 (*cluster 2*)

D'après ces deux derniers profils, les jeunes hommes adultes et fumeurs, réalisant leur devis sur la période de mai à août ainsi que les personnes âgées de 47 ans et plus, souhaitant un crédit immobilier supérieur à 135 000 € sans la garantie MNO et sans avoir modifié leur devis ; ont un taux de transformation quasiment inférieur au double de celui du portefeuille.

Limites des profils types obtenus

Il est difficile de faire ressortir certaines caractéristiques telles que les horaires, les jours de la semaine, les mois ou la catégorie socio-professionnelle ... car la base de données est initialement complexe à traiter au vu de la période d'observation des années 2020 à 2023, bouleversée par de nombreux faits économiques et financiers. Cela peut impacter la structure des données où les variables présentent un fort déséquilibre du nombre de devis au sein de ces modalités.

Conclusion

L'exploration des données en provenance d'un comparateur d'assurance en ligne a permis de faire ressortir des devis similaires en caractéristiques et en taux de transformation.

Les algorithmes de *clustering* présentés dans ce mémoire, en alternative à l'approche classique, ont conduit à former des groupes homogènes de devis par la jointure des bases de données de devis et de contrats. Lors de l'étude des *clusters*, les approches ont pu démontrer leur réussite quant à la meilleure capacité à distinguer les groupes de devis par des taux de transformation différents. En identifiant les caractéristiques communes des devis à taux de transformation élevés, des profils types se sont démarqués. Ils représentent les devis les plus propices à se convertir en contrat. Les acteurs de l'assurance emprunteur pourraient tirer partie de ces analyses pour améliorer la compréhension des comportements des clients, adapter leurs produits et leurs stratégies commerciales.

En raison des années instables de 2020 à 2023 : période fortement impactée par les taux bas, la crise sanitaire covid-19 et la remontée des taux induite en partie par la guerre en Ukraine ; les profils types obtenus restent assez complexes à identifier et ne font pas ressortir de tendances particulières. Il serait intéressant de réaliser à nouveau cette étude dans quelques années sur une période "plus stable" pour voir si les résultats obtenus s'améliorent.

Il semble également important de souligner que les contrats d'assurance emprunteur sont majoritairement souscrits par deux personnes, et l'étude a pris en compte les caractéristiques de l'emprunteur principal désigné. L'omission des caractéristiques du second emprunteur pourrait avoir un impact sur la représentativité des profils identifiés.

De plus, les données ne permettent pas de différencier les emprunteurs cherchant une assurance pour un crédit nouvellement souscrit de ceux souhaitant changer leur assurance existante. Cette donnée aurait pu apporter davantage d'informations et mener à des profils différents.

A noter que ces résultats ne doivent pas être considérés comme revêtant un caractère général car une même étude réalisée sur les données issues d'autres comparateurs d'assurances aurait pu produire des résultats différents.

L'étude présentée dans ce mémoire pourrait être enrichie et approfondie par une jointure supplémentaire avec une base de données sur les sinistres. Pour justifier l'utilisation de ces données, les actuaires devront respecter le Règlement Général sur la Protection des Données (RGPD) dans le cadre des directives émises par la Commission Nationale de l'Informatique et des Libertés (CNIL) visant à protéger l'usage des données personnelles. L'ajout de cette donnée pourrait permettre de ressortir des profils plus ou moins risqués parmi ceux qui sont plus susceptibles de transformer leur devis en contrat.

Cette analyse présenterait un double intérêt : dans un premier temps intégrer l'indicateur de rentabilité du ratio S/P pour chacun des clusters afin d'orienter la démarche des commerciaux sur les contrats les plus rentables, puis dans un second temps permettre aux actuaires d'utiliser cette segmentation dans une optique tarifaire. Cependant, les modèles de clustering seront à retravailler notamment pour supprimer les variables discriminantes pour une tarification.

Le comparateur pourrait également partager les résultats des profils types avec les assureurs qui proposent leurs contrats, ce qui aiderait les assureurs à se positionner selon la cible commerciale souhaitée. Cela nécessiterait au préalable de retravailler les profils afin d'exclure la variable sexe notamment. Si besoin, les assureurs pourraient ainsi adapter la tarification des produits à destination des profils qui n'offrent pas d'indicateur de rentabilité optimale sur les dits contrats ou d'adapter les garanties afin de mieux répondre aux attentes de leur clientèle cible.

Il est important de noter que le recours au "profilage" (traitement utilisant les données personnelles d'un individu en vue d'analyser et de prédire son comportement), tel que défini dans les articles 4 et 22 du RGPD, peut présenter des défis significatifs pour les travaux actuariels lors de l'identification de profils types et de la segmentation des assurés en fonction de leurs risques. En effet, les restrictions imposées par le RGPD visent à garantir que les individus ne soient pas soumis à des décisions automatisées ayant un impact majeur sur leurs droits, et en particulier ce qui concerne le coût de leur contrat. Cela peut limiter l'utilisation de modèles de clustering qui prédisent automatiquement le comportement des assurés sans leur consentement explicite. Les actuaires doivent donc être conscients de ces contraintes légales pour continuer à exploiter au mieux les données et offrir des produits et des tarifs adaptés à leurs clients.

Finalement, cette étude menée dans le contexte de l'assurance emprunteur a permis de tester l'application de nouvelles méthodes qui peuvent aussi s'appliquer à des contrats Santé ou Prévoyance.

Annexes

1 Le remboursement par annuités constantes : démonstration

Par définition, le flux initial doit être égal à la somme actualisée de tous les flux futurs. Ainsi le capital initial C emprunté à l'instant $t = 0$ pour un remboursement à annuités constantes a peut s'écrire comme la suite des annuités actualisées au taux i :

$$\begin{aligned} C &= \frac{a}{(1+i)} + \frac{a}{(1+i)^2} + \ldots + \frac{a}{(1+i)^T} \\ &= a\left(\frac{1}{(1+i)} + \frac{1}{(1+i)^2} + \ldots + \frac{1}{(1+i)^T}\right) \\ &= \frac{a}{(1+i)}\left(1 + \frac{1}{(1+i)} + \ldots + \frac{1}{(1+i)^{T-1}}\right) \end{aligned} \tag{11.1}$$

Ce qui s'exprime par l'expression d'une suite géométrique de premier terme 1 et de raison $\frac{1}{(1+i)}$. Le capital emprunté peut donc s'écrire :

$$\begin{aligned} C &= \frac{a}{(1+i)} \frac{1 - \frac{1}{(1+i)^T}}{1 - \frac{1}{1+i}} \\ &= \frac{a}{i}((1 - (1+i)^{-T}) \end{aligned} \tag{11.2}$$

Le montant de l'annuité pour chaque période est ainsi déterminé par :

$$a = C\frac{i}{((1 - (1+i)^{-T})}$$

Avec :
— i l'intérêt composé du prêt
— C le capital initial emprunté
— T la durée totale de l'emprunt, en période
— a le montant de l'annuité constante d'une période

2 La présentation de l'algorithme t-SNE

Voici une simple illustration de la méthodologie t-SNE :

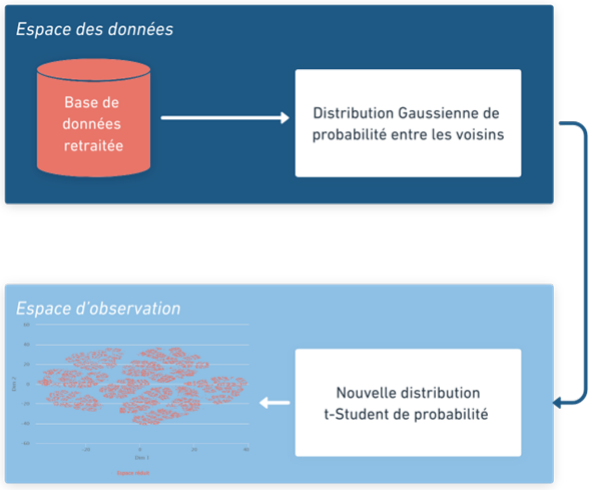

FIGURE 11.18 – Illustration simplifiée de l'algorithme t-SNE

Les trois grandes étapes de l'algorithme t-SNE sont résumées ci-dessous :

1. Calcul des similarités en dimension initiale : pour chacune des observations, une probabilité conditionnelle représentant son voisinage est calculée. Ce voisinage indique la similarité entre cette observation et ses observations voisines. L'ensemble de ces probabilités conditionnelles calculées pour chaque observation le sont par la densité Gausienne. Elles sont regroupées dans une matrice notée $S1$.
2. Calcul des similarités en dimension réduite : au début, les coordonnées idéales sur ce nouvel espace ne sont pas connues. Les points sont donc répartis aléatoirement. Puis, par l'utilisation d'une distribution t-*Student*, est obtenu pareillement qu'à l'étape 1, une liste de probabilités conditionnelles reflétant les similarités entre chaque observations, notée $S2$.
3. Optimisation : pour minimiser l'écart entre ces distributions de probabilités, les similarités $S1$ et $S2$ obtenues sont comparées en utilisant la mesure *Kullback Leibler*. Cette mesure est rendue minimale par des méthodes d'optimisation numérique, en confère la descente de gradient, non expliquée ici.

3 Le rééquilibrage des données

3.1 *Oversampling*

L'approche du sur-échantillonnage, dit "*oversampling*", consiste à augmenter la proportion de devis de la classe minoritaire en créant de nouveaux devis synthétiques.

Dans le cadre de l'*oversampling*, la méthode *Synthetic Minority Oversampling Technique* (SMOTE) est souvent utilisée. Elle vise à augmenter le nombre de devis transformé en contrat, en générant de nouveaux devis qui ressemblent aux autres sans être pour autant identiques.

L'algorithme SMOTE s'applique aux variables continues. Pour s'adapter aux variables catégorielles, une variante de cet algorithme est utilisée : le SMOTE-*Nominal Continuous* (SMOTE-NC). Il permet de traiter à la fois les variables numériques préalablement normalisées et les variables catégorielles.

Les étapes du SMOTE-NC

Les différentes étapes du SMOTE-NC sont résumées ci-dessous :

1. Sélection aléatoire d'un premier devis, dev_1, de la classe minoritaire
2. Identification des k plus proches voisins de dev_1, parmi les devis de la classe minoritaire
3. Sélection aléatoire d'un second devis, dev_2, parmi ces k plus proches voisins
4. Création d'un devis synthétique, $dev_{\text{créé}}$, à partir des variables caractéristiques de dev_1 et dev_2, par l'attribution de nouvelles caractéristiques :
   — Pour les variables numériques, génération d'un coefficient $\alpha \in [0; 1]$ et attribution de la valeur numérique située entre dev_1 et dev_2 selon la valeur du coefficient
   — Pour les variables catégorielles, attribution de la modalité qui est majoritaire parmi les k plus proches voisins.
5. Répétition de ces opérations sur les devis minoritaires jusqu'à atteindre le pourcentage souhaité de devis transformés en contrats

Application aux données

Le SMOTE-NC est appliqué à la base de données retraitée mais non regroupée en nouvelles modalités et sur des données normalisées. La fonction $SMOTE\_NC()$ du package *RSBID*[4] permet l'implémentation de cette méthode sur R.

4. RSBID : Resampling Strategies for Binary Imbalanced Datasets : Implémentation de la variante SMOTE-NC, disponible depuis 2021 dans le package RSBID, GitHub

3.2 *Undersampling*

L'approche du sous-échantillonnage, dit "*undersampling*", consiste à réduire la proportion de devis de la classe majoritaire en éliminant certains de ces devis.

La méthode de *Tomek Links* peut être utilisée dans le cadre de l'"*undersampling*". Elle vise à éliminer les points ambigus (appelés bruits) des données en identifiant des paires de devis transformés et non transformés très proche entre elles. Le devis de la classe majoriatire est alors supprimé. Cependant, dans certains cas, l'algorithme peut considérer que les devis de la classe minoritaire sont également ambigus et les supprimer.

La fonction $step\_tomek()$ du *package themis* permet l'implémentation de cette méthode sur R. Les résultats obtenus sur les données ne sont pas satisfaisant car un grand nombre de devis de la classe minoritaire est également supprimé. Par conséquent, une autre méthode est utilisée : le tirage aléatoire.

Le tirage aléatoire est une technique simple et rapide à mettre en œuvre. Elle implique la sélection aléatoire d'un sous-ensemble de devis en réduisant la part de devis non transformés.

4 Les dendrogrammes

Voici les dendrogrammes du scénario 1 :

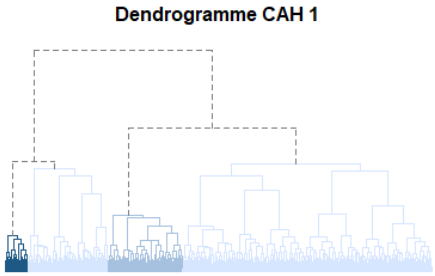

FIGURE 11.19 – Dendrogramme CAH0 - Scénario 1

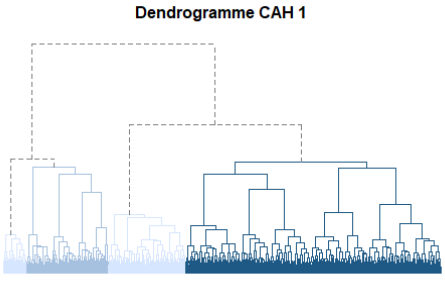

FIGURE 11.20 – Dendrogramme CAH1 - Scénario 1

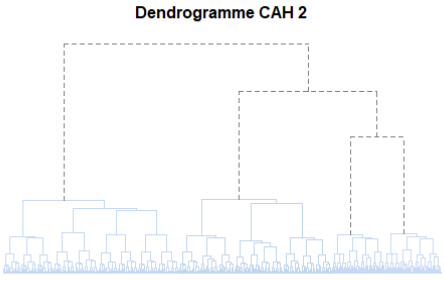

FIGURE 11.21 – Dendrogramme CAH2 - Scénario 1

5 L'indice de Fowlkes-Mallows

Définition

L'indice de Fowlkes-Mallows mesure, comme l'indice de Rand, la similitude entre deux partitions. A la différence de l'indice de Rand, l'indice de Fowlkes-Mallows se concentre davantage sur les paires de devis classées de manière cohérente, et ne prend pas en compte toutes les paires de devis, soit les paires discordantes (b). S'il est reprit les termes définis par l'indice de Rand, l'indice de Fowlkes-Mallows s'exprime par :

$$I_{FM} = \frac{a}{\sqrt{(a+c)\,(a+d)}}$$

L'indice de Fowlkes-Mallows varie également entre 0 et 1. Une valeur de 1 indique une parfaite similarité entre les partitions, tandis qu'une valeur de 0 indique une absence de similarité.

Application aux données

La fonction $FM\_index()$ du *package dendextend* sur R est utilisée. Les résultats sont les suivants :

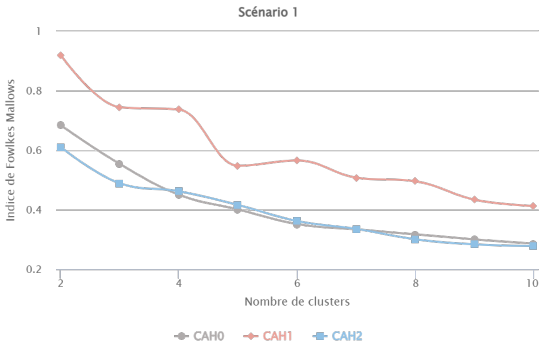

FIGURE 11.22 – Indice de Fowlkes-Mallows pour scénario 1

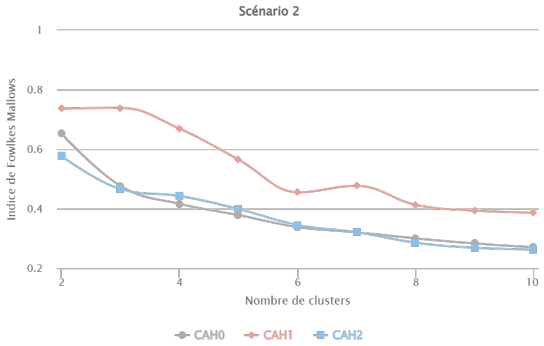

FIGURE 11.23 – Indice de Fowlkes-Mallows pour scénario 2

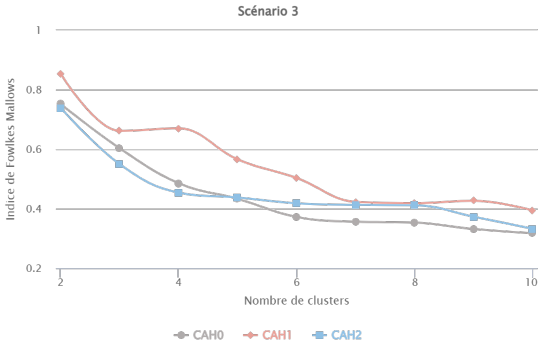

FIGURE 11.24 – Indice de Fowlkes-Mallows pour scénario 3

L'analyse des résultats de l'indice de Fowlkes-Mallows est globalement similaire à celle des résultats de l'indice de Rand ; en raison de leur proximité de mesure.

6 Le tableau de construction des profils types

Profil moyen	CAH1 sur Scénario 2 (4 *clusters*)				Portefeuille
Cluster	Devis transformés	Devis transformés	Devis transformés	Devis non transformés	Tous
	2	3	4	1	
Taux de transformation	38,4%	97,8%	11,7%	0,5%	10,2%
Part de la population	72,9%	3,8%	2,9%	20,4%	100,0%
Variables propre au devis					
Quadrimestre : 1 / 2 /3	37% / 30% / 33%	37% / 28% / 35%	40% / 32% / 28%	32% / **40%** / 28%	38% / 32% / 30%
Semaine / Week-end	**99%** / 1%	88% / 12%	2% / **98%**	96% / 3%	96% / 4%
Horaire libre / Horaire travail	24% / 76%	35% / 65%	30% / 70%	43% / 56%	23% / 77%
Modification devis : Oui / Non					16% / 84%
Variables propre à l'utilisateur					
Homme / Femme	68% / 32%	48% / **52%**	7% / **93%**	**85%** / 14%	76% / 24%
Région : 1 / 2 /3	**44%** / 32% / 24%	**41%** / 33% / 26%	34% / 37% / 29%	33% / 35% / 32%	37% / 37% / 26%
Fumeur / Non Fumeur	10% / **90%**	28% / 72%	10% / **90%**	**30%** / 70%	15% / 85%
Garantie MNO : Oui / Non	67% / 33%	58% / 42%	48% / 52%	58% / 41%	75% / 25%
Cadres / Non-Cadres / Autres	30% / 55% / **14%**	30% / 45% / 25%	30% / 61% / 9%	33% / 43% / 24%	31% / 58% / 11%
Age : [20;27] / [28;30] / [31,47] / [48;64]	8% / 9% / **52%** / 31%	9% / 11% / **47%** / 32%	11% / **12%** / **34%** / 41%	**23%** / 13% / 48% / 16%	9% / 9% /55% / 27%
Variables propre au crédit					
Durée : 7 / 10 / 15 / 20 / 25 ans	1% / 7% / 21% / 31% / 40%	5% / 14% / 25% / 26% / 30%	1% / 5% / 17% / 28% / 49%	8% / 17% / 23% / 29% / 23%	1% / 6% / 15% / 28% / 50%
Taux : Faible / Moyen / Elevé	23% / 44% / 33%	25% / 37% / 38%	22% / 37% / 41%	30% / 40% / 30%	24% / 28% / 48%
Montant : 1 / 2 / 3 / 4 /5	37% / 29% / 17% / 1% / 16%	31 % / 23% / 20% / 3% / 23%	43% / 28% / 14% / 2% / 12%	29% / 22% / 21% / 6% / 22%	48% / 28% / 12% / 2% / 10%
Profil type simplifié	*Personne agée de 31 à 47 ans, non fumeur, autres csp, habitant la région 1 et réalisant le devis en semaine*	*Femme agée de 31 à 47 ans habitant en région 1*	*Femmes agées de 28 à 47 non fumeur réalisant le devis le week-end*	*Jeune homme agé de 20 à 27 ans, fumeur et réalisant le devis sur la période d'été*	
Taux de consommation associé	13,7%	15,2%	13,4%	6,4%	
Part de la population associé	1,8%	4,6%	0,5%	2,0%	

FIGURE 11.25 – Tableau de construction des profils types - CAH1 Scénario 2

Profil moyen	CAH2 sur Scénario 3 (8 *clusters*)				Portefeuille
Cluster	Devis transformés	Devis transformés	Devis transformés	Devis transformés	Tous
	3	4	6	8	
Taux de transformation	29,0%	31,0%	25,0%	27,0%	10,2%
Part de la population	10,0%	13,0%	3,0%	3,0%	100,0%
Variables propre au devis					
Quadrimestre : 1 / 2 /3	38% / 30% / 32%	36% / 31% / 33%	36% / 31% / 33%	39% / 30% / 30%	38% / 32% / 30%
Semaine / Week-end	94% / 6%	88% / 12%	90% / 10%	92% / 8%	96% / 4%
Horaire libre / Horaire travail	30% / 70%	34% / 66%	30% / 70%	33% / 66%	23% / 77%
Modification devis : Oui / Non	**100%** / 0%	**100%** / 0%	**100%** / 0%	**100%** / 0%	16% / 84%
Variables propre à l'utilisateur					
Homme / Femme	67% / 33%	66% / 34%	68% / 32%	**72%** / 28%	76% / 24%
Région : 1 / 2 /3	40% / 34% / 26%	40% / 34% / 26%	40% / 33% / 27%	40% / 34% / 26%	37% / 37% / 26%
Fumeur / Non Fumeur					15% / 85%
Garantie MNO : Oui / Non	54% / 46%	68% / 32%	**85%** / 15%	**81%** / 19%	75% / 25%
Cadres / Non-Cadres / Autres	32% / 48% / 20%	33% / 48% / 17%	24% / **67%** / 9%	33% / 54% / 13%	31% / 58% / 11%
Age : [20;27] / [28;30] / [31,47] / [48;64]	0% / 0% / 0% / **100%**	0% / 0% / **100%** / 0%	**100%** / 0% / 0% / 0%	0% / **100%** / 0% / 0%	9% / 9% /55% / 27%
Variables propre au crédit					
Durée : 7 / 10 / 15 / 20 / 25 ans	**4%** / **14%** / 28% / 33% / 20%	3% / 9% / **21%** / **32%** / 35%	0% / 1% / 11% / 21% / **63%**	0% / 3% / 12% / 25% / **60%**	1% / 6% / 15% / 28% / 50%
Taux : Faible / Moyen / Elevé	21% / 39% / 40%	26% / **40%** / 34%	26% / 45% / 29%	26% / 44% / 30%	24% / 28% / 48%
Montant : 1 / 2 / 3 / 4 / 5	35% / 26% / 18% / **3%** / **18%**	37% / 26% / 17% / 3% / 17%	36% / 28% / 17% / 2% / 17%	45% / 28% / 15% / 1% / 11%	48% / 28% / 12% / 2% / 10%
Profil type simplifié	*Personnes agées de 48 à 64 ans pour une durée d'emprunt 7 ou 10 ans, montant <100 000 et avec modification devis*	*Personnes agées de 31 à 47 ans pour une durée d'emprunt 15 ou 20 ans à taux moyen et avec modification devis*	*Jeune adulte de 20 à 27 ans non cadre, prêt de 25 ans avec modification devis et garantie MNO*	*Personnes agées de 28 à 30 ans, prêt de 25 ans avec modification et garantie MNO*	
Taux de transformation associé	33,2%	29,1%	24,1%	34,4%	
Part de la population associé	0,2%	1,6%	1,1%	0,2%	

FIGURE 11.26 – Tableau de construction des profils types - CAH2 Scénario 3 *(Partie 1/2)*

Profil moyen	CAH2 sur Scénario 3 (8 *clusters*)				Portefeuille
Cluster	Devis non transformés	Devis non transformés	Devis non transformés	Devis non transformés	Tous
	1	2	5	7	
Taux de transformation	7,6%	5,8%	6,0%	5,0%	10,2%
Part de la population	27,0%	26,0%	9,0%	9,0%	100,0%
Variables propre au devis					
Quadrimestre : 1 / 2 /3	35% / 32% / 33%	36% / 32% / 32%	35% / 31% / 33%	36% / 33% / 31%	38% / 32% / 30%
Semaine / Week-end	84% / 16%	89% / 11%	90% / 10%	87% / 12%	96% / 4%
Horaire libre / Horaire travail	37% / 63%	36% / 64%	34% / 66%	34% / 65%	23% / 77%
Modification devis : Oui / Non	0% / **100%**	0% / **100%**	0% / **100%**	0% / **100%**	16% / 84%
Variables propre à l'utilisateur					
Homme / Femme	62% / 38%	64% / 36%	66% / 34%	64% / 36%	76% / 24%
Région : 1 / 2 /3	36% / 35% / 29%	37% / 34% / 29%	36% / 36% / 26%	35% / 37% / 28%	37% / 37% / 26%
Fumeur / Non Fumeur					15% / 85%
Garantie MNO : Oui / Non	60% / 40%	53% / **47%**	68% / 32%	70% / 30%	75% / 25%
Cadres / Non-Cadres / Autres	32% / 46% / 22%	32% / 43% / 25%	31% / 55% / 14%	27% / 61% / 11%	31% / 58% / 11%
Age : [20;27] / [28;30] / [31,47] / [48;64]	0% / 0% / **100%** / 0%	0% / 0% / 0% / **100%**	0% / **100%** / 0% / 0%	**100%** / 0% / 0% / 0%	9% / 9% /55% / 27%
Variables propre au crédit					
Durée : 7 / 10 / 15 / 20 / 25 ans	5% / **12%** / **22%** / 30% / 31%	6% / 18% / 28% / 28% / 20%	0% / 5% / 15% / 32% / 48%	0% / 5% / 15% / 28% / 52%	1% / 6% / 15% / 28% / 50%
Taux : Faible / Moyen / Elevé	26% / 38% / 36%	23% / 40% / 37%	25% / **43%** / 32%	25% / **45%** / 30%	24% / 28% / 48%
Montant : 1 / 2 / 3 / 4 / 5	30% / 25% / 20% / 5% / 20%	**30%** / **24%** / 19% / 7% / 20%	36% / 28% / 17% / 4% / 15%	30% / 28% / 20% / 5% / 17%	48% / 28% / 12% / 2% / 10%
Profil type simplifié	*Personnes agées de 31 à 47 ans, prêt de 10 ou 15 ans et sans modification devis*	*Personnes agées de 48 à 64 ans, montant > 135 000€ qui ne modifient pas leur devis et sans garantie MNO*	*Personnes agées de 28 à 30 ans avec emprunt à taux moyen sans modification devis*	*Personnes agées de 31 à 47 ans avec emprunt à taux moyen sans modification devis*	
Taux de transformation associé	7,5%	6,3%	4,9%	3,8%	
Part de la population associé	1,9%	6,3%	3,9%	4,2%	

FIGURE 11.27 – Tableau de construction des profils types - CAH2 Scénario 3 *(Partie 2/2)*

7 Les autres profils types

Les autres profils ayant un taux de transformation élevés

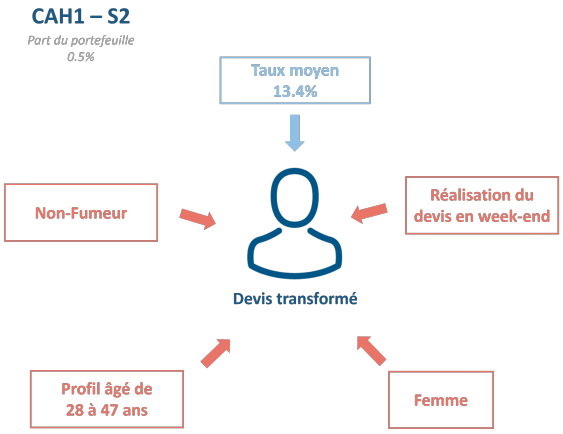

FIGURE 11.28 – Profil type d'un devis transformé - CAH1 Scénario 2 (*cluster 4*)

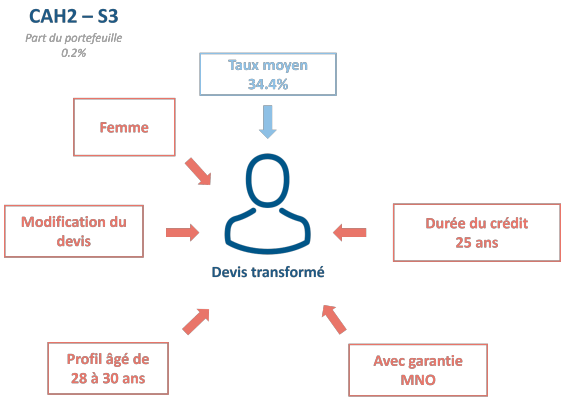

FIGURE 11.29 – Profil type d'un devis transformé - CAH2 Scénario 3 (*cluster 8*)

Les autres profils ayant un taux de transformation faibles

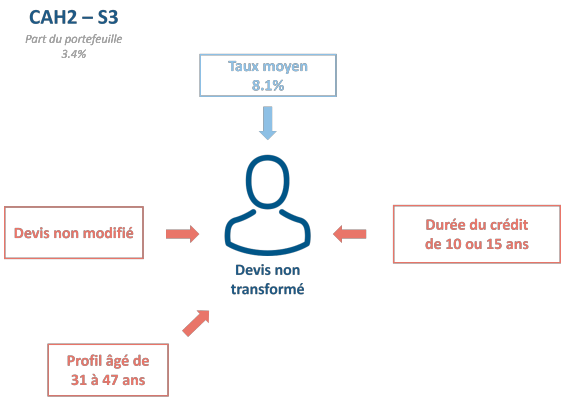

FIGURE 11.30 – Profil type d'un devis non transformé - CAH2 Scénario 3 (*cluster 1*)

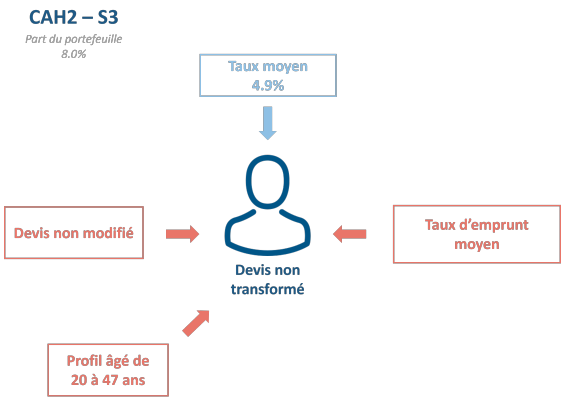

FIGURE 11.31 – Profil type d'un devis non transformé - CAH2 Scénario 3 (*cluster 5 et 7*)

Table des figures

Bibliographie

[1] Leo Breiman. Random forests in machine learning, 45, 5–32. https://link.springer.com/article/10.1023/a:1010933404324, 2001.

[2] Tianqi Chen & Carlos Guestrin. Xgboost : A scalable tree boosting system. https://arxiv.org/pdf/1603.02754.pdf, 2016.

[3] Lise Bellanger & Arthur Coulon & Philippe Husi. Une méthode de classification ascendante hiérarchique par compromis : hclustcompro - p52. https://hal.science/hal-03280918/document, 2021.

[4] Vol. 27 No. 4. J. C. Gower, Biometrics. A general coefficient of similarity and some of its properties. https://members.cbio.mines-paristech.fr/~jvert/svn/bibli/local/Gower1971general.pdf, 1971.

[5] Peter J.Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis, journal of computational and applied mathematics, 20 :53-65, 1987. https://www.sciencedirect.com/science/article/pii/0377042787901257, 1987.

[6] Christine Malot-Tuleau. Introduction à la sélection de variables, cours de master mass. https://math.unice.fr/~malot/CART.pdf, 2006.

[7] Dimitri Delcaillau & Antoine Ly & Franck Vermet & Alizé Papp. Model transparency and interpretability : survey and application to the insurance industry. https://doi.org/10.1007/s13385-022-00328-y, 2022.

[8] Marie Chavent & Vanessa Kuentz-Simonet & Amaury Labenne & Jérôme Saracco. Clustgeo : an r package for hierarchical clustering with spatial constraints. https://hal.science/hal-01664018/document, 2020.

[9] Geoffrey Hinton & Laurens van der Maaten. Visualizing data using t-sne, hal open science, journal of machine learning research 9 2579-2605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf, 2008.