Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Julie DI FALCO

Titre du mémoire :

Modélisation du comportement client à la souscription

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

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

Membres présents du jury de la filière : Signature :

Entreprise :

Nom : PACIFICA

Signature :

Directeur de mémoire en entreprise

Membres présents du jury de l'Institut des Actuaires : Signature :

Nom : Juan CALDERON
Signature :

Invité :
Nom : Audrey MAHUZIER
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 :

Institut Statistique de l'Université de Paris

Par : Julie DI FALCO

Mémoire de fin d'étude

Modélisation du comportement client à la souscription

Entreprise : Pacifica
Service : Direction Marché des Particuliers
Tuteur en entreprise : Juan CALDERON
Tuteur académique : Olivier LOPEZ

- Année 2019/20 -

Résumé

Devant un marché de l'assurance automobile de plus en plus concurrentiel, les organismes d'assurance sont tenus de suivre et d'adapter leurs offres en permanence, notamment en termes de position tarifaire de manière à répondre au mieux aux besoins des clients et ainsi fidéliser ses clients et conquérir de nouvelles parts de marché. Le taux de transformation est un indicateur important en assurance, puisqu'il permet d'évaluer l'attractivité d'un produit, mais aussi d'ajuster et piloter la stratégie tarifaire de la compagnie de façon à pouvoir réagir rapidement aux demandes du marché et s'adapter à la concurrence. En termes de stratégie, les organismes d'assurances disposent de plusieurs méthodes pour encourager les ventes, parmi lesquelles se trouvent les promotions commerciales, véritable levier des ventes. Maîtriser leur usage demeure essentiel, que ce soit pour maximiser leurs effets en ciblant les clients qui y sont sensibles, ou pour mesurer et contrôler les coûts associés.

Ce mémoire se décompose en deux parties : la première partie permet d'appréhender l'environnement dans lequel les promotions commerciales sont utilisées, en caractérisant la stratégie commerciale actuellement mise en place dans le réseau d'agences. La seconde partie s'axe sur le comportement client à la souscription afin d'optimiser l'utilisation des promotions commerciales. Ainsi, l'objectif de ce mémoire est de caractériser la stratégie commerciale développée dans le réseau, avant d'identifier et de préconiser des critères optimaux de l'usage des promotions.

Mots clés:
Assurance automobile, Taux de transformation, Promotion commerciale, Clustering, Analyse en Composantes Principales, Gradient Boosting Machine, Modèle Linéaires Généralisés, Régression pénalisée - Lasso, Recherche approximative (distance de Levenshtein, Algorithme Soundex)

Abstract

Faced with an increasingly competitive car insurance market, insurance companies are required to constantly monitor and adapt their offers, particularly in terms of their pricing position, in order to best meet the needs of their customers and thus build loyalty and win new market shares. The transformation rate is an important indicator in insurance, as it allows to evaluate the attractiveness of a product, but also to adjust and steer the pricing strategy of the company in order to be able to react quickly to market demands and adapt to competition. In terms of strategy, insurance companies have several methods at their disposal to encourage sales, among which are commercial promotions, a real sales lever. Mastering their use remains essential, whether to maximize their effects by targeting sensitive customers, or to measure and control the associated costs.

This master's dissertation is divided into two parts: the first part allows us to understand the environment in which commercial promotions are used, by characterizing the commercial strategy currently implemented in the agency network. The second part focuses on the customer's behaviour at the time of subscription in order to optimize the use of commercial promotions. Thus, the objective of this master's thesis is to characterize the commercial strategy developed in the network, before identifying and recommending optimal criteria for the use of promotions.

Keywords :
Car insurance, Transformation rate, Sales promotion, Clustering, Principal Component Analysis, Gradient Boosting Machine, Generalized Linear Model, Penalized regression - Lasso, Approximate search (Levenshtein distance, Soundex algorithm)

Note de synthèse

Face à un marché de l'assurance automobile de plus en plus concurrentiel, soutenu par la mise en place de réglementations telles que la libéralisation des assurances, la loi Consommation (dite loi Hamon)… les organismes d'assurances sont contraints de suivre et d'améliorer sans cesse leurs offres. Ceci passe par des revues tarifaires visant à fidéliser les clients déjà en portefeuille, et à conquérir de nouvelles parts de marché en réalisant des affaires nouvelles.

Pour mesurer l'efficacité de leurs offres, les organismes d'assurances disposent d'indicateurs de performance, tels que le taux de transformation, qu'ils peuvent suivre.
Outre le fait d'évaluer la performance des offres, le suivi et l'analyse de ce taux de transformation permettent également de piloter leurs stratégies commerciales. Au sein de ces dernières, se trouvent l'emploi de promotions commerciales, dont l'objectif est de stimuler les ventes. Dans le monde de l'assurance, elles constituent, en effet, un véritable levier des ventes et de renforcement de la concurrence.
Dès lors, il est nécessaire de comprendre et de maîtriser l'usage de la promotion commerciale. Au premier abord, il est courant de penser que les promotions ne présentent que des avantages, comme la fidélisation des clients, la conquête ou la création de *plus-value*. Toutefois, elles présentent des coûts associés. En effet, un usage excessif aura tendance à annihiler l'objectif défini en ayant un impact réduit, sans compter le coût qui lui est associé. Optimiser l'usage des promotions prend alors tout son sens, et cela passe par l'identification et le ciblage des clients sensibles aux offres commerciales. *In fine*, la segmentation des clients selon leur sensibilité aux promotions commerciales permet de maximiser le taux de transformation étant donné un taux de promotion fixé selon la stratégie commerciale de l'entreprise.

C'est dans ce cadre que ce mémoire s'inscrit : **modéliser le comportement client au moment de la souscription, en vue d'optimiser l'utilisation des promotions commerciales**.
Il se décompose en deux parties. La première vise à caractériser la stratégie actuelle d'utilisation des promotions commerciales, tout en prenant en compte la pluralité des stratégies commerciales qui existent dans le réseau d'agences de Pacifica. La seconde partie cherche à déterminer des critères, issus des caractéristiques des profils de clients sensibles à la présence d'une promotion de façon à améliorer le processus de vente en les ciblant, étant donné un budget de promotions commerciales déterminé. Ainsi, le taux de transformation est maximisé grâce à l'optimisation de l'allocation des promotions.

Construction de la base de données

Notre étude se base sur les projets (devis et propositions) automobile du marché des particuliers, réalisés entre 2017 et 2019.
Dans un premier temps, il a fallu constituer une base de données en fusionnant les informations relevant des devis et celles des propositions. Avant d'ajouter des informations que nous avons jugé utiles pour notre étude, comme des variables externes telles que des données socio-démographiques de l'INSEE.

Parmi les variables, l'une d'entre elle a requis un traitement particulier, une correction d'orthographe selon une technique de recherche approximative (ou *fuzzy matching*). Cette étape était nécessaire pour intégrer la notion de concurrence à notre base. Peu de données relatives à ce sujet étaient disponibles, mis à part une variable déclarative, valorisée par le nom de l'ancien organisme d'assurance. C'est la saisie libre de cette variable qui constituait une difficulté, car cela engendre une multitude d'orthographes possibles. Corriger ces chaînes de caractères requiert de les considérer de deux façons différentes, comme une suite de caractères alphanumériques ou comme une suite de sons (représentation phonétique). Par conséquent, le processus de correction doit tenir compte de ces deux représentations, en utilisant deux approches : la distance de Levenshtein et l'algorithme du *Soundex*.

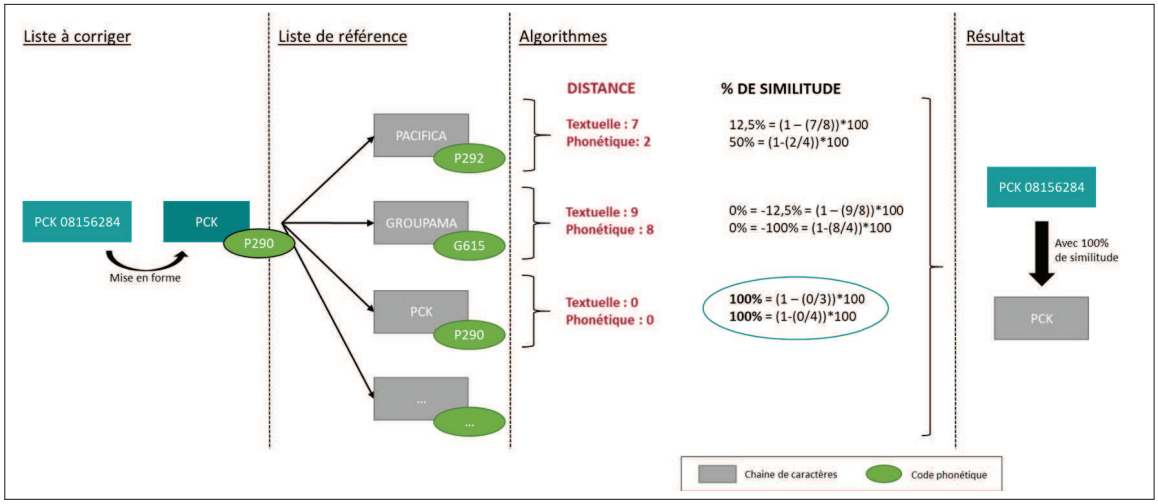


Figure 1 – *Schéma illustrant la correction orthographique pour une chaîne de caractères*

Partie 1

Appréhension de la stratégie commerciale et segmentation des caisses régionales

Une fois tous les traitements réalisés et la base de données exploitable, nous commençons par analyser le contexte et l'environnement dans lequel notre étude s'intègre. Cette étape de compréhension de la stratégie commerciale actuellement mise en place est nécessaire, afin de comprendre l'usage des promotions commerciales au sein du réseau d'agences Pacifica.
Cette appréhension de l'environnement stratégique commence par la création d'un tableau de bord interactif qui synthétise différents indicateurs caractéristiques de la stratégie commerciale, parmi lesquels on trouve le taux de transformation, le taux d'usage des offres promotionnelles (taux de promotion) ou encore la part de dépassement du budget promotionnel.

Actuellement, le taux de transformation du produit automobile est stable depuis 3 ans. A l'inverse du taux de promotion qui croît chaque année.
En revanche, si on descend à l'échelle des caisses régionales (CR), des réseaux vendeurs, des disparités apparaissent. Le taux de transformation devient moins stable et il varie beaucoup d'une CR à une autre, de même que le taux de promotion. Cette hétérogénéité se justifie en partie par une intensité concurrentielle différente dans les régions, mais aussi par des comportements clients et des pratiques commerciales différents. En effet, chaque CR possède sa propre politique commerciale ; cette absence de stratégie nationale complique notre analyse, car il existe autant de typologies comportementales qu'il y a de CR, soit 40. Appréhender les promotions commerciales et leur usage devient complexe. Pour contourner cette multitude de stratégies, nous réalisons une segmentation des CR selon leurs pratiques promotionnelles, à l'aide d'un *clustering* et d'une ACP (Analyse en Composantes Principales). Ainsi, nous espérons réduire le nombre de typologies comportementales et réussir à caractériser les stratégies commerciales actuellement déployées par le réseau.

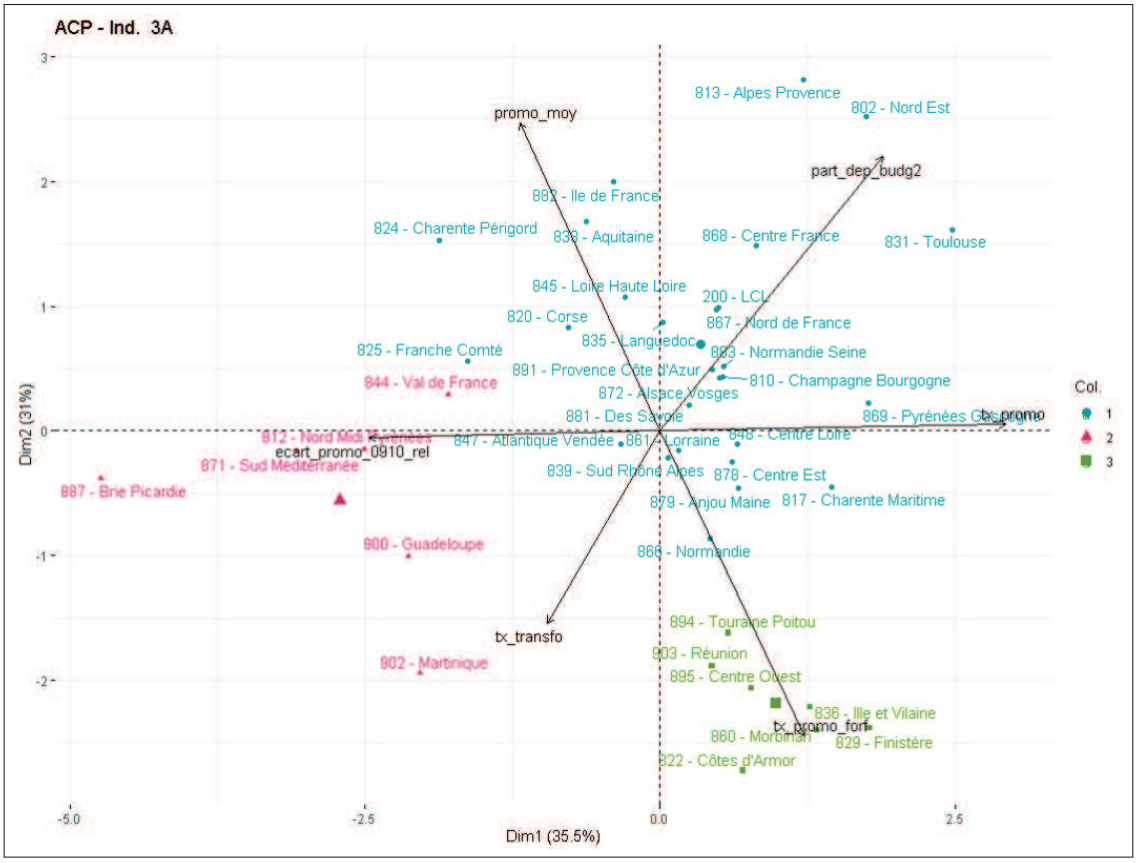


Figure 2 – Graphe des variables et des individus (données moyennes observées sur 3 ans)

A partir de cette segmentation des CR (réalisée sur des données moyennes observées sur 3 ans), nous identifions 3 typologies comportementales caractérisables :

- Le *cluster* n°1, regroupant des CR utilisant beaucoup de promotions (dépassement du budget), un montant moyen de promotion élevé et un taux de transformation peu élevé.
- Le *cluster* n°2, avec des CR utilisant peu de promotions, sauf en période « temps fort » (période marquée par une hausse des promotions chaque année), et avec un taux de transformation élevé.
- Le *cluster* n°3, contenant des CR utilisant beaucoup de promotions de type « forfait », avec un montant moyen de promotion bas et un « bon » taux de transformation.

Partie 2

Modélisation du taux de transformation et identification des critères optimaux

Maintenant que nous avons identifié 3 principales approches commerciales, nous procédons à la seconde partie du mémoire, à savoir la modélisation du comportement client à la souscription.
En préambule de la modélisation par GLM, on utilise le *Machine Learning* pour identifier les variables les plus influentes, qui seront intégrées dans le GLM.
A partir d'une modélisation du taux de transformation en utilisant un GBM (*Gradient Boosting Machine*), nous obtenons un classement des variables ayant le plus d'influence dans la construction des arbres (i.e. dans l'explication du taux de transformation).

Cependant, avant d'intégrer ces variables explicatives au GLM, il faut mesurer la qualité du modèle pour s'assurer de la crédibilité de la sélection des variables.
Différents indicateurs existent parmi lesquels se trouve l'AUC, le *recall*, la *precision*... Ces indicateurs permettent de valider notre modèle avec un AUC de 0,75 et un *recall* de 0,77. Des valeurs plus importantes auraient été appréciables, elles auraient renforcé nos certitudes sur

la qualité du modèle.
Il est possible d'aller plus loin et de regarder dans quelle mesure les variables identifiées précédemment interviennent dans le modèle. De cette manière, nous pallions au côté « boîte noire » du GBM en en apprenant plus sur l'impact d'une variable sur la prédiction, tout en pouvant juger la cohérence de l'information. Un impact contradictoire avec les connaissances métiers aura tendance à nous interroger et à remettre en question le modèle.
Les graphes de dépendance partielle (DP) concèdent à cela, en permettant d'interpréter globalement une variable et de visualiser son impact.

Une fois la sélection de variables effectuée, il est possible de passer à la modélisation du taux de transformation par GLM. Cette modélisation permet de simuler l'appétence d'un client à la souscription, information non-négligeable.
Après vérification de la qualité de l'ajustement, l'une des premières variables à regarder, est la prédiction de la variable indicatrice d'une promotion (*indpromo*). Elle permet de confirmer la légitimité de l'étude sur les promotions et la recherche d'optimisation, car les promotions entrent bien en jeu dans l'acte de vente. La présence d'une promotion augmente la probabilité de transformer de 72% (toute chose égale par ailleurs).

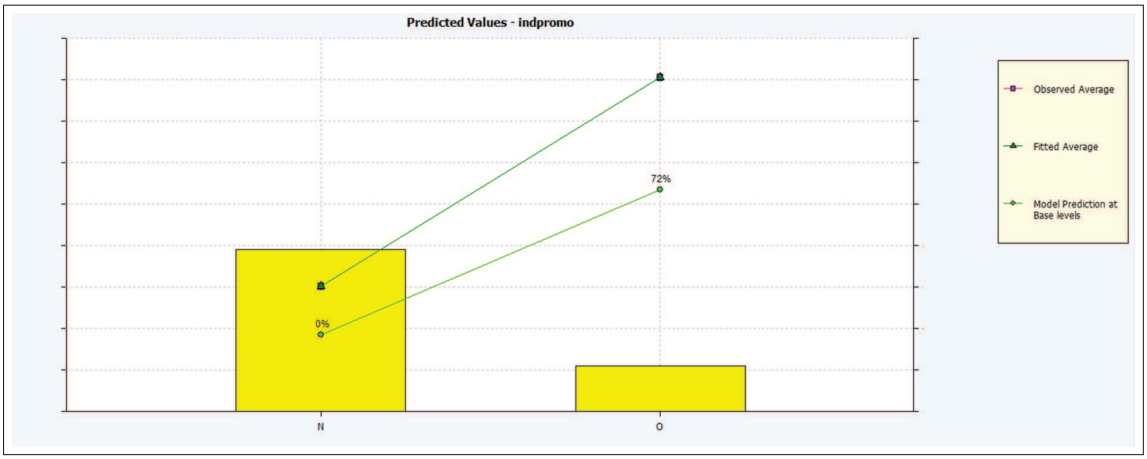


Figure 3 – Modélisation du taux de transformation - Valeurs prédites *indpromo* - GLM

Toutefois, définir le contexte et modéliser le taux de transformation ne suffisent pas pour optimiser l'usage des promotions. Étudier les interactions entre variables, et plus particulièrement celles avec la variable *indpromo* est nécessaire. Pour cela, nous mettons en œuvre une régression pénalisée avec la méthode Lasso. Cette régression permet d'établir un bon compromis entre parcimonie et significativité, en tenant compte de la complexité de chaque variable. Elle permet de tester toutes les interactions entre une variable d'intérêt et les autres, et de ne conserver que les plus significatives et moins coûteuses. Le modèle ainsi sélectionné est optimal en complexité et significativité. Les interactions qui se démarqueraient, constitueraient nos critères d'usage de la promotion.
Cette régression pénalisée est dans un premier temps réalisée sur l'ensemble des données, toutes stratégies commerciales confondues. Il en ressortira certains critères. Cependant, la présence de toutes les CR affaiblit certainement certaines interactions, c'est pourquoi, cette régression pénalisée est réitérée sur les CR du *cluster* n°1, celui caractérisé par un faible taux de transformation malgré un usage intensif de promotions. De cette manière, nous espérons trouver une explication aux résultats de transformation de cette typologie, tout en identifiant de nouveaux critères d'usage.

Finalement, après plusieurs régressions pénalisées, axées sur différents aspects de la promotion (présence/absence, montant), il nous est possible de préconiser un ensemble de critères

optimaux d'utilisation de celle-ci.
Ainsi, la promotion commerciale semble influencer positivement la transformation lorsque :

- le véhicule est déjà assuré ;
- le conducteur dispose d'une ancienneté de permis ;
- le conducteur est âgé ;
- le projet est réalisé par un vendeur réseau.

Néanmoins, ces critères ne constituent qu'une préconisation et ne sont pas absolus.

Toutes les étapes qui constituent ce mémoire et qui ont permis la modélisation du comportement client à la souscription et d'aboutir à des critères optimaux de la promotion commerciale sont résumées dans le schéma récapitulatif ci-dessous.

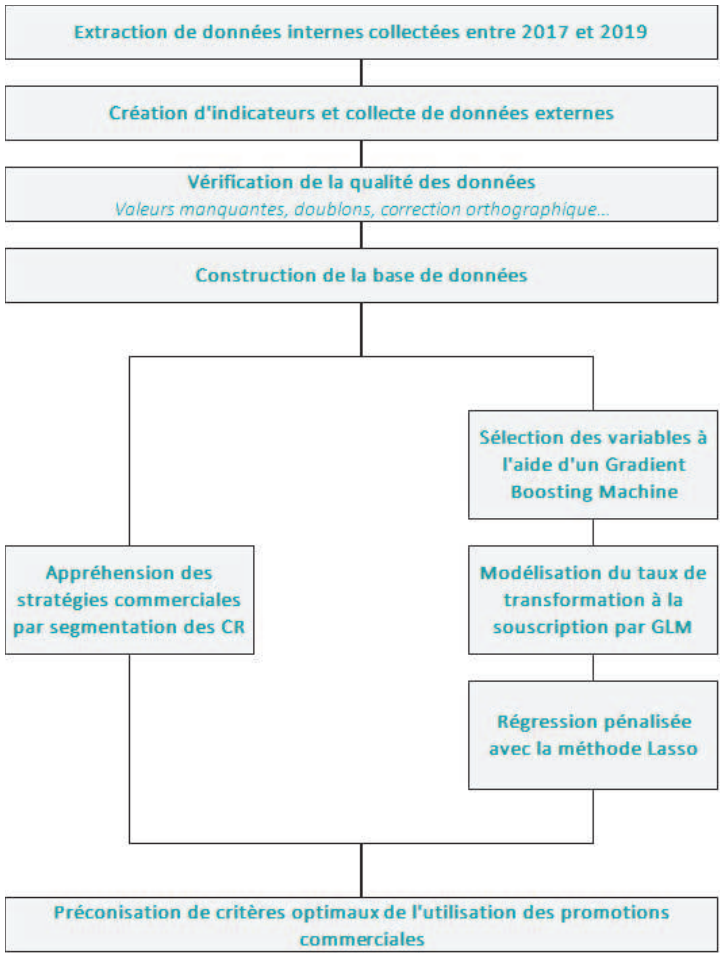


Figure 4 – Étapes de construction dans le processus d'optimisation de la promotion par préconisation de critères optimaux

Enfin, des axes d'amélioration sont envisageables afin d'affiner nos préconisations. Pour cela, il pourrait être intéressant de compléter ces critères par une étape de *Market Basket Analysis* (MBA), afin d'approfondir ces premiers résultats. Il serait également judicieux de mesurer l'impact réel des promotions sur les critères identifiés et ainsi de les valider en procédant à une approche *uplift*. Dès lors, on serait en capacité de distinguer les segments clients sur lesquels la promotion est efficace, et ceux qui auraient souscrit même en absence de promotions.
Enfin, calculer la marge associée à chaque segment client permettrait d'identifier les segments les plus rentables et ainsi de les favoriser dans l'allocation des promotions.

Remerciements

Tout d'abord, je tiens à remercier Lionel FERAUD responsable de la Direction Marché des Particuliers, et Yann MERCUZOT, responsable du service Actuariat Produits, Réassurance et CEDS de m'avoir permis de réaliser mon apprentissage au sein de Pacifica, et pour leur apport professionnel.

Merci à Olivier LOPEZ, mon tuteur académique, pour ses conseils et son accompagnement.

Je tiens à remercier tout particulièrement Audrey MAHUZIER, ma manager et responsable du Pôle Particuliers, pour son encadrement, son temps, ses explications et sa bienveillance à mon égard. Et Juan CALDERON mon tuteur, pour son aide et sa disponibilité qui m'ont été utiles au quotidien dans mon travail et dans la rédaction de ce mémoire.

A toute l'équipe DPART - Actuariat Produits, Réassurance et CEDS, un grand merci pour leur disponibilité, leur aide et leur convivialité.

Enfin, je remercie toutes les personnes de Pacifica avec lesquelles j'ai pu travailler, pour leur accueil cordial et leur bienveillance.

Sommaire

Introduction

En France, l'assurance automobile est obligatoire depuis 1958, et représente un budget moyen de 660€ [1] par an et par véhicule pour un foyer en 2017. Cette obligation d'assurance génère une véritable économie, confrontée depuis quelques années à une vive concurrence (renforcée par des réglementations) et dépendant d'un consommateur de plus en plus exigeant et mobile.

Suivre et maîtriser son taux de transformation (appelé également taux de concrétisation) demeure essentiel pour une compagnie d'assurances. Cette analyse permet entre autres de mesurer la performance d'un produit, mais également d'appréhender le comportement des consommateurs et d'optimiser les stratégies commerciales, notamment au travers des promotions commerciales.
Ces promotions commerciales constituent un véritable levier des ventes pour les compagnies d'assurances. Cependant, un usage excessif affaiblit leurs portées en affectant les aspects positifs et stimulants. L'appréciation de la promotion dans le processus de vente est donc nécessaire, et s'intègre parfaitement à l'analyse de la sensibilité au prix.

Une des particularités du réseau Pacifica, est la décentralisation de l'offre commerciale : cette dernière est à la main de chaque caisse régionale (CR), réseau d'agences distributrices, engendrant ainsi autant de stratégies commerciales qu'il y a de caisses.

L'objectif de ce mémoire est de comprendre quelles sont les stratégies commerciales actuellement mises en place dans le réseau, puis, d'identifier les clients sensibles aux promotions afin d'optimiser l'usage de ces dernières en ciblant ce segment de client, en préconisant des critères optimaux d'utilisation de la promotion.

Dans un premier temps, nous exposerons l'environnement dans lequel cette étude a été menée, avant de présenter succinctement les différents traitements mis en œuvre pour l'élaboration de la base de données. Puis, nous caractériserons la situation actuelle concernant l'usage des promotions commerciales, afin de mettre en lumière les différentes stratégies commerciales qui existent au sein du réseau Pacifica. Pour finir, nous modéliserons le taux de transformation en identifiant quelles sont les variables qui l'expliquent le plus, avant d'identifier les clients sensibles aux promotions, et de proposer des critères optimaux d'utilisation de la promotion.

1. D'après l'ACA (*Automobile Club Association, Budget de l'Automobiliste* – mars 2018)

Chapitre I

Éléments théoriques

1 Les modèles

Afin d'expliquer le taux de transformation, différentes méthodes ont été mises en œuvre dans ce mémoire. Le principe des arbres de décision et des méthodes d'agrégation (*GBM*, *Random Forest*), ainsi que les modèles linéaires généralisés et la régression pénalisée seront présentés dans cette partie. Ces derniers offrent l'avantage d'être facilement interprétables, tandis que les premiers vont nous permettre d'identifier les variables qui sont importantes et celles qui sont en interactions implicites sans hypothèses *a priori*.

1.1 *Random Forest* et *Gradient Boosting Maching*

Avant de présenter les méthodes d'agrégation, il faut présenter brièvement les arbres de décision et le *Bagging* (*Boostrap Aggregation*), qui constituent la base des méthodes de *Random Forest* et *Gradient Boosting Maching*. Cette section a pour vocation de présenter le principe de ces méthodes. Pour plus d'informations et de détails, se référer au livre *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*[2].

1.1.1 Arbre de décision

En apprentissage supervisé, l'arbre de décision permet une première visualisation de nos données sous la forme d'une classification des données. Elle se base sur un algorithme consistant à construire un arbre en partant du sommet, et en le subdivisant jusqu'à ses feuilles. Cette construction repose sur une succession de tests sur les valeurs d'un ensemble de variables, c'est la variable qui maximisera le critère qui sera sélectionnée pour constituer un nœud.

Les arbres de décision nous permettent de faire une analyse rapide de notre jeu de données ; cependant, l'un des inconvénients de cette méthode est qu'elle est « instable ». Le moindre changement, aussi léger soit-il, produira des arbres très différents. Et si ces changements ont lieu au niveau des nœuds proches de la racine, alors l'arbre final sera très différent lui aussi. Pour résoudre ce problème, on peut réaliser deux autres méthodes : le *Bagging* et/ou les *Random Forest*.

Ces méthodes visent à réduire la variance d'estimation.

1.1.2 *Boostrap*

A la base de ces méthodes se trouve la technique de *Boostrap*, qui consiste à créer de «nouveaux échantillons», par tirage au hasard dans l'ancien échantillon avec remise.

2. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction* de Trevor Hastie, Robert Tibshirani et Jerome Friedman

1.1.3 *Bagging*

Afin d'améliorer la classification d'un arbre de décision, on peut réaliser un *Bagging*. En *Bagging*, on considère que les arbres de décision sont des « classifieurs faibles » (c'est-à-dire pas beaucoup plus efficace qu'une classification aléatoire). Le *Bagging* a donc pour but de réduire la variance de l'estimateur, et ainsi, corriger l'instabilité des arbres de décision : chaque classifieur est construit à partir d'un tirage aléatoire des variables avec remise , afin de diminuer la variance.

Pour chaque sous-ensemble de données (créé à l'aide d'un échantillonage *Boostrap*) , on entraîne un arbre de décision, qui nous donne un estimateur. Ce sont ces estimateurs qui seront moyennés dans le cas d'une régression, ou estimés par un « vote » à la majorité dans le cas d'une classification. C'est ici qu'intervient l'étape *aggregation*. C'est pendant cette dernière, qu'on sera en mesure de réduire la variance.

1.1.4 *Random Forest* (RF)

Toujours en vue d'améliorer notre modèle, et dans la continuité du *Bagging*, on peut réaliser un *Random Forest* (RF), également appelé « forêts aléatoires ». Les forêts aléatoires, consistent, elles aussi, à réduire la variance des prévisions d'un arbre de décision, afin d'améliorer la performance.

Il s'agit d'un apprentissage en parallèle : c'est-à-dire que les différents estimateurs peuvent être construits parallèlement, que l'algorithme apprend sur plusieurs arbres de décision construits aléatoirement et entraînés sur des sous-ensembles de données différents. Autrement dit, chaque arbre apprend sur un sous-ensemble aléatoire de données selon le principe du *Boostrap*, avec un sous-ensemble aléatoire de *features* (variables tirées aléatoirement) selon le principe des « projections aléatoires ». Les prédictions sont ensuite moyennées, lorsque les données sont quantitatives ou utilisées pour un vote pour des données qualitatives, dans le cas des arbres de classification, ce qui est notre cas.

1.1.5 *Gradient Boosting Maching* (GBM)

L'algorithme GBM est un cas particulier du *Boosting*. Tout comme le RF, on utilise plusieurs modèles au lieu d'un seul. La différence avec les RF est qu'au lieu de faire la moyenne de l'ensemble des prédicteurs, le *Boosting* travaille de manière séquentielle.

Le *Boosting* commence par construire un premier modèle qu'il va évaluer. En fonction de l'erreur de prédiction, chaque individu sera pondéré. L'objectif est de donner un poids plus important aux individus pour lesquels la valeur a été mal prédite pour la construction du modèle suivant. Ainsi, les arbres vont s'ajouter progressivement à l'ensemble, l'arbre final étant obtenu en tenant compte de l'apprentissage qui a été réalisé sur les arbres précédents.

L'algorithme GBM utilise le gradient de la fonction de perte, pour le calcul des poids des individus lors de la construction de chaque nouveau modèle.

1.1.6 *XGBoost*

Un dernier algorithme, non exploité dans ce mémoire mais utilisant également les arbres de décision pour résoudre des problèmes de classification et de régression, l'XGBoost (*ou eXtremeGradientBoosting*).

Sa différence avec les méthodes vues précédemment, est qu'au lieu d'utiliser un seul modèle, l'algorithme va en utiliser plusieurs, qui seront ensuite combinés pour obtenir un seul résultat.

Cet algorithme est donc plus lent que les forêts aléatoires, mais il permet à l'algorithme de s'améliorer par capitalisation par rapport aux exécutions précédentes. Le premier modèle va

nous fournir son évaluation, et à partir de cette évaluation, chaque observation sera pondérée en fonction de la performance de la prédiction.

1.2 Quelques rappels sur la régression linéaire

Soit y une variable quantitative que l'on souhaite prédire, et $(x_{i,j})_{1\leq j\leq p}$, p variables explicatives (pouvant être réelles ou catégorielles). L'objectif est de trouver les coefficients $\beta = (\beta_j)_{0\leq j\leq p}$ tel que $\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$ approxime au mieux la variable y.
Ainsi le modèle peut s'écrire, $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \epsilon$ où ϵ correspond à l'erreur du modèle, c'est-à-dire à la différence entre la valeur observée et la valeur prédite. Et sous forme matricielle on a, $Y = X\beta$.

Pour trouver les coefficients $(\beta_j)_{0\leq j\leq p}$ on utilise la méthode des moindres carrés, c'est-à-dire qu'on cherche à minimiser l'erreur générée par le modèle.

$$\beta = argmin_{(\beta_j)_{0\leq j\leq p}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2$$

où N représente le nombre d'individus.
Finalement, on a $\hat{y}_i = \sum_{j=1}^{p} \beta_j x_{i,j}$ pour $i \in 1, ..., N$ la prédiction donnée par le modèle pour y_i.
Intuitivement, lorsqu'on souhaite expliquer une variable Y par rapport à plusieurs variables explicatives $X^j (j = 1, ..., p)$, on pense aux modèles linéaires gaussiens classiques. Cependant, ces modèles supposent que la variable réponse suit une loi normale, que son espérance dépend linéairement des variables explicatives et que sa variance est constante (homoscédasticité). Ces hypothèses pouvant être restrictives, on souhaite s'en émanciper. Pour cela, on utilise une extension des modèles linéaires classiques, les GLM (*Generalized Linear Models*) qui permettent de gérer d'autres distributions.

Pour la suite on pose,
Y la variable aléatoire réelle qu'on observe et qu'on souhaite expliquer et/ou prédire.
Soit $Y = (Y_1, ..., Y_n)'$ le vecteur aléatoire de $\mathbb{R}^n$.
Et $X = (X^1, ..., X^p)$ l'ensemble des variables explicatives [3].
Enfin, on conserve l'hypothèse selon laquelle les variables aléatoires Y|X sont générées de manières indépendantes.

1.2.1 Définition du modèle

Les modèles linéaires généralisés (GLM) sont présentés sous ce nom pour la première fois en 1972 par Nelder et Wedderburn, et ils se caractérisent par 3 composantes.

1.2.2 La composante aléatoire

La composante aléatoire, permet d'identifier la loi de probabilité de la variable à expliquer Y, dont les densités appartiennent à la famille exponentielle et s'écrivent

$$f(y_i|\mu_i, \phi) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right)$$

3. Les variables explicatives sont également appellées : les prédicteurs, les variables indépendantes ou encore les variables exogènes. Et pour les variables à expliquer, on parle de variable réponse, variable dépendante, ou de variable endogène.

où a, b et c, sont des fonctions (connues et dérivables) et dépendent du type de la loi exponentielle. ϕ est le paramètre de dispersion et est supposé connu, tandis que θ_i, le paramètre canonique, est inconnu et sera développé un peu plus bas.

Le plus souvent, la nature du problème étudié influe sur le choix de la loi de probabilité. Dans le cas de notre étude, on est sur un problème de type « succès/échecs » («õui/non »), avec une variable réponse binaire. Y_i peut donc être distribué selon une loi de Bernoulli, $\mathcal{B}$(1,p) dont la densité s'écrit :

$$\begin{aligned} f(y) &= p^y(1-p)^{1-y} \\ &= e^{yln(p)+(1-y)ln(1-p)} \\ &= e^{y(ln(p)-ln(1-p))+ln(1-p)} \\ &= e^{yln\left(\frac{p}{1-p}\right)+ln(1-p)} \end{aligned}$$

et est de la forme espérée avec $\theta = ln\left(\frac{p}{1-p}\right)$, $a(\theta) = nln(1+e^{\theta})$ et $\phi = 1$.

1.2.3 La composante déterministe

La composante déterministe, également appelée prédicteur linéaire, précise quels sont les prédicteurs. Elle s'exprime sous la forme d'une combinaison linéaire, où les $\beta_j (j = 1, ..., p)$ sont des coefficients, des paramètres non observés qui seront à estimer. Soit η le prédicteur linéaire, le vecteur à n composantes,

$$\eta = X\beta$$

où $X = (x_{i,j})_{1 \leq i \leq n, 1 \leq j \leq p}$ la matrice n*p des variables explicatives, et β un vecteur de p paramètres.

1.2.4 La fonction de lien

Cette dernière composante détermine la relation entre la composante aléatoire et la composante déterministe. Elle permet de spécifier comment l'espérance mathématiques de Y est liée au prédicteur linéaire construit à partir des variables explicatives.

En effet en GLM, on estime l'espérance conditionnelle de Y aux covariables.

La fonction de lien g, bijective et différentiable telle que :

$$g(\mu_i) = x_i'\beta$$

qui lie le prédicteur linéaire $\eta_i = x_i'\beta$ à la moyenne μ_i de Y_i (car on rappelle, $\mu_i = \mathbb{E}(Y_i|X_i)$).

Chacune des lois de la famille exponentielle possède une fonction de lien spécifique, appelée fonction canonique, qui permet de relier l'espérance μ au paramètre naturel (=canonique), θ.

$$g_\star(\mu) = \theta$$

Ainsi, on peut résumer le modèle GLM comme la relation linéaire entre $\mathbb{E}(Y_i|X_i)$ et les variables explicatives X^j où :

- Y |X = x $\sim$P$_{\theta,\phi}$ appartient à une famille exponentielle ;
- $g(\mathbb{E}(Y|X)) = g(\mu(X)) = X\beta$ pour une certaine fonction de lien g, bijective.

1.2.5 Estimation des paramètres

1.2.5.1 Estimation de ϕ

Son estimation est secondaire car c'est un paramètre de nuisance. Sa valeur n'influencera pas la maximisation de la vraisemblance de β, donc on ne s'attardera pas sur ce point. Cependant, si besoin ϕ peut être estimé par maximum de vraisemblance.

1.2.5.2 Estimation des β

Lorsque Y est distribuée selon une loi appartenant à une famille exponentielle, on montre en Annexe I.A, que :

$$\mathbb{E}(Y) = b'(\theta) \;\; et \;\; Var(Y) = \phi b''(\theta) \tag{I.1}$$

Pour n observations supposées indépendantes, (et en tenant compte que θ dépend de β), la log-vraisemblance s'écrit :

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

où $\ell(\theta_i, \phi; y_i)$ correspond à la contribution de la i^{eme} observation à la log-vraisemblance.

L'estimation des paramètres β_j est calculée en maximisant la log-vraisemblance du modèle linéaire généralisé, pour cela il nous suffit de dériver la log-vraisemblance par rapport au paramètre β, et d'écrire les conditions du premier ordre. Cette partie est développée en Annexe I.B.

On obtient ainsi les équations suivantes,

$$\sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \eta_i}{Var(Y_i)} * x_{i,j} * \frac{\partial \mu_i}{\partial \eta_j} \tag{I.2}$$

$$= 0 \qquad pour\ tout\ j \tag{I.3}$$

Dans le cas général, on ne sait pas résoudre explicitement cette équation. L'EMV (Estimation du Maximum de Vraisemblance) est donc calculée numériquement par méthodes itératives, souvent grâce à l'algorithme Newton-Ramphson. On obtient ainsi le score et l'information de Fisher.

Définition 1.1 (Score et information de Fisher)**.** *Soit $U_n(\beta) = (u_1, ..., u_p)'$ où*

$$u_j = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} logf\left(y_i; g^{-1}(X_i\beta)\right)$$

On dit que $U_n(\beta)$ est le vecteur des scores. L'information de Fisher est définie par

$$\mathcal{I}_n(\beta) = \mathbb{E}[U_n(\beta)U_n(\beta)']$$

Théorème 1.1 (Consistance et normalité asymptotique)**.** *Sous des hypothèses de régularité ($\beta \in \Theta$ ouvert, convexe, g deux fois continûment différentiable, conditions sur X pour que $I_n(\beta)$ soit définie positive)*

1. *$\hat{\beta}$ existe et est consistant*
2. *$\hat{\beta}$ est asymptotiquement normal*

$$\mathcal{I}_n(\beta)^{1/2}(\hat{\beta}\text{-}\beta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_P(0, \mathcal{I}_P)$$

où, $\mathcal{I}_P$ la matrice identité dans $\mathbb{R}^P$

Ce qui permet par la suite de construire des intervalles de confiance pour β et de tester la significativité des variables explicatives en regardant la non nullité des coefficients $\hat{\beta}_1, ..., \hat{\beta}_p$

1.2.6 Adéquation du modèle et calcul de la déviance

Une fois nos paramètres estimés, il nous faut juger de la qualité du modèle. Reflète-t-il bien la réalité ? Est-il en adéquation avec les données ?
Pour cela, on va comparer le modèle qu'on a ajusté $\mathcal{M}$ à un modèle plus général, dit modèle saturé que l'on note $\widetilde{\mathcal{M}}$. Ce dernier est également un modèle linéaire généralisé, avec la même distribution et la même fonction de lien que le modèle $\mathcal{M}$, sauf qu'il a autant de paramètres que de variables réponses.

Notons, $\tilde{\ell}$ la vraisemblance du modèle saturé et $\hat{\ell}$ la vraisemblance du modèle estimé $\mathcal{M}$.
La méthode la plus intuitive pour comparer des vraisemblances est de faire un test de rapport de vraisemblance. Notons alors $log(\lambda) = \tilde{\ell} - \hat{\ell}$.
Cependant avec cette méthode, nous ne connaissons pas le seuil à partir duquel on peut affirmer que la valeur de $log(\lambda)$ est grande, et donc, affirmer que le modèle estimé décrit mal les données par rapport au modèle saturé.

La solution est d'utiliser la déviance, définie comme :

$$\Delta = 2log(\lambda) = 2log(\tilde{\ell} - \hat{\ell})$$

qu'on peut réécrire comme :

$$\Delta = 2\sum_{i=1}^{n} \frac{Y_i(\tilde{\theta}_i - \hat{\theta}_i) + a(\tilde{\theta}_i) - a(\hat{\theta}_i)}{\phi}$$

De plus, d'après le théorème de Wilks, la déviance peut être approchée par une loi du χ^2_{n-p} où p correspond au nombre de paramètres à estimer dans le modèle estimé $\mathcal{M}$.

Théorème 1.2 (Théorème de Wilks)**.** *Si les hypothèses du modèle GLM sont vérifiées, alors*

$$\Delta \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_{n-p}$$

avec p le nombre de paramètres à estimer.

Ainsi, comparer le modèle estimé au modèle saturé, revient à comparer la déviance à une loi du χ^2 au bon nombre de degrés de liberté.

1.2.7 Les limites du GLM

Même si les GLM permettent d'aborder des problèmes statistiques plus complexes que le modèle linéaire classique en s'émancipant d'hypothèses contraignantes. Il subsiste la nécessité de faire des hypothèses *a priori*, que ce soit sur la loi de la variable d'intérêt ou sur les interactions entre les variables explicatives.

Même si dans notre cas, la loi d'une variable binaire est nécessairement une Bernoulli et que la fonction de lien, *logit* est connu, en général, le choix de la loi de la variable à expliquer constitue une hypothèse, car nous n'avons aucune certitude. Si l'hypothèse est fausse, on prend donc un risque dans la modélisation. De plus, nous sommes restreints à des lois de probabilité appartenant à la famille exponentielle, et cette dernière impose de ne pas avoir de valeurs extrêmes. Ce qui peut constituer une limite, quand on tente de modéliser une variable réponse continue (ex: un coût).

Enfin, les interactions entre les variables explicatives sont modélisables, mais elles reposent fortement sur la connaissance du métier, de même pour la sélection des variables.
Même si cette dernière s'est faite sur des critères cohérents tels que l'AIC ou le BIC, nous ne regardons pas le cas des variables couplées. Enfin, le nombre d'interactions est potentiellement

très important, et il n'est pas toujours possible de les tester toutes, en raison de capacités informatiques limitées.

Cette obligation de faire des hypothèses *a priori* dans les modèles statistiques « classiques » nous amène à envisager d'autres méthodes de modélisation, notamment en nous intéressant aux méthodes non-paramétriques, et plus particulièrement aux méthodes d'ensembles (aggrégation de prédicteurs).

1.2.8 L'effet relatif en lien logistique

On a vu précédemment que les modèles linéaires généralisés peuvent s'écrire de façon générale comme

$$Y|X \sim \mathcal{L}(\theta_x, \phi)$$

avec $\theta_x = h[\mathbb{E}(Y|X = x) = g(x^t\beta)]$.

Dans le cas de notre étude, sommes sur une distribution de Bernoulli, dont la fonction de lien canonique est *logit*, d'où

$$Y \sim B(\pi_i) \qquad Y|X = x \sim \mathcal{B}(H(x^t\beta)) \qquad avec\ H(s) = \frac{e^s}{1+e^s}$$

$$g(\pi_i) = logit(\pi_i) = log\left(\frac{\pi_i}{\pi_i}\right) = \beta_0 + \sum_{k=1}^{n} \beta_k x_{i,k} = \eta_i$$

ainsi,

$$\mathbb{P}(Y_i|X_i = x) = \mathbb{E}\left[Y_i|X_i = x\right] = \frac{exp(\eta_i)}{1 + exp(\eta_i)} = \frac{exp[\beta_0 + \sum \beta_k x_{i,k}]}{1 + exp\left[\beta_0 + \sum \beta_k x_{i,k}\right]} \tag{I.4}$$

Les β ne sont pas directement interprétables, car ce n'est pas une structure multiplicative, contrairement à la régression de Poisson, $Y|X \sim P(\lambda_i)$, où la fonction de lien canonique est le logarithme. En effet, l'équation (I.1) n'est pas simplifiable ; il n'est pas possible d'isoler β_k. Contrairement au lien logarithme où β_k est isolable, autrement dit, il y a un effet pur par construction.

Cet effet pur par construction est montré juste ci-dessous. La fonction de lien d'une loi de Poisson, $P(\lambda_i)$, s'écrit :

$$g(\lambda_i) = log(\lambda_i) = \eta_i \qquad où\ \eta_i = \beta_0 + \sum \beta_j x_{i,j}$$

dès lors, on peut isoler λ_i, et on trouve $\lambda_i = exp(\eta_i)$ ainsi :

$$\mathbb{P}(Y_i|X_i = x) = \mathbb{E}(Y_i|X_i = x) = exp(\eta_i) = exp(\beta_0 + \sum \beta_j x_{i,j}) = exp(\beta_0) \prod exp(\beta_j x_{i,j})$$

On reconnaît là une structure multiplicative. A partir de laquelle il est facile d'isoler un β et d'en connaître son effet relatif, ce qui revient à connaître l'effet pur de la variable associée.

Pour X_k la $k^{\text{ième=}}$ variable aléatoire à valeur dans $\mathbb{N}$, l'écart relatif peut s'écrire :

$$\begin{aligned}
Ecart\ relatif &= \frac{\mathbb{E}[Y_i|(X_{i,j})_{j\neq k} = x_{i,j}, X_{i,k} = \ell + 1]}{\mathbb{E}[Y_i|(X_{i,j})_{j\neq k} = x_{i,j}, X_{i,k} = \ell]} - 1 \\
&= \frac{\mathbb{E}[Y_i|X_{i,j} = x_{i,j} * \mathbf{1}_{j\neq k} + (\ell + 1) * \mathbf{1}_{j=k}]}{\mathbb{E}[Y_i|X_{i,j} = x_{i,j} * \mathbf{1}_{j\neq k} + \ell * \mathbf{1}_{j=k}]} - 1 \\
&= \frac{exp(\beta_0 + \beta_k(\ell + 1) + \sum_{j\neq k}^{n} \beta_j x_{i,j})}{exp(\beta_0 + \beta_k \ell + \sum_{j\neq k}^{n} \beta_j x_{i,j})} - 1 \\
&= exp(\beta_k) - 1
\end{aligned}$$

Ainsi, l'écart relatif lié à la variable X_k prenant la valeur $\ell + 1$, vaut $exp(\beta_k)$-1, ce qui revient à l'effet pur de cette variable. Soit, on a $exp(\beta_k)$-1 % de chance en plus d'impacter Y.

En théorie, modéliser un problème de classification (binaire) avec la fonction de lien logarithme est faux, car inapproprié : le logarithme n'est pas la fonction de lien canonique d'une distribution de Bernoulli.
En revanche, en pratique cela permet de faire une première lecture, en donnant un premier aperçu de l'effet pur par construction d'une variable : « l'effet pur de la variable, c'est $exp(\beta) - 1$ ».

1.3 La régression pénalisée - La méthode *Lasso*

La méthode *Lasso* (*Least Absolute Shrinkage and Selection Operator*), constitue une méthode de régression pénalisée. Elle consiste à ajouter une pénalité au modèle, afin de favoriser des solutions parcimonieuses.

Ainsi, deux usages se dégagent : améliorer la robustesse du modèle, et réaliser une sélection de variables. En effet, en contraignant les coefficients « petits » à être nuls, les autres coefficients sont rendus plus significatifs, générant ainsi un modèle de plus petite dimension, un modèle plus parcimonieux et d'écarter les interactions inutiles.

Dans le cas de la régression linéaire, l'estimateur *Lasso* de β est définit par :

$$\hat{\beta}_{Lasso} = argmin_{\beta\in\mathbb{R}^p} \left(\sum_{i=1}^{n} (y_i - \sum_{j=0}^{p} \beta_j x_{i,j})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

où λ est un paramètre positif, à choisir.

2 Optimisation des hyperparamètres

L'optimisation des hyperparamètres consiste à chercher pour un algorithme donné, les paramètres qui minimisent le risque du modèle. En pratique, on cherche les paramètres qui maximisent la performance sur les échantillons de validation. On peut le faire en entraînant son modèle sur une base d'apprentissage et en l'évaluant sur une base de test, ou en réalisant une validation croisée [4] (ou *cross validation*).

Cette étape d'optimisation des hyperparamètres a été réalisé pour chaque méthode présentée ci-dessus.

4. Procédure consistant à diviser en k sous-échantillons (ou plis) le jeu de données. Chaque pli sera ensuite utilisé une fois comme validation tandis que les k - 1 plis restants forment l'ensemble d'apprentissage.

3 Mesure de performance et *Odds ratios*

Dans cette section, nous allons présenter des outils permettant de calculer la performance des algorithmes de *Machine Learning* et de valider nos modèles. Habituellement en régression, on calcule l'erreur quadratique moyenne pour valider les modèles. Tandis qu'en classification, on regarde des indicateurs fournis par la matrice de confusion et la courbe ROC.
Cependant, il n'existe pas d'indicateur unique ou meilleur qu'un autre. Pour valider un modèle, il nous faut donc en regarder plusieurs à la fois, avec une préférence pour ceux en lien avec la spécificité du problème.

3.1 Matrice de confusion

En classification binaire, la plupart des métriques sont évaluées à partir de la matrice de confusion résumant les résultats d'un modèle appliqué à l'échantillon *test*.
De manière générale, on peut écrire un classifieur binaire comme,

Définition 3.1.

$$\hat{y}(x) = \mathbf{1}\left\{\pi(X) > \tau\right\}$$

où $\pi(X)$ désigne la probabilité associée au classifieur, et τ un seuil (par défaut, 0,5) tel que si $\hat{\pi}(X) > \tau$ alors $\hat{y}_i = 1$ et si $\pi(X)\tau$

Dans le cadre d'une variable réponse binaire, la matrice de confusion se présente ainsi :

	$\hat{y}_i = 1$	$\hat{y}_i = 0$
$y_i = 1$	VP (vrai positif)	FN (faux négatif)
$y_i = 0$	FP (faux positif)	VN (vrai négatif)

A partir de cette dernière et des éléments qu'elle contient, on peut calculer plusieurs indicateurs. Nous nous limiterons à définir certains critères, ceux les plus souvent utilisés.

Cependant, la matrice de confusion s'appuie seulement sur les prédictions $\hat{y}_i$, et elle ne prend pas en considération les probabilités estimées $\hat{\pi}(x_i)$.

3.2 La précision totale, ou *accuracy*

Elle correspond au nombre de prédictions « correctes » rapportées au nombre total de prédictions. En classification binaire, on peut calculer en termes de positif et négatif.

$$Precision\ totale = \frac{VP + VN}{VP + VN + FP + FN}$$

Cet indicateur est très généraliste, on ne peut pas se fier à lui seul, car il ne prend pas en considération les faux négatifs, et les faux positifs. De plus, si l'ensemble de données est déséquilibré, la précision n'est pas fiable.

3.3 *Precision*

Cette métrique répond à la question : Quelle est la part de prédictions positives vraiment correctes?, autrement dit, elle mesure la capacité du modèle à refuser les solutions non pertinentes. La *precision* s'écrit :

$$Precision = \frac{VP}{VP + FP}$$

3.4 Rappel, ou *recall*

A l'inverse, le *recall* mesure la capacité du modèle à donner toutes les solutions pertinentes, et tend à répondre à la question : Quelle est la part de prédictions positives correctement identifiées ? Et s'écrit :

$$Recall = \frac{VP}{VP + FN}$$

Precision et *recall* sont deux indicateurs à regarder ensemble pour avoir une bonne interprétation des résultats d'un modèle.
En effet, ces derniers sont antagonistes, c'est-à-dire que l'amélioration de l'un se fait souvent au détriment de l'autre ; il faut donc trouver un équilibre.
Un modèle avec une très bonne prédiction (i.e une probabilité proche de 1), se traduira comme une quasi absence de FP, ce qui peut sembler très bien pour un modèle, mais en contrepartie si le *recall* est très faible (proche de 0), cela voudra dire que l'identification des positifs est très faible.
En conclusion, plus ces deux indicateurs sont élevés, plus la qualité du modèle l'est.

3.5 Spécificité et sensibilité

La spécificité mesure la capacité à prédire un résultat négatif lorsque l'hypothèse n'est pas vérifiée, autrement dit, parmi tous les négatifs réels.
Elle s'oppose à la sensibilité, qui correspond à la capacité de prédire un résultat positif parmi les positifs réels.

$$Spécificité = \frac{VN}{VN + FP} \quad et \quad Sensibilité = \frac{VP}{VP + FN}$$

Comme dans le cas de la *precision* et du *recall*, ces deux indicateurs doivent être regardés ensemble.

3.6 *F-mesure*

La *F-mesure* ou score F_1 est un compromis entre le *recall* et la *precision*, elle permet de trouver un équilibre entre ces deux métriques et de donner la performance du système. En effet, il s'agit de la moyenne harmonique de la *precision* et du *recall*.
Elle mesure la capacité du modèle à donner toutes les solutions pertinentes, et à refuser les autres.
On définit la *F-mesure* comme suit,

$$F_{mesure} = F_1 = 2 * \frac{P * R}{P + R}$$

Elle pondère de façon égale le *recall* et la *precision*. Elle rend compte de la qualité d'une classification en fonction des classes, mais ne tient pas compte de l'éventuel déséquilibre entre les classes. Et elle ne tient pas non plus compte des faux négatifs.

3.7 Courbe ROC et AUC

Pour construire la courbe ROC (*Receiver Operating Characteristic*), on a besoin de calculer pour différents seuils $\tau \in [0,1]$ le taux de vrai positif, TVP, et le taux de faux positif, TFP, qu'on définit respectivement comme:

$$RVP_\tau = \frac{nombre\ de\ bonnes\ estimations\ positives}{nombre\ de\ positif} = \frac{\sum_{i=1}^{N} \mathbf{1}_{\{y_i=1 \cap \hat{y}_i\}}}{N_p}$$

$$RVP_\tau = \frac{nombre\ de\ fausses\ estimations\ positives}{nombre\ de\ négatif} = \frac{\sum_{i=1}^{N} \mathbf{1}_{\{y_i=1 \cap \hat{y}_i\}}}{N_n}$$

où N_p correspond au nombre de positif, et N_n au nombre de négatif. On rappelle également que $\hat{y}_i$ est défini positif (i.e prend la valeur de 1) quand $\hat{p}_i \geqslant \tau$.

La courbe ROC est obtenue en traçant les points ($RFP_{(\tau)}$,$RVP_{(\tau)}$) pour différents seuils de τ.

Par construction de la courbe, aux coordonnées (0,0), le classifieur est toujours négatif et il n'y a aucun faux positif et aucun vrai positif. Aux coordonnées (1,1), le classifieur est toujours positif. Et en (0,1), il n'y a aucun faux positif, ni faux négatif c'est-à-dire que chaque classe est prédite correctement : le classifieur est donc de meilleure qualité.

En intégrant la fonction ROC, on obtient l'AUC (*Area Under the Curve*) qu'on appelle également aire sous la courbe. Ce dernier indicateur peut être vu comme la qualité d'un « bon » classement d'individu, et il est complémentaire de la courbe ROC.
Si deux courbes ROC sont semblables, on regarde alors l'AUC. L'AUC le plus élevé indique le meilleur modèle de prédiction. Cependant, les valeurs AUC et les courbes ROC sont à regarder ensemble, car les courbes peuvent être performantes sur certaines régions seulement.
Dans le cas où l'on a un classifieur aléatoire, on aura une courbe ROC qui sera égale à la première bissectrice et un AUC de 0,5.

Remarque. *Dans le cas de certains jeux de données (présence d'un déséquilibre dans la représentation de la variable d'intérêt), l'AUC n'est pas toujours fiable.*

3.8 Les *Odds ratios*

Les *Odds ratios* (OR) ou rapport des cotes, permettent de montrer la force d'association entre deux variables binaires / dichotomiques.
Plus la valeur de l'OR est élevée, plus l'association sera importante, et plus il sera probable que tout changement sur l'une des deux variables impactera l'autre.

L'*odds* ou la cote, est un rapport de probabilité entre évènement et non évènement, et s'écrit :

$$C(x) = \frac{p(x)}{1 - p(x)}$$

où *p* est la probabilité de l'évènement.

On peut l'interpréter comme, pour un groupe de X individus présentant l'évènement étudié, *X*C* ne le présentent pas.

L'*Odds ratios* est toujours un nombre positif compris entre 0 et l'infini, et s'écrit :

$$R(x, x') = \frac{C(x)}{C(x')} = \frac{p(x) * (1 - p(x'))}{(1 - p(x)) * p(x')}$$

Son interprétation est la suivante :

- si OR < 1, alors il est peu probable qu'il y ait un lien entre les 2 variables ;
- si OR = 1, il n'y a aucune association entre les 2 variables ;
- si OR > 1, l'association est d'autant plus forte que la valeur de l'OR est grande.

Cependant, il faut faire attention avec l'OR, ce dernier suggère seulement s'il y a une association. Il n'implique pas de rapport de cause à effet.

Dans le cas d'une régression logistique, on a $ln\left(\frac{p(x)}{1-p(x)}\right) = x\beta$, ainsi

$$C(x) = e^{x\beta} \quad et \quad R(x, x') = e^{x\beta - x'\beta}$$

4 Interprétation globales et locales des modèles

4.1 Les graphes de dépendance partielle (DP)

4.1.1 La dépendance partielle

A l'inverse des GLM / modèles linéaires qui sont facilement interprétables grâce aux coefficients, l'interprétation est plus difficile quand les modèles sont obtenus par construction non paramétrique, car cela les rend plus complexes.
Afin de résoudre ce problème, on peut utiliser les graphes de dépendance partielle (*partial dependence plot*, en littérature anglaise) proposés par Friedman.
Ils tendent à montrer la dépendance entre la variable à expliquer Y, et un ensemble de caractéristiques « cibles », x_S, marginalisant les valeurs de toutes les autres caractéristiques, dites « complémentaires », x_C.
Intuitivement, nous pouvons interpréter ces graphes comme la réponse cible Y, plus exactement la variation de cette dernière en fonction de caractéristiques « cibles ».

Si on reprend notre modèle présenté dans la partie 1.2, à *p* variables explicatives X et N observations, nous avons N prédictions de la forme :

$$\hat{y}_i = F(x_{i,1}, x_{i,2}, ..., x_{i,p})$$

La fonction de dépendance partielle pour la régression s'écrit :

$$\hat{f}_{x_S}(x_S) = \mathbb{E}_{x_C}[\hat{f}(x_S, x_C)] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$$

où x_S désigne les caractéristiques « cibles » pour lesquelles la fonction de dépendance partielle doit être tracée, et x_C les autres caractéristiques utilisées dans le modèle. Soit x_S et x_C des sous-ensembles de X, tel que $x_S \cup x_C = X$.

Il s'agit de mesurer l'effet d'une variable x_S (ou de deux) sur les prédictions $\hat{y}_i$ en prenant en compte l'effet moyen de toutes les autres variables explicatives complémentaires x_C.

Elle est approximée par :

$$\hat{f}_{x_S}(x_S) = \frac{1}{N} \sum_{i=1}^{N} \hat{f}(x_S, x_{iC})$$

où x_{iC} correspond à la valeur réelle des variables explicatives complémentaires des N observations.

Ainsi, en moyennant toutes les prédictions sur toutes les observations selon la modalité x prise par la variable explicative x_S, il est possible de voir son impact sur les prédictions, soit son effet pur.

L'extension multivariée est simple en principe, mais en pratique on se limite à un ensemble S constitué de 2 variables maximum pour avoir des graphes interprétables. Ces deux variables sont généralement choisies parmi les variables les plus importantes.

Les graphes de dépendance peuvent être construits pour n'importe quel modèle prédictif quelle que soit sa complexité. Dans le cas d'un modèle linéaire, le modèle prédictif s'écrit $\hat{y}_i = \sum_{j=1}^{p} \beta_j x_{i,j}$. L'approximation de la dépendance partielle d'une variable $x_S = (x_k)$ est

$$\hat{f}_{x_S}(x) = \beta_k x + \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq k} \beta_j x_{i,j} = \beta_k x + \sum_{j \neq k}^{N} \beta_j \bar{x}_j$$

où $\bar{x}_j$ est la valeur moyenne de la j^{ème} variable.

La fonction de dépendance partielle nous permet donc de connaître comment la valeur de la variable $x_S = (x_k)$ influence les prédictions du modèle $\hat{y}_i$ après avoir moyenné l'influence de toutes les autres variables. Cela revient donc à connaître l'effet pur de cette variable $x_S = (x_k)$.

4.1.2 Les limites de la dépendance partielle

L'interprétation des graphes de DP est simple et intuitive, en plus d'être facilement implémentable.
Cependant, cette méthode présente des inconvénients dont le plus important est l'hypothèse de non-corrélation entre les variables de l'ensemble S et celles de C sur laquelle repose cette méthode.
En pratique, cette hypothèse est difficilement vérifiée, ce qui génère des points de données peu probables ou impossibles (comme observer une habitation de $30m^2$ avec 30 pièces).

De plus, cette méthode présente un coût de calcul potentiellement long, car il faut la calculer pour chaque valeur prise par les variables de l'ensemble S.
On est également limité à une représentation bidimensionnelle (c'est-à-dire à un maximum de 2 variables pour l'ensemble S), car au-delà, le graphe devient illisible et donc ininterprétable.
Enfin en cas d'effets hétérogènes, ces derniers peuvent êtres masqués, car les graphes de dépendance partielle ne montrent que les effets marginaux moyens. Par exemple, si on a la moitié d'une variable associée positivement à la prédiction versus l'autre moitié associée négativement, alors la courbe pourrait être horizontale par « annulation » des deux moitiés, nous amenant à conclure (à tort) à l'absence d'effet entre la variable explicative et la prédiction.

4.2 Les courbes d'espérance conditionnelle individuelle (ICE)

Pour contourner ce risque, on peut tracer les courbes d'espérance conditionnelle individuelle, appelées également ICE (*Individual Conditional Expectation*).
Cette méthode est rapidement expliquée ci-dessous, et ne sera pas développée plus tard.
Une courbe correspond à une observation pour laquelle on fait varier la valeur de la variable explicative (de l'ensemble S), avant de récupérer la prédiction associée à chaque valeur. De ce fait, chaque ligne illustre l'effet de la variation de la variable sur la prédiction.
Ce processus est réitéré pour chaque individu composant la base de données. Ainsi, les graphes ICE sont une décomposition de l'effet moyen fourni par les graphes de DP, car ce dernier n'est autre que la moyenne des lignes individuelles d'un tracé ICE.

Par conséquent, le graphe ICE permet de mettre à jour des effets hétérogènes qui seraient imperceptibles sur un graphe de DP.
Ces deux graphes sont donc complémentaires et ils apportent une explication globale des prédictions. On parle de méthode d'interprétation globale quand elle permet d'expliquer un modèle dans son ensemble.

Cependant les courbes ICE, tout comme le graphe de DP, reposent sur l'hypothèse d'indépendance des variables et sont limitées en termes de dimension de représentation (on se limitera à 1 ou 2 variables). De plus, un nombre de courbes à tracer trop important surchargera le graphe, il y a donc une limite en terme d'observations étudiées.

4.3 Les valeurs de Shapley

Développées par un mathématicien américain, Lloyd Shapley (1923 – 2016), les valeurs de Shapley (ou *SHAP Values*) reposent sur la théorie des jeux et permettent de décomposer un gain entre joueurs de façon « équitable ». Cette notion est présentée brièvement dans cette partie, pour plus d'informations se référer au document *La valeur de Shapley – Comment individualiser le résultat d'un groupe*[5].

On note E un ensemble de n joueurs qui s'associent pour obtenir un résultat, d'une valeur ν(E) (dans $\mathbb{R}$), où ν(E) est la fonction « valeur », la fonction caractéristique du jeu. Et on suppose que l'on connaît pour chaque sous-ensemble F de E, le gain ν(F). On cherche la valeur ϕ(i) associée à chaque joueur i et qui représente sa contribution au résultat ν(E) ?

Les valeurs de Shapley sont solutions uniques de 4 axiomes, axiome :

- **de symétrie**
  Si deux joueurs i et j ayant le même impact marginal[6] sur tous les sous-ensembles se substituent dans le jeu, alors ils doivent avoir les mêmes contributions, les mêmes valeurs de Shapley. C'est en cela que la solution est « équitable ».
  Cette propriété s'écrit comme : si $\forall$F $\not\ni$ i, j ; ν(F $\cup \{i\}$) = ν(F $\cup \{j\}$) alors $\phi_i(\nu)$ = $\phi_j(\nu)$
- **d'efficience**
  La somme des valeurs de Shapley de chaque joueur de l'ensemble doit être égale à ce que l'ensemble des joueurs peut obtenir : $\sum_{i=1}^{n} \phi_i(\nu) = \nu(E)$
- **dit du « joueur nul »**
  Un joueur qui ne génère jamais aucun gain ni aucune perte pour les sous-ensembles dont il est membre ne reçoit ni ne paie rien.
  Cela s'écrit comme : si $\forall$F $\not\ni$ i ; ν(F $\cup \{i\}$) = ν(F) alors $\phi_i(\nu)$ = 0
- **d'additivité**
  Si un joueur i participe à deux jeux identiques (c'est-à-dire avec les mêmes joueurs) dont les fonctions caractéristiques sont respectivement ν et ω, et qu'on considère ces deux jeux comme un seul alors l'axiome dit : $\phi_i(\nu + \omega) = \phi_i(\nu) + \phi_i(\omega)$

Ainsi, pour obtenir la valeur de Shapley d'un individu i, il faut calculer pour tous sous-ensembles F possible pour i, la contribution marginale de i au sein de ce sous-ensemble F pour ensuite faire la moyenne des contributions marginales de i sur tous ces sous-ensembles F.
La valeur de Shapley du joueur i est donnée par la formule suivante :

$$Sh(i) = \phi_i(\nu) = \frac{1}{n!} \sum_{F \subset E} (n-f)!(f-1)! \left[\nu(F) - \nu(F \backslash \{i\})\right]$$

5. *La valeur de Shapley – Comment individualiser le résultat d'un groupe* (Alexis Eidelman, janvier 2012)
6. C'est la valeur apportée au groupe si le joueur arrive le dernier

Où $[\nu(F) - \nu(F \backslash \{i\})]$ correspond à la contribution marginale de l'individu *i* au sous-ensemble F : c'est-à-dire la différence entre la valeur de F contenant l'individu *i* et la valeur de F \{*i*}, incluant tous les membres de F hormis l'individu *i*. Et *f*, le nombre d'éléments d'un sous-ensemble contenant *i*, f varie donc de 1 à *n*.

On peut donc résumer la valeur de Shapley comme l'apport moyen d'un individu aux autres. C'est cette capacité de la valeur de Shapley, celle de connaître la contribution de chaque individu à un résultat qui nous intéresse, car elle peut être adaptée aux modèles prédictifs. Dans le cas de modèle de prédictions, elle permet d'expliquer la prédiction localement, c'est-à-dire que pour un individu donné on peut voir la contribution de chaque variable à sa prédiction. On parle alors d'interprétation locale.

5 L'analyse en composantes principales (ACP)

L'Analyse en Composantes Principales, ACP, est une méthode de statistique descriptive multidimensionnelle qui vise à faciliter l'exploration statistique de données quantitatives complexes. Elle permet de synthétiser, résumer, et hiérarchiser les données en réduisant le nombre de dimensions d'un nuage de points tout en conservant un maximum d'information. Cette méthode d'analyse factorielle de données peut être considérée comme une méthode de projection d'un nuage de points de dimension p (si p variables) vers k dimensions ($k < p$), de façon à conserver un maximum d'informations sur un minimum de dimensions.
L'étude des variables permet d'explorer les liaisons entre les variables : quelles sont celles qui se ressemblent, lesquelles sont très différentes … A partir de l'ensemble des variables on souhaite avoir une vision synthétique sans avoir besoin de passer par l'ensemble des couples de variables. Concernant l'étude des individus, elle tend à mettre en évidence des groupes homogènes du point de vue des variables, c'est-à-dire à mettre en évidence les individus qui se ressemblent.

Si les variables étaient qualitatives nous pourrions utiliser une extension de l'ACP et effectuer une AFC (Analyse Factorielle des Correspondances), ou une ACM (Analyse des Correspondances Multiples) dans le cas de variables multiples.

Soit un tableau de données brutes noté X, de n individus et p variables.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

L'étude des individus revient à considérer un nuage de points qui évolue dans un environnement à dimension élevée, à p dimensions si p variables.
Or, la représentation devient difficile quand le nombre de dimension vaut 3, et elle est impossible au-delà.

Comment visualiser le nuage des individus dans un espace très grand ? Avant d'essayer de visualiser, il nous faut d'abord définir la ressemblance entre deux individus.
Deux individus se ressembleront s'ils prennent des valeurs proches dans l'ensemble p variables, on parle de distance (distance euclidienne au carré) entre deux individus i et i'.

$$d^2(i, i') = \sum_{p=1}^{p} (x_{i,p} - x_{i',p})^2$$

Ainsi, étudier le nuage des individus revient à analyser la forme du nuage de points, et on cherchera à visualiser ce nuage dans un espace à deux dimensions.

Au préalable, deux pré-traitements sont recommandés : centrer et réduire les données. La réduction est une étape indispensable en présence de variables exprimées dans des unités différentes. Elle permet d'éviter les effets d'échelle, et d'accorder plus d'importance aux variables avec une grande variance.
Ainsi, $x_{i,j}$ s'exprime

$$x_{i,j} = \frac{x_{i,j} - \bar{x}_j}{\sigma_j} \qquad où, \sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)^2 \; et \; \bar{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$$

L'ACP va permettre de visualiser un tableau à p dimensions, en cherchant une image simplifiée, la plus fidèle possible (i.e. chercher un sous-espace F_k de $\mathbb{R}^p$ de dimension k, où k petit) qui résume au mieux les données. En effet, les distances entre individus sont calculées dans un espace à p dimensions, avant d'être projetées. Il faut donc qu'elles soient le moins déformées possibles, que la forme générale du nuage soit fidèlement restituée et ainsi la variabilité des individus conservée.
Autrement dit, on cherche à définir k nouvelles variables, combinaisons linéaires des p variables initiales et qui conservent le maximum d'informations par rapport aux variables initiales. Les axes sont appelés composantes principales.

Pour qualifier la qualité d'une image, on utilise le terme de dispersion; plus le nuage est dispersé, plus la variabilité est importante.
Cette variabilité est sur plusieurs dimensions, on parle alors d'inertie.
Le sous-espace F_k de $\mathbb{R}^p$ de dimension k (avec k petit) optimal, sera celui sur lequel le nuage projeté aura une inertie maximale (i.e. que le nuage sera très dispersé). On définit l'inertie de la projection du nuage de dimension k comme,

$$I(F_k) = \sum_{i=1}^{i} p_i ||x_i - \hat{x}_i||^2 \qquad où, \; \hat{x}_i \; est \; la \; projection \; de \; x_i \; sur \; F_k$$

Ce sous-espace est également celui qui contient l'axe déformant le moins le nuage, autrement dit, celui maximisant la variabilité; c'est-à-dire, celui résumant au mieux l'ensemble des individus.
En cherchant le second axe orthogonal au premier et qui synthétise le reste de l'information non synthétisée par le premier, on obtient le plan optimal.
La recherche des autres axes repose sur le même principe : ils doivent être orthogonaux entre eux et maximiser l'inertie.

Pour interpréter les graphes d'individus, les variables sont nécessaires. En calculant la corrélation entre les variables et chaque axe, on obtient ainsi le cercle des corrélations. Selon la valeur et le signe de chaque corrélation obtenue, on peut caractériser chaque axe.

Pour vérifier si un individu est bien représenté sur le nuage, et par quel axe il est le mieux représenté, différents instruments de mesure existent :

- **Le pourcentage d'inertie**, qui correspond au pourcentage d'information expliquée par chaque axe. Le pourcentage d'un plan peut également être connu en sommant le pourcentage d'inertie des deux axes associés, car les axes sont orthogonaux.

- **La qualité de représentation** d'un individu j sur l'axe s, correspond au cos^2 de l'angle entre cet individu j et son projeté. Plus le cos^2 est proche de 1, plus l'angle est proche de 0, et donc l'individu est extrêmement bien projeté.
  Si on veut connaître la qualité de représentation à un axe, il suffit de regarder le cos^2 associé à l'individu et l'axe en question. Si on souhaite la connaître pour un plan, il faut sommer les qualités de représentation (car les axes sont orthogonaux).
  Remarque : Seuls les individus correctement projetés peuvent être interprétés, car si deux individus sont mal projetés, alors cela signifie qu'ils sont éloignés du plan de projection, sans que l'on sache s'ils sont proches ou non dans l'espace (même s'ils le semblent sur le plan).

- **La contribution** d'un individu j à la construction d'un axe s (i.e. le pourcentage d'inertie de l'axe s lié à l'individu j) correspond aux coordonnées au carré de l'individu j sur l'axe s divisé par la somme des coordonnées de l'ensemble des individus.

$$Ctr(s) = \frac{F_{js}^2}{\sum_{i=1}^{i} F_{js}^2} \in [0, 1]$$

Ainsi l'ACP permet de visualiser la structure des corrélations entre plusieurs variables. Cependant, la représentation bidimensionnelle peut distordre les projections factorielles, notamment en présence de variables nombreuses et complexes.
Une solution pour corriger ces distorsions, consiste à faire figurer sur les projections factorielles du nuage des individus, la partition obtenue par *clustering*.
En effet, ACP et *clustering* (et partitionnement) sont complémentaires, car l'ACP appréhende l'ensemble des dimensions du nuage : elle prend en compte les individus tels qu'ils sont réellement, et non tels qu'ils sont en projection.

5.1 Méthode de *clustering* et de partitionnement des données

Les méthodes de *clustering* et de partitionnement sont d'autres méthodes d'analyse de données. Elles consistent à regrouper en classe des objets (individus, points…) similaires, tel que chaque objet soit dans une classe, et que ces classes forment une partition (en littérature anglaise, on parle de *cluster*). Ainsi, chaque groupe se veut homogène, mais très différent des groupes restants.
Ces méthodes appartiennent aux algorithmes non supervisés (i.e. les groupes n'existent pas avant d'être créés ; on n'apprend pas à partir des données pour construire les *clusters*).

Deux grandes techniques de *clustering* existent : la classification hiérarchique et le partitionnement. Pour chacune de ces techniques, on abordera une méthode, respectivement la Classification Ascendante Hiérarchique (CAH) et la méthode des *k-means*.
Ces deux méthodes sont complémentaires.

5.1.1 La classification ascendante hiérarchique

Les méthodes de classification hiérarchique fournissent un ensemble de partitions de moins en moins fines, obtenu par regroupement successifs de partitions. *In fine*, ce n'est pas

une partition qu'on obtient, mais une hiérarchie de partition en $n, \ldots, 1$ classes, où l'inertie inter-classe diminue à chaque agrégation.

L'algorithme commence avec n singletons, chaque élément (individu, point...) constituant un groupe, puis on réunit les groupes considérés comme les plus proches, selon un critère de regroupement, et ce jusqu'à l'obtention d'un groupe contenant tous les éléments.
Nous avons donc besoin de définir une distance entre les éléments, ainsi que d'un critère de regroupement également appelé stratégie d'agrégation.

5.1.1.1 Distance entre éléments

La distance entre éléments revient à choisir comment les dissimilitudes entre éléments vont être mesurées. Le choix de cette mesure entre éléments va dépendre des données étudiées ainsi que de l'objectif à atteindre.

Il existe plusieurs distances, parmi lesquelles on peut citer :

- **La distance euclidienne**
  Il s'agit d'une distance géométrique dans un espace multidimensionnel.
  $$dist(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$
- **La distance euclidienne au carré**
  Elle permet de « surpondérer » les objets atypiques (i.e. éloignés) en élevant la distance euclidienne au carré.
  $$dist(x, y) = \sum_i (x_i - y_i)^2$$
- **La distance de Manhattan**
  Cette distance correspond à la somme des différences entre les dimensions.
  $$dist(x, y) = \sum_i |x_i - y_i|$$

5.1.1.2 Critère de regroupement, ou stratégie d'agrégation

Après avoir défini la distance entre éléments, il faut définir la distance entre classes. Plusieurs stratégies existent, telles que :

- **La stratégie du saut minimum ou *single linkage***
  On regroupe les deux classes présentant la plus petite distance entre éléments des deux classes.
  $$\Delta(A, B) = min_{i \in A, j \in B}\ d(i, j)$$
- **La stratégie du saut maximum ou du diamètre ou *complete linkage***
  On regroupe les deux classes présentant la plus grande distance entre éléments des deux classes.
  $$\Delta(A, B) = max_{i \in A, j \in B}\ d(i, j)$$

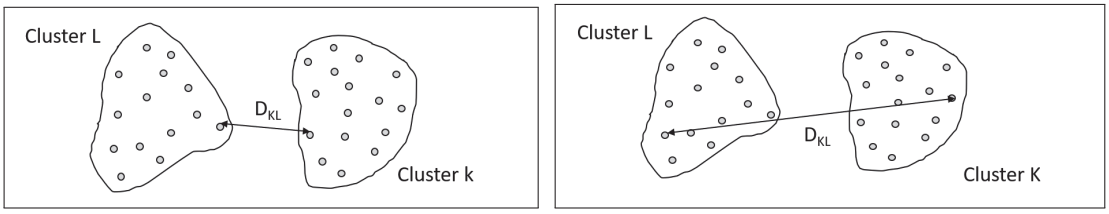


Figure I.1 – *Stratégie du saut minimum (à gauche) et Stratégie du saut maximum (à droite)*

- **La méthode de *Ward***
  C'est la méthode la plus connue et la plus utilisée dans la pratique, car elle donne de meilleurs résultats. Elle correspond à une optimisation pas à pas. On agrège à chaque itération les classes dont l'agrégation fait perdre le moins d'inertie inter-classe.
  Avec cette méthode, les indices d'agrégation sont calculés de la manière suivante : « *Si une classe C est obtenue en regroupant les classes A et B, sa distance à la classe D est donnée par la distance entre les barycentres de la classe A et D* ».

  $$\Delta(C_A, C_D) = \sqrt{\frac{p_A p_D}{p_A + p_D} ||g_{C_A} - g_{C_D}||^2} \qquad où,\ p_i = poids\ du\ groupe\ C_i$$

  A chaque étape de la CAH, on rassemble les deux classes présentant la plus petite distance.

La dernière étape d'une CAH, consiste à choisir la partition finale, celle qui semble être la meilleure. On dit qu'une classification est bonne si la variabilité inter-classe est grande, et la variabilité intra-classe petite.
Pour prendre en compte ces deux critères, on utilise l'inertie totale d'un groupe.

Définition 5.1 (Inertie totale)**.** *Soient $X_1, ..., X_n$ les individus et g le centre de gravité.*
Soit une classification en k groupes $G_1, ..., G_k$ d'effectifs $n_1, ..., n_k$ et de centres de gravités $g_1, ..., g_k$. On définit l'inertie totale de la manière suivante

$$\begin{aligned} Inertie\ totale &= Inertie\ interclasse + Inertie\ intraclasse \\ &= \frac{1}{n}\sum_{i=1}^{k} d^2(g_i, g) + \frac{1}{n}\sum_{i=1}^{k}\sum_{e \in G_i} d^2(e, g_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} d^2(X_i, g) \end{aligned}$$

Généralement, la meilleure partition sera celle qui précédera une distance inter-classe brutalement plus faible.
Le graphe du *R square semi partiel* et le dendrogramme sont les deux meilleures représentations pour visualiser ce saut.
La Figure I.2, est un exemple de dendrogramme (représentation graphique sous forme d'arbre binaire) associé à un nuage de points.
Cette représentation permet de mettre en évidence une hiérarchie entre éléments et groupes d'éléments. Au sommet, se trouve la racine de l'arbre contenant tous les éléments. La hauteur d'une branche est proportionnelle à la distance entre les deux objets regroupés (plus la branche est haute, plus la distance est grande et la distance *inter* l'est également).

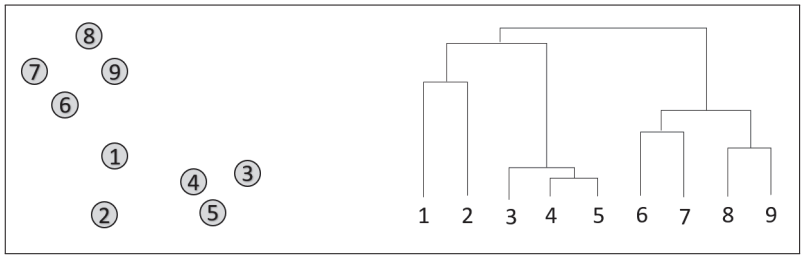


Figure I.2 – *Dendrogramme (à droite) obtenu à partir des données 2D (à gauche)*

5.1.2 La méthode des *k-means*

La classification *k-means* a été introduite en 1967 par MacQueen, elle consiste à regrouper les objets (individus, points…) en k groupes distincts, k étant fini.
La méthode des *k-means* repose sur la minimisation de la somme des distances euclidiennes au carré entre chaque élément et le centroïde (point central) de la classe.
Dans la méthode des *k-means*, les classes ne se chevauchent pas et les éléments ne peuvent appartenir qu'à une seule classe.

5.1.2.1 La distance euclidienne

La distance euclidienne est à la base de la méthode des *k-means*, puisqu'il s'agit d'attribuer chaque élément à une classe, de façon à ce que la somme des distances euclidiennes au carré entre chaque point et le centroïde de sa classe, soit la plus faible. On parle de minimisation inter-classe.

On rappelle que la distance euclidienne dans un espace à n dimensions $(u, v, \ldots, z)$ entre deux éléments i et i', est définie par $dist(i, i') = \sqrt{(u_i - u_{i'})^2 + (v_i - v_{i'})^2 + \ldots + (z_i - z_{i'})^2}$

Ainsi, deux éléments identiques auront une distance nulle.

De même qu'en ACP ou CAH, il est préférable de centrer et réduire les données pour s'émanciper des unités des variables. La présence d'unités différentes, augmente le risque de sur-pondération.

5.1.2.2 L'algorithme

L'algorithme du *k-means* tend à minimiser la variance intra-classe. C'est un algorithme itératif qui cherche à minimiser la somme des distances entre chaque individu et le centroïde de la classe. L'algorithme peut être résumé en 5 étapes :

1. Attribuer une classe à chaque élément (individu, point…) de façon aléatoire.
2. Calculer le centroïde de chaque classe.
3. Pour chaque élément, calculer sa distance euclidienne avec les centroïdes de chaque classe.
4. Calculer la somme de la variabilité intra-classe
5. Réitérer les étapes 2 à 5, jusqu'à atteindre la stabilisation de la somme de la variabilité intra-classe (i.e. qu'il n'y a plus de changement de classe).

L'algorithme converge toujours vers une solution, car à chaque itération, l'inertie inter-classe diminue. Cependant, à l'inverse des CAH qui aboutissent à un arbre dit optimal, la solution de l'algorithme *k-means* ne sera pas toujours la même si cet algorithme est réitéré plusieurs fois car il dépend du choix aléatoire des premiers centres.
Pour contourner cette dépendance à l'attribution initiale aléatoire des classes, il est nécessaire de lancer plusieurs fois l'algorithme et de regarder la stabilité.

Sous *R*, la fonction *kmeans()* permet de réaliser automatiquement ces attributions aléatoires multiples et de sélectionner les meilleurs résultats.

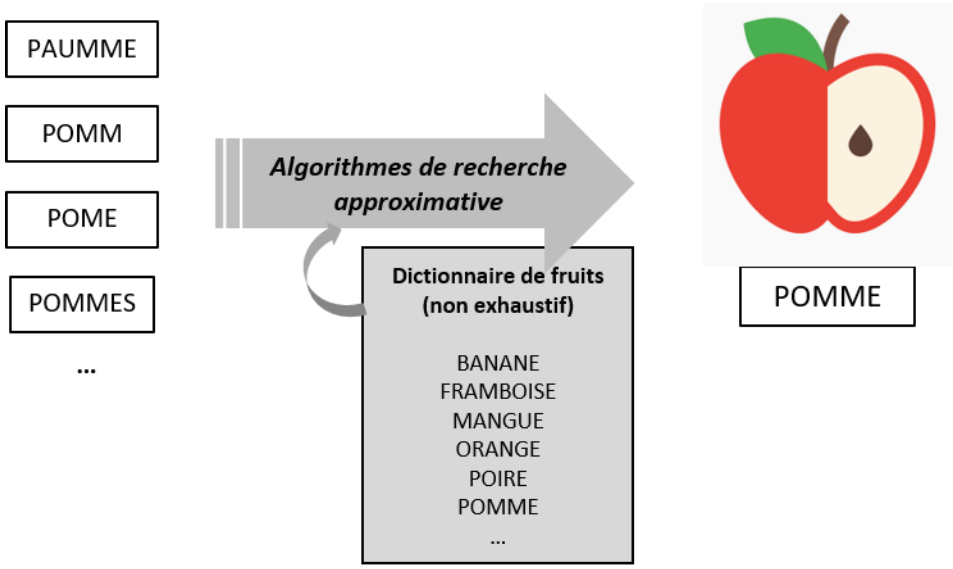
Cette méthode de partitionnement présente l'avantage d'être rapide, facile à comprendre et de pouvoir traiter des données volumineuses.

6 Correction orthographique et recherche approximative

Dans l'objectif d'améliorer la qualité de nos données, on cherche à corriger l'une de nos variables qui est en saisie libre et à la main de plusieurs utilisateurs, générant ainsi une multitude d'orthographes.
Le nombre de possibilités d'orthographe nous empêche de réaliser une correction manuelle, nous obligeant à nous tourner vers des algorithmes, et plus précisément vers de la recherche approximative (RA) ou *fuzzy matching*.

6.1 Recherche approximative, ou *Fuzzy matching*

La recherche approximative ou *fuzzy matching* est une technique permettant de rechercher des correspondances entre chaînes de caractères, en se basant sur un motif approximatif plutôt qu'exact. Plus une chaîne de caractères présente des similitudes avec une autre, plus la correspondance sera avérée.
Cette technique suppose la création d'un dictionnaire le plus exhaustif possible : il doit référencer toutes les chaînes de caractères orthographiées correctement, auxquelles seront comparées les chaînes de caractères à corriger.



6.2 Représentation d'un mot

Un mot, ou plus largement un groupe de mots peut être représenté de deux façons différentes :

- comme une suite de caractères : on parlera de représentation basée sur les caractères alphanumériques ;
- ou comme un son : on parlera de représentation basée sur les phonèmes.

Dès lors, il est important de prendre en considération ces deux représentations qui se veulent complémentaires, afin d'avoir une correction orthographique la plus juste.

C'est pour cela qu'on a considéré notre variable à corriger comme une chaîne de caractères dans un premier temps, avant de considérer son contenu comme un son. Ainsi deux approches de RA ont été appliquées.
Avant de présenter les deux méthodes que nous avons utilisées d'un point de vue théorique, il faut savoir que des traitements sont à réaliser en amont sur la variable à corriger, car les méthodes qu'on va utiliser sont sensibles à la casse et aux caractères spéciaux. Le premier traitement concerne donc la casse, il nous faut l'uniformiser en mettant tout en majuscule ou

en minuscule. Puis, il faut supprimer tous les accents, les symboles, les parenthèses, les tirets… et remplacer les espaces multiples par des espaces simples.

6.3 Distance de Levenshtein

La distance de Levenshtein, appelée également distance d'édition ou distance de similarité est une méthode inventée en 1965 par le mathématicien russe Vladimir Levenshtein.
Elle permet de comparer deux mots ou chaînes de caractères, en calculant leur degré de ressemblance. On peut définir la distance entre deux mots m_1 et m_2 comme étant le nombre minimal d'opération d'édition pour transformer m_1 en m_2.

Les opérations sont de trois types :

- la substitution d'une lettre de m_1 par une lettre de m_2 ;
- la suppression d'une lettre de m_1 ;
- l'insertion d'une lettre de m_2.

6.3.1 Calcul de la distance

Définition 6.1 (Distance de Levenshtein)**.** *A partir de ces deux mots m_1 et $m_2 \in A^*$, l'ensemble des mots finis sur l'alphabet, de longueur respective ℓ_1 et ℓ_2, on définit la table T à $\ell_1 + 1$ lignes et $\ell_2 + 1$ colonnes par*

$$T[i,j] = lev(m_1[0,\ldots,i], m_2[0,\ldots,j])$$

Pour $i = -1, 0, \ldots, \ell_1 - 1$ et $j = -1, 0, \ldots, \ell_2 - 1$

Définition 6.2. *Pour calculer $T[i,j]$, on utilise la formule de récurrence suivante :*

Pour $i = -1, 0, \ldots, \ell_1 - 1$ et $j = -1, 0, \ldots, \ell_2 - 1$, on a :

$$\begin{aligned} T[-1,-1] &= 0 \\ T[i,-1] &= T[i-1,-1] + 1 \\ T[-1,j] &= T[-1,j-1] + 1 \end{aligned}$$

$$T[i,j] = min \begin{cases} T[i-1,j-1] + \textit{Sub}(m_1[i], m_2[j]) \\ T[i-1,j] + 1 \\ T[i,j-1] + 1 \end{cases}$$

Où

$$\textit{Sub}(a,b) = \begin{cases} 0 & \textit{si } a = b \\ 1 & \textit{sinon.} \end{cases}$$

La valeur à la position [i, j] dans la matrice *T*, avec $i, j \geqslant 0$, ne dépend ainsi que des valeurs aux positions $[i-1, j-1]$, $[i-1, j]$ et $[i, j-1]$.

De cette matrice *T*, on peut extraire deux informations : la distance minimale entre m_1 en m_2 et les suites d'opérations (insertion, substitution et suppression) permettant de passer d'un mot à l'autre.
La valeur de la distance minimale entre m_1 et m_2, se trouve dans la cellule en bas à droite de la matrice *T*. A partir de cette même cellule (en bas à droite), il est possible d'expliciter les suites d'opérations, en remontant de cellule en cellule, en prenant à chaque fois la ou les cellules à l'origine de la valeur minimale. Plusieurs cellules pouvant être à l'origine de

cette valeur minimum, plusieurs chemins peuvent être déduits. Mais ils seront toujours de longueur minimale.
Plus le nombre d'opérations est important, plus la distance est élevée, et moins les chaînes sont similaires. Deux chaînes de caractères identiques auront une distance de Levenshtein de 0, car elles ne nécessiteront aucune transformation.

Pour illustrer le principe de calcul de la distance de Levenshtein, l'exemple ci-dessous va montrer comment calculer la matrice *T*, comment déterminer la distance, et quelles sont les opérations nécessaires pour transformer le mot *CHIEN* en *NICHE*.

La première étape consiste donc à calculer la matrice *T* en respectant les règles définies dans la **Définition 6.2**.

		N	I	C	H	E
	0	1	2	3	4	5
C	1					
H	2					
I	3					
E	4					
N	5					

		N	I	C	H	E
	0	1	2	3	4	5
C	1	1	2	2	3	4
H	2	2	2	3	2	3
I	3	3	2	3	3	3
E	4	4	3	3	4	3
N	5	4	4	4	4	4

Figure I.3 – Matrice T associée à la distance de Levenshtein

Pour trouver la valeur contenue dans la cellule $T[1,1]$, on procède de la manière suivante :
Comme C est différent de N, le coût est de 1 ($sub(C,N) = 1$). En lui ajoutant l'élément diagonal *0* (en haut à gauche), le premier élément vaut 1.
Si on considère l'élément à gauche de $T[1,1]$ et qu'on ajoute 1, on obtient 2 pour le second élément. De même pour l'élément du dessus, à savoir *1*, qui nous donne une valeur de 2 pour le troisième élément.
La plus petite valeur entre les trois éléments 1, 2, 2 est 1. On inscrit donc 1 dans la case $T[1,1]$.

Mathématiquement, cela donne :

$$T[1,1] = min \begin{cases} T[0,0] + Sub(C,N) \\ T[0,1] + 1 \\ T[1,0] + 1 \end{cases} = min \begin{cases} 0+1 \\ 1+1 \\ 1+1 \end{cases} = 1$$

Pour chaque cellule, on procède de la même manière, et on obtient la matrice finale *T* (matrice de droite sur la Figure I.3).

Enfin, à partir de cette même matrice, on peut connaître la distance minimale entre les mots *CHIEN* et *NICHE*, à savoir 4 (valeur de la cellule en bas à droite) mais également savoir quelles sont les suites d'opérations qui ont permis de passer du mot *CHIEN* au mot *NICHE*.

Début de lecture →

		N	I	C	H	E
	0	1	2	3	4	5
C	1	1	2	2	3	4
H	2	2	2	3	2	3
I	3	3	2	3	3	3
E	4	4	3	3	4	3
N	5	4	4	4	4	4

On rappelle les 3 opérations possibles :

- la substitution d'une lettre de m_1 par une lettre de m_2 ;
- la suppression d'une lettre de m_1 ;
- l'insertion d'une lettre de m_2.

Et leur sens de lecture,

→ Insertion
↘ Substitution
↓ Suppression

On part du mot CHIEN

		N	I	C	H	E
	0	1	2	3	4	5
C	1	1	2	2	3	4
H	2	2	2	3	2	3
I	3	3	2	3	3	3
E	4	4	3	3	4	3
N	5	4	4	4	4	4

Insertion de la lettre N
→ NCHIEN

		N	I	C	H	E
	0	1	2	3	4	5
C	1	1	2	2	3	4
H	2	2	2	3	2	3
I	3	3	2	3	3	3
E	4	4	3	3	4	3
N	5	4	4	4	4	4

Insertion de la lettre I
→ NICHIEN

		N	I	C	H	E
	0	1	2	3	4	5
C	1	1	2	2	3	4
H	2	2	2	3	2	3
I	3	3	2	3	3	3
E	4	4	3	3	4	3
N	5	4	4	4	4	4

Substitution de la lettre C
→ NICHIEN

		N	I	C	H	E
	0	1	2	3	4	5
C	1	1	2	2	3	4
H	2	2	2	3	2	3
I	3	3	2	3	3	3
E	4	4	3	3	4	3
N	5	4	4	4	4	4

Substitution de la lettre H
→ NICHIEN

		N	I	C	H	E
	0	1	2	3	4	5
C	1	1	2	2	3	4
H	2	2	2	3	2	3
I	3	3	2	3	3	3
E	4	4	3	3	4	3
N	5	4	4	4	4	4

Suppression de la lettre I
→ NICHEN

		N	I	C	H	E
	0	1	2	3	4	5
C	1	1	2	2	3	4
H	2	2	2	3	2	3
I	3	3	2	3	3	3
E	4	4	3	3	4	3
N	5	4	4	4	4	4

Substitution de la lettre E
→ NICHEN

		N	I	C	H	E
	0	1	2	3	4	5
C	1	1	2	2	3	4
H	2	2	2	3	2	3
I	3	3	2	3	3	3
E	4	4	3	3	4	3
N	5	4	4	4	4	4

Suppression de la lettre N
→ NICHE

		N	I	C	H	E
	0	1	2	3	4	5
C	1	1	2	2	3	4
H	2	2	2	3	2	3
I	3	3	2	3	3	3
E	4	4	3	3	4	3
N	5	4	4	4	4	4

Figure I.4 – Illustration du calcul de la distance de Levenshtein entre les mots *Chien* et *Niche*

Sur la Figure I.4, on peut voir que pour transformer le mot *CHIEN* en *NICHE*, 4 opérations sont nécessaires : 2 insertions et 2 suppressions (l'opération substitution n'a pas de coût, car il s'agit du même caractère). Et 4 correspond également à la distance entre ces deux chaînes de caractères.

6.3.2 Degrés de similitude

Pour faciliter l'exploitation des résultats de la distance de Levenshtein, on peut calculer le rapport de similitude basé sur la distance de Levenshtein :

$$\frac{(|a| + |b|) - lev_{a,b}(i, j)}{|a| + |b|}$$

où |a| et |b| correspondent aux longueurs des chaînes de caractères a et b.

Ainsi, il est plus simple de quantifier la similitude entre deux chaînes via un pourcentage de similitude, qu'un nombre minimal d'opération qui ne tient pas compte de la longueur des chaînes : une distance de 4 n'aura pas la même signification si on compare des chaînes de caractères de longueur inférieure à 10 versus supérieure à 20.

Si on reprend notre exemple, le rapport de similitude vaut $\frac{(5+5)\text{ - }4}{(5+5)} = 60\%$.

6.3.3 Complexification de l'algorithme

A noter, qu'il est possible de complexifier l'algorithme original de Levenshtein, en affectant des poids différents aux différentes opérations par exemple. Ainsi, on peut pénaliser en augmentant ou en diminuant le coût de telle ou telle opération, comme diminuer le coût lorsqu'il s'agit d'une substitution entre deux caractères similaires (m et n, b et p...)
De plus, la distance de Levenshtein n'est pas la seule approche pour comparer des chaînes de caractères entre elles : il en existe d'autres, telles que la distance de Jaro-Winckler, qui va s'appuyer sur le nombre de caractères communs et non le nombre de modifications à faire ; la distance de Hamming, qui comptabilise le nombre de caractères à la même position dans les deux chaînes...

La distance de Levenshtein permet d'aborder un mot ou une séquence comme une suite de caractères. Elle est donc bien adaptée à la mise en évidence des fautes de frappe.
En revanche, elle l'est moins lorsqu'il s'agit d'erreurs d'orthographe, d'écriture en abrégé ou en présence d'homonymes. Pour prendre en considération ces « erreurs » de saisies, il nous faut considérer les mots ou séquences comme une suite de sons, c'est-à-dire de façon consonantique.

6.4 L'algorithme *Soundex*

Le terme *Soundex* est apparu au début du XXe siècle, avec un premier algorithme inventé par Margaret O'Dell et Robert C. Russell. Les algorithmes de *Soundex* font partie des algorithmes phonétiques.
Tous les algorithmes de *Soundex* reposent sur un même principe, à savoir la décomposition d'un mot en sons plutôt qu'en une suite de caractères. Ces algorithmes utilisent un ensemble de règles pour représenter une chaîne de caractères à l'aide d'un code court : qui prend la forme d'une lettre suivie de 3 chiffres.
Ce code contient alors les informations « essentielles » sur la sonorité de la chaîne si elle est lue à voix haute. Ainsi, en comparant ces codes, il est possible d'avoir une correspondance entre deux chaînes qui s'écriraient différemment mais se prononceraient de la même façon.
Initialement, cet algorithme a été élaboré pour la langue anglaise, avant d'être adapté par la suite à d'autres langues, dont le français. En effet, les règles varieront en fonction de la langue d'étude des mots, car on n'utilise pas le même lexique, les sons peuvent être différents... mais le procédé reste le même, et comporte 6 étapes :

1. Retranscription du mot en majuscule, après suppression et remplacement des accents, espaces multiples…
2. Conservation de la première lettre du mot.
3. Elimination des voyelles, des lettres muettes (h et w, en version française).
4. Transcodification des lettres restantes à l'aide des règles suivantes (en version française).

Lettres	**B P**	**C K Q**	**D T**	**L**	**M N**	**R**	**G J**	**S X Z**	**F V**
Code	1	2	3	4	5	6	7	8	9

5. Suppression de toutes les paires consécutives de chiffres dupliqués.
6. Conservation des quatre premiers caractères du *Soundex*, à compléter par des 0 si besoin (1 lettre + 3 chiffres).

Cependant, cette représentation présente deux limites. La première concerne la représentation en elle-même : en effet, la représentation *Soundex* ne se fait que sur 4 caractères. Cela réduit donc drastiquement la représentation des mots par rapport au nombre de caractères originellement distincts.

Pour illustrer ce propos, on peut utiliser les mots *Ment*, *Menthe* et *Mante* qui ont le même code, *M530*. L'espace de représentation des mots est réduit de 3 à 1.

Deuxièmement, certaines correspondances phonétiques ne se font pas, notamment quand le caractère est en début de mot comme les sons «F» et «Ph».

Enfin, le *Soundex* n'est pas le seul algorithme de *Soundex* : d'autres algorithmes existent comme le *Soundex 2* ou le *Phonex*. Le premier repose sur le même principe, avec 4 caractères ; cependant, cette version conserve les lettres, on ne transcode pas en chiffres. Le second est

encore plus performant, car il reconnaît certains sons comme «ein», «ai», «on» … Mais cette performance a un coût : le temps de calcul est deux fois plus long que le *Soundex*.

Dès lors, l'algorithme *Soundex* sera pratique pour corriger les erreurs phonétiques, ainsi que les saisies en abrégé, mais moins efficace pour traiter les erreurs de frappes. C'est pourquoi les deux approches (alphanumériques et phonétiques) permettent de mieux appréhender une chaîne de caractères, et se veulent complémentaires dans notre démarche de correction d'orthographe, même si ces méthodes ne sont pas uniques.
Dans notre cas, la distance de Levenshtein est suffisante pour comparer sur la base de caractères alphanumériques, car nos chaînes de caractères sont relativement courtes. Si ça n'avait pas été le cas, si nos chaînes de caractères avaient été plus longues, telles que des phrases, voire des paragraphes, alors il y aurait eu le risque d'avoir des chaînes de caractères très différentes d'un point de vue orthographique et pourtant ayant le même sens.

Exemple : Il neige, la température est très basse, vs Il fait très froid, il neige.

Il aurait donc fallu considérer d'autres algorithmes plus performants, tel que celui du package *fuzzywuzzy* qui calcule la distance et le rapport de similitude de Levenshtein comme précédemment, mais présente des fonctionnalités supplémentaires permettant de comparer et chercher la correspondance dans des sous-chaînes. Ou alors des algorithmes de prolongement de mots (le *words embeddings*), permettant de contextualiser une chaîne de caractères en la projetant dans un espace vectoriel.
Ces aspects ne seront pas plus approfondis, car non traités dans ce mémoire, mais ils illustrent la multitude d'approches existantes pour aborder un problème de correction orthographique, et/ou de contextualisation.

7 *Market Basket Analysis* (MBA)

Le *Market Basket Analysis* (MBA), ou analyse du « panier de consommation » est une technique cherchant à prédire les comportements d'achat.
C'est un ensemble de calculs statistiques d'affinités mettant en évidence des modèles de fréquences dans les données.
En d'autres termes, le MBA permet d'identifier les combinaisons (d'achat) de produits se produisant souvent ensemble dans les commandes. En général, cette technique est utilisée en marketing, et plus précisément sur les transactions, d'où le terme de panier. Il repose sur des règles d'associations (*Association Rule Mining*).

7.1 Les règles d'association

7.1.1 Qu'est-ce qu'une règle d'association ?

On cherche à savoir quels articles (*items*) sont fréquemment achetés ensemble par les clients. Ce qui nous donne un ensemble de règles appelées règles d'association. On peut résumer cela en « si ceci, alors cela ».
Le MBA est donc utilisé en marketing et peut répondre à de nombreuses stratégies comme :

- analyser le comportement des clients ;
- identifier les articles les plus sollicités ;
- cibler les courriers personnalisés ;
- changer l'agencement d'un magasin ;
- ...

Une règle entre un produit A et un produit B, s'écrira de la manière suivante :

$$A \Rightarrow B = [Support, Confiance]$$

Et pour $A \Rightarrow B = [Support = 20\%, Confiance = 60\%]$, elle se lira comme suit:

20% des achats montrent que le produit B et le produit A sont achetés ensemble

60% des clients qui achètent le produit B l'ont acheté avec le produit A

7.1.2 L'exploration des règles d'association

Le Support et la Confiance, sont deux mesures permettant de mesurer l'intérêt de la règle (sa force).

$$Support\ (A \Rightarrow B) = \frac{fréquence(A, B)}{N} \qquad où\ N : nombre\ de\ transaction$$

$$Confiance\ (A \Rightarrow B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{fréquence(A, B)}{fréquence(A)}$$

On nommera *minSupp* et *minConf* les seuils minimums du Support et de la Confiance, et seront à fixer.
On parlera de règle d'association forte quand elle satisfait *minSupp* et *minConf*.

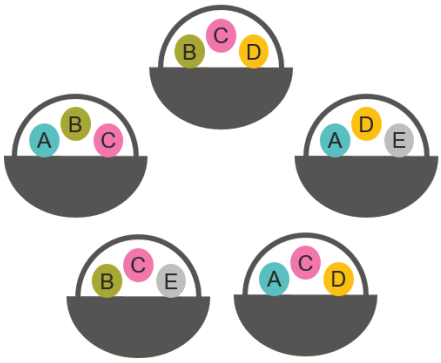
Enfin, pour mesurer la corrélation entre A et B dans la règle $A \Rightarrow B$, on calcule le *Lift*. C'est-à-dire, on montre comment un ensemble d'éléments *A* affecte l'ensemble d'éléments *B*.

$$Lift\ (A \Rightarrow b) = \frac{Support}{Supp(A) * Supp(B)}$$

Ce dernier s'interprète de la façon suivante :

- si = 1, alors A et B sont indépendants et aucune règle ne peut être exploitée ;
- si >1, alors A et B dépendent l'un de l'autre et le degré est donné par la valeur du lift ;
- si < 1, la présence de A aura un effet négatif sur B.

Pour illustrer ces notions de Support, de Confiance et de *Lift*, nous l'illustrons par un exemple. Soit 5 transactions, composées d'un ensemble de 5 *items*.



Règles potentielles	Support	Confiance	Lift
{A} => {D}	1/5	2/3	10/9
{C} => {B}	3/5	3/4	**5/4=1,25**
...	...	...	...
{B,C} => {D}	1/5	1/3	5/9

L'achat des produits B et C ensemble, est 1,25 fois plus susceptible de se réaliser que le hasard. Ces deux produits ont une relation positive entre ces 2 produits

7.1.3 Identifier les règles d'association

On cherche à identifier des règles d'association qui répondent à deux conditions : un Support supérieur à un seuil fixé et de même pour la Confiance, autrement dit :

$$Support \geqslant minSupp$$

et

$$Confiance \geqslant minConf$$

Cependant, le volume de données généralement traité est très important. Il est donc impossible de chercher toutes les règles dans un ensemble de transactions T, et de sélectionner celles répondant aux conditions de seuils fixés. Cela reviendrait à calculer le Support et la Confiance pour chaque règle, et à ne conserver que celles répondant aux seuils. Ce serait extrêmement coûteux en termes de calculs. Pour un ensemble de d items, le nombre de règles d'association possibles est de :

$$R = 3^d - 2^{d+1} + 1$$

Dans notre exemple, d = 5 soit R = 180 règles possibles. Et si on augmente d de 1, soit d = 6, on passe à 602 règles possibles.

Or, les bases de données sont généralement de taille importante, ainsi la recherche de règles d'association et plus particulièrement la recherche des ensembles d'*items* est très coûteux numériquement. De plus, toutes les associations ne sont pas forcément intéressantes et peuvent être dues au hasard. C'est pourquoi la plupart des algorithmes de recherche de règles d'associations se décomposent en deux étapes :

1. Itérativement, trouver les *itemsets* (= ensemble d'*items*) fréquents dont la cardinalité varie de 1 à k (*k-itemset*)
2. Utiliser les *itemsets* fréquents pour générer les règles d'association.

7.1.4 La génération des ensembles d'*items* fréquents

Souvent, cette première étape est très coûteuse numériquement et constitue une limite. En termes de calcul, l'algorithme analyse la base de données plusieurs fois, ce qui réduit les performances. En effet, une fois les ensembles d'*items* énumérés, il faut calculer le Support pour chacun d'eux. Pour ce faire, il faut rapprocher chaque ensemble d'*items* à toutes les transactions observées, et implémenter celui-ci lorsqu'il est observé dans une transaction.
Cette approche peut donc être très coûteuse, car la complexité, autrement dit le nombre de comparaisons, est de O(NMω), où N est le nombre de transactions, M = 2^k − 1 le nombre d'ensemble d'*items* et ω la plus grande taille de transaction.
Dans notre exemple, on aurait $O(5 * 2^5 * 4)$ comparaisons à faire, alors qu'on est dans une base de données de 5 transactions avec 5 *items* possibles, ce qui est très loin des bases de données avec des millions de lignes de transactions, sans parler du nombre d'*items* possibles.

Il est donc nécessaire de diminuer le coût de la recherche d'ensemble d'*items*. Pour cela, il existe deux stratégies :

- réduire le nombre d'ensemble d'*items* fréquents (M) sans avoir besoin de calculer le Support. Pour cela, on utilise le principe *Apriori*, qu'on développera dans la partie suivante ;
- réduire le nombre de comparaisons (NM). Au lieu de comparer chaque ensemble d'*items* à toutes les transactions, on peut réduire le nombre de comparaisons en utilisant des structures de données plus complexes, comme utiliser une fonction de hachage. (Notion non abordée dans ce mémoire.)

7.2 Qu'est-ce que le principe *Apriori* ?

Le principe *Apriori* repose sur l'idée que si un ensemble est non fréquent, alors tous ses sur-ensembles ne sont pas fréquents :

- si A n'est pas fréquent, alors A,B ne peut pas l'être ;
- si A,B est fréquent, alors A et B le sont.

La réciproque est également vraie : si un ensemble est fréquent, alors tous ses sous-ensembles le sont aussi. Dès lors, si on sait que le sous-ensemble n'est pas fréquent, alors on peut supprimer tous les ensembles le contenant. Cette stratégie s'appelle l'élagage basé sur le Support, et elle repose sur la propriété d'anti-monotonie : le Support d'un ensemble d'items n'est jamais supérieur au Support de ses sous-ensembles.

Définition 7.1 (Propriété de monotonie)**.** *Soit I un ensemble d'items et $\mathcal{J} = 2^I$ la puissance de l'ensemble I.*

Une mesure f est monotone si

$$\forall X, Y \in J : X \subseteq Y \rightarrow f(X) \leq f(Y)$$

Une mesure f est anti-monotone si

$$\forall X, Y \in J : X \subseteq Y \rightarrow f(X) \geq f(Y)$$

En pratique, on utilise l'algorithme *APRIORI*, qu'on peut implémenter sous R. Cet algorithme repose sur deux étapes, ou plutôt l'exploration des règles d'association se fait en deux étapes :

1. Identifier les ensembles d'objets fréquents : rechercher tous les ensembles d'objets fréquents avec $Support \geqslant minSupp$.
2. Générer les règles, c'est-à-dire répertorier toutes les règles d'association à partir d'ensembles d'éléments fréquents. Calculer le Support et la Confiance pour toutes ces règles. Puis éliminer les règles qui échouent aux seuils *minSupp* et *minConf*.

Pour illustrer ces étapes, et plus généralement le principe *Apriori*, on peut regarder les schémas présentés ci-dessous Figure I.5.

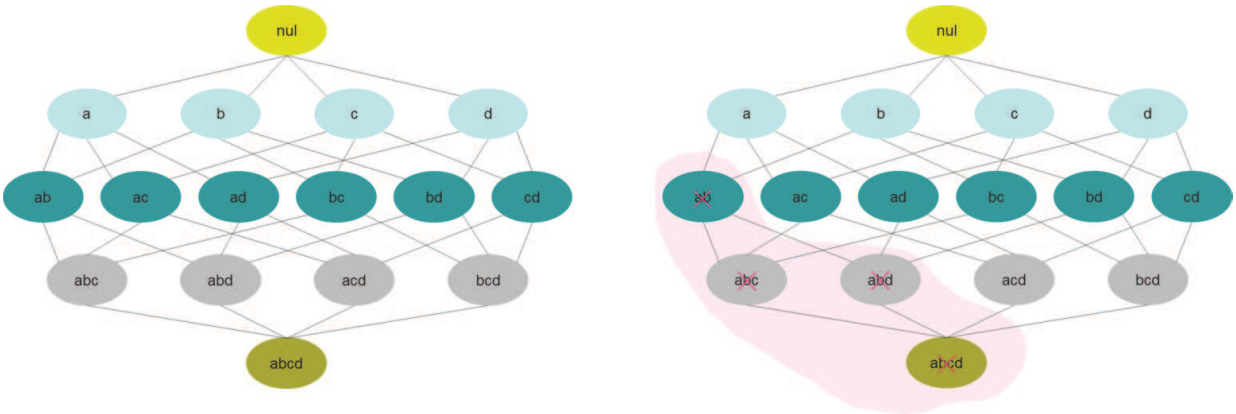


Figure I.5 – Représentation graphique du principe *Apriori*

On part du bas, d'un nœud d'ensemble d'éléments k (toutes les données), puis on commence à monter vers le haut en créant des sous-ensembles jusqu'à l'ensemble nul. Pour d nombre d'articles la taille du réseau sera de 2^d.

Pour la figure de droite : si l'ensemble d'*items* $\{a, b\}$ est peu fréquent, alors on n'a pas besoin de prendre en compte tous ses super-ensembles, et par conséquent de calculer le Support et la Confiance de ces super-ensembles.

Chapitre II

Contexte de l'étude

Le produit d'assurance sur lequel ce mémoire s'appuie, est le produit automobile de Pacifica sur le marché des particuliers.
Afin de contextualiser l'étude, et de mieux appréhender les enjeux de la modélisation du taux de transformation et de l'effet des promotions commerciales, nous commencerons par présenter le produit, et plus précisément le marché actuel de l'assurance automobile des particuliers.

1 Le marché de l'assurance automobile

1.1 L'assurance automobile, une obligation légale

Depuis 1958, tout propriétaire d'un véhicule (terrestre à moteur) est dans l'obligation de l'assurer (*a minima*), et ce, pour chaque véhicule en sa possession, même les non-roulants. Cette loi tend à la protection des tiers.
En effet, la garantie responsabilité civile, appelée également « assurance au tiers » est la garantie minimale obligatoire à souscrire. Elle couvre l'ensemble des dommages qu'un véhicule pourrait causer à des tiers : piéton, passager, autre véhicule, bâtiment… et tout individu autre que le conducteur du véhicule et/ou responsable de l'accident. Cette obligation génère une économie favorable à l'assurance automobile.

1.2 Le marché de l'assurance automobile, un marché très concurrentiel

En 2018 : 2,173 millions de voitures particulières neuves furent vendues en France. Soit presque 4 voitures par minute. 85% des ménages français possèdent au moins un véhicule, et 37% sont multi-motorisés [7]. Cependant, ces dernières années, on constate une stabilisation du parc automobile. Avec une légère hausse de 0,9% du parc des véhicules de 1ere catégorie (hors flottes) en 2018, soit une estimation de 42,562 millions de véhicules assurés.
Cette stabilisation du parc automobile accentue l'intensité concurrentielle déjà très présente sur le marché français, avec plus de 600 acteurs agréés en France en 2018, où chaque concurrent est obligé d'améliorer ses produits et d'optimiser ses prix pour préserver sa présence sur le marché.

De plus, cette concurrence est renforcée par certaines réglementations en faveur du consommateur, parmi lesquelles on peut citer la libéralisation des assurances, la *Gender Directive*, ou encore la loi Hamon…

7. D'après *L'INDUSTRIE AUTOMOBILE FRANÇAISE - ANALYSE ET STATISTIQUES 2019* publié par le Comité des Constructeurs Français d'Automobiles (CCFA)

1.3 Une concurrence renforcée par la réglementation

1.3.1 La libéralisation des assurances

Mise en place au 1er juillet 1994, la libéralisation des assurances offre la possibilité aux compagnies d'assurance ayant leur siège en Europe, d'exercer leur activité partout dans la Communauté européenne, et ainsi, de commercialiser leur produit au-delà de leur frontière. Cette ouverture de marché est rendue possible par la mise en place d'un marché unique qui se traduit par un système d'agrément unique.
La libéralisation des assurances a deux principaux objectifs. Le premier, concerne l'offre au public. Le consommateur se voit offrir un choix plus large de produits, lui permettant de trouver la police correspondant au mieux à ses besoins. Poussant ainsi les organismes d'assurance à se livrer une concurrence pour offrir des produits qui répondent aux demandes du consommateur ; à pénétrer sur de nouveaux marchés, et donc à générer une concurrence qui va s'accroître entre les entreprises. La mise en concurrence est le second objectif de l'existence du marché unique. Ainsi, les organismes d'assurance doivent améliorer sans cesse la qualité de leurs prestations, et réagir rapidement aux demandes du marché.

1.3.2 La *Gender Directive*

La *Gender Directive*, entrée en vigueur le 21 décembre 2012, est un arrêt de la Cour de justice de l'Union européenne visant à mettre fin à toute discrimination basée sur le genre des assurés. Cet arrêt interdit la tarification des primes et des prestations selon le genre de l'assuré.

1.3.3 La loi Hamon, ou loi Consommation

Entrée en vigueur le 1er janvier 2015, elle instaure de nouvelles règles assouplissant la procédure de résiliation d'un contrat d'assurance. Elle donne au consommateur le pouvoir de résilier son contrat auto, moto, ou multirisque habitation [8] quand il le souhaite, dès lors que son contrat d'assurance à une ancienneté d'au moins 1 an. En plus de cette facilité de résiliation, c'est au nouvel assureur de prendre en charge les démarches de résiliation et de veiller à la continuité de la couverture de l'assuré entre les deux contrats.
L'objectif derrière cette facilité de résiliation, est de permettre à l'assuré de changer plus facilement d'assureur, et ainsi de mieux faire jouer la concurrence extérieure.

Cependant, en France, la concurrence n'est pas tant venue de l'extérieure avec la libéralisation du marché, ni des autres assureurs européens, mais plutôt avec l'arrivée sur le marché de nouveaux distributeurs comme les banques et les groupes financiers français, qui depuis 1984, peuvent créer des filiales d'assurance ou distribuer des produits d'assurance ; mais elle est aussi venue des comparateurs d'assurance, du commerce en ligne, des mutuelles sans intermédiaires, de la vente directe, de la grande distribution …

Ces environnements économique et réglementaire accentuent la concurrence entre acteurs. D'autant que certains acteurs mènent une concurrence agressive, en pratiquant des gels tarifaires ou des tarifs très bas et connus sur le marché. Ces pratiques ont pour vocation de fidéliser les clients déjà en portefeuille, et de conquérir de nouvelles parts de marché, en réalisant des nouvelles affaires.

Cette concurrence oblige ainsi les organismes d'assurance à suivre et à réviser continuellement leurs offres, notamment en termes de tarification, afin d'attirer davantage de clientèle et en s'ajustant au mieux à ses besoins.

8. sauf pour les propriétaires, l'assurance n'étant pas obligatoire, ils devront se charger des démarches de résiliation seuls.

2 L'acte de souscription

Le contrat d'assurance est le lien juridique unissant une compagnie d'assurance à un souscripteur, l'assuré. C'est un engagement entre deux parties, où la compagnie d'assurance s'engage à verser des prestations en cas de réalisation d'un risque, moyennant le paiement d'une prime ou cotisation par le souscripteur.
La souscription est la première étape dans la vie d'un contrat d'assurance : c'est l'instant où le contrat d'assurance est établi. Cette étape est généralement précédée par la réalisation de devis et/ou de propositions.

2.1 Qu'est-ce qu'un devis ?

Un devis est une confirmation de prix pour un bien ou un service. Il contient généralement des informations pré-contractuelles à l'achat. Dans le cadre d'un devis pour une assurance, le prospect est informé des caractéristiques de l'offre et du prix, sans que ce dernier ne soit forcément adapté à ses besoins.

Un devis n'est pas un contrat, c'est un engagement unilatéral de la part de l'assureur. C'est une confirmation de prix pour un bien ou un service, devant laquelle le vendeur s'engage à n'effectuer aucune modification tant que l'acheteur n'a pas exprimé son intention de renoncer à en faire l'acquisition.

Un acheteur peut faire plusieurs devis, un pour chaque bien ou service qu'il veut acquérir. En assurance, un prospect peut donc réaliser un devis pour chaque formule proposée. Dès lors, si un des devis qui lui a été fait lui correspond, il peut le reprendre et en faire une proposition, ou souscrire. On parle alors de « devis transformé » dans ce dernier cas.

2.2 Qu'est-ce qu'une proposition commerciale ?

Une proposition commerciale ou « offre commerciale » est un document analysant les besoins du client, afin de présenter une offre élaborée qui réponde aux besoins ou à la problématique du client.
La proposition commerciale est plus détaillée, plus personnalisée que le devis, car elle replace l'offre dans un contexte, et prend la forme d'une argumentation logique.
Faire une proposition n'est pas obligatoire ; ce n'est pas une étape systématique dans le processus de vente. Elle constitue une solution plus détaillée, plus personnalisée à un besoin exprimé par le prospect.

2.3 Quel est le parcours client « type » au sein de Pacifica ?

Lorsqu'un prospect veut souscrire un contrat pour assurer un bien, il a la possibilité de faire une ou plusieurs simulations de couvertures, appelé « devis » ou « proposition ». Cette étape de simulations est nécessaire pour pouvoir souscrire un contrat.
Généralement, la première simulation faite est un devis, mais il est également possible que ce soit une proposition.
En effet, en fonction de la saisie et de l'avancée de la simulation, l'objet créé ne sera pas le même.

- Si la simulation contient seulement les données *a minima* avant d'être enregistrée, alors l'objet créé sera un devis.
- Si la simulation contient les données *a minima* ainsi que les données complémentaires - dont les données administratives - avant d'être enregistrée, alors l'objet créé sera une proposition.

- Enfin, si on a les données *a minima*, les données complémentaires et la signature au cours de la simulation alors on créé une affaire nouvelle (AN).

Cette dernière configuration est très marginale : elle représente moins de 1% de la base d'étude. Mis à part les AN, cette première simulation peut être reprise pour modification, et donner lieu à une proposition si c'était un devis, ou bien être transformée en contrat. On identifie alors 3 parcours client présentés en Figure II.1, après exclusion des prospects transformant directement en AN.
Dans 75,9% des cas, le parcours client « type » pour un prospect consistera à réaliser un ou plusieurs devis, avant une éventuelle transformation.

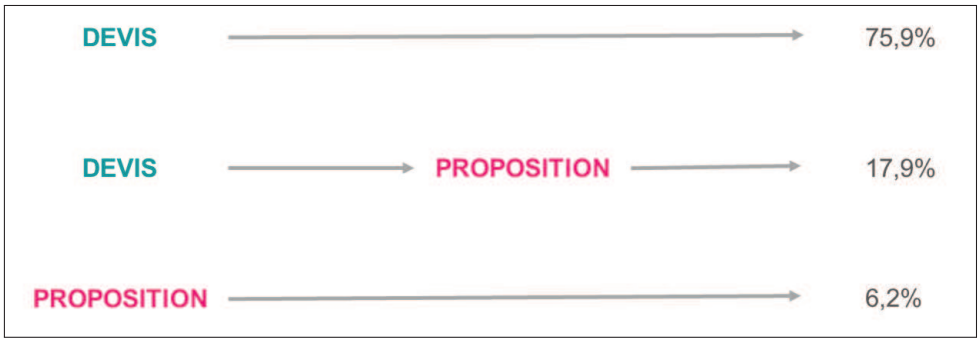


Figure II.1 – Parcours client

Actuellement, lorsqu'un prospect veut souscrire un contrat pour assurer un bien, plusieurs possibilités s'offrent à lui pour réaliser un devis et/ou une proposition. Il peut :

- aller en agence ;
- aller sur le site internet Crédit Agricole ;
- se connecter à son espace client (s'il en possède un).

Ces trois options constituent les canaux de distribution majoritaires de Pacifica. Le canal agence correspond au parcours client majoritaire : 90% des AN sont faites en agences du réseau Crédit Agricole. Le canal « en ligne », correspond aux projets fait sur des plateformes en ligne : le site internet Crédit Agricole ou l'espace client Crédit Agricole, qu'on appellera respectivement E-IARD et SNBAM.
Les projets ont donc des origines différentes en fonction du canal emprunté. Ce canal d'entrée définira l'origine d'un projet tout au long de sa vie, même si ce dernier est repris pour modification, ou pour transformation en contrat. Son origine restera le canal par lequel le prospect est rentré.

Pour la suite de ce mémoire, on utilisera le terme de « projets » pour désigner indifféremment un devis ou une proposition, et on ne tiendra compte que des projets issus du canal agence ou SNBAM.

2.4 Le suivi de l'acte de souscription

La souscription étant la première étape dans la vie d'un contrat, son étude, à l'instar de la résiliation, constitue un axe important dans le suivi et le pilotage commercial d'un produit. Elle permet de surveiller un produit, qu'il soit en déploiement ou déjà présent sur le marché, d'identifier des événements pouvant l'impacter, et de prendre des mesures correctrices si nécessaire.
Pour effectuer ces suivis, une entreprise peut mettre en place et suivre des tableaux de bord ainsi que des indicateurs de performance (KPI, *Key Performance Indicator*) afin de suivre, de comprendre, et de mesurer au mieux et de façon objective les performances de ses produits.

En assurance, il existe plusieurs indicateurs comme le taux de résiliation, qui désigne le nombre de contrats résiliés rapporté au nombre de contrats présents en portefeuille à date d'observation, ou encore le taux de transformation.

2.5 Le taux de transformation

Le taux de transformation également appelé taux de conversion, taux de concrétisation, est un indicateur très utilisé pour mesurer l'efficacité et la rentabilité d'un produit. L'analyse de ce taux permet un suivi, ainsi qu'une optimisation de la commercialisation du produit.
Il s'exprime comme le rapport entre le nombre de nouveaux clients et le nombre de prospects. Autrement dit, le nombre d'individus ayant réalisé un projet versus le nombre de clients ayant souscrit un contrat.

$$Taux\ de\ transformation = \frac{Nombre\ de\ projets\ transformés\ en\ contrat}{Nombre\ de\ projets\ réalisés}$$

Le suivi de ces indicateurs permet entre autres à une entreprise d'évaluer sa performance, et d'ajuster et piloter ses stratégies commerciales en vue de répondre au mieux aux attentes de ses clients.

Cependant, le suivi de ces indicateurs n'est pas le seul outil à la disposition des compagnies d'assurance. Elles peuvent utiliser des offres commerciales, dont l'objectif est de stimuler les ventes et le chiffre d'affaires, en incitant les consommateurs au moyen de promotions sur un produit ou un service.

3 Promotion commerciale et ciblage marketing

3.1 Les promotions commerciales

Une promotion commerciale est une technique de communication organisée de façon ponctuelle pour encourager les ventes d'un produit ou d'un service sur le court terme. Même si l'objectif final est le même, i.e. de stimuler les ventes, l'objectif pour y parvenir peut-être de plusieurs ordres :

- minimiser le risque à l'achat (surtout en cas de manque de pouvoir d'achat) ;
- fidéliser la clientèle ;
- conquérir de nouveaux clients ;
- élargir sa visibilité.

Les promotions commerciales peuvent également intervenir à des moments différents dans le cycle de vie d'un produit, afin de répondre à des objectifs différents :

- promouvoir un nouveau produit lors de son introduction sur le marché ;
- stimuler les ventes d'un produit en pleine croissance ;
- relancer l'activité d'un produit ayant atteint le stade de maturité ;
- écouler des stocks d'un produit en fin de vie.

Enfin, elles peuvent prendre différentes formes :

- des jeux concours ;
- des essais, des échantillons ;
- des ventes avec primes directes (à l'achat), différées (après l'achat), de contenants (i.e. réutilisable), de produits (vente de produits en plus) ;
- ou des réductions, telles que des bons de réduction, des offres spéciale ou ventes groupées, des offres de reprises ou d'échange de produits.

Dans le monde de l'assurance, les promotions commerciales constituent un vrai levier des ventes, et prennent le plus souvent la forme d'un avantage tarifaire : réductions de cotisations, mois gratuits, offres découvertes, jeux concours…
Pacifica ne fait pas exception à cette règle. Il est donc nécessaire de comprendre et maîtriser l'utilisation de la promotion commerciale au sein du réseau. Dans un premier temps, il est courant de penser que les promotions ne présentent que des avantages, comme la fidélisation des clients, la conquête ou la création de *plus-value*. Toutefois, elles présentent des coûts associés ; un usage intensif aura tendance à annihiler l'objectif défini en ayant un impact réduit, sans compter le coût qui lui est associé.
Par ailleurs, l'une des particularité du réseau de Pacifica en termes d'utilisation des promotions, c'est l'absence d'uniformité dans l'usage de ces dernières. Elles sont multiples et spécifiques aux agences distributrices ; il n'existe pas de stratégie nationale.

Plus généralement, les promotions commerciales s'intègrent dans une stratégie commerciale s'articulant en plusieurs volets qu'on ne développera pas dans ce mémoire, mais parmi lesquels se trouve la notion de ciblage marketing. Cette notion sera brièvement survolée dans la partie qui suit, afin de faciliter la compréhension de l'étude notamment sur l'optimisation du ciblage des opérations promotionnelles.

3.2 Le ciblage marketing

Le ciblage fait suite à l'étape de segmentation. Avec cette première étape, les consommateurs sont affectés à des groupes précis selon une étude et analyse de plusieurs critères et facteurs d'influences. On peut entre autres citer les informations clients, socio-démographiques…
Une fois cette segmentation faite, la stratégie de ciblage intervient et consiste à sélectionner les segments lui correspondant le mieux et auxquels elle souhaite s'adresser.
Cette étape de ciblage présente plusieurs avantages, tels que cibler les clients les plus appétents pour être plus efficace sur les marchés, l'optimisation des ressources…
Dans le cadre de ce mémoire, l'utilisation du ciblage a trait à optimiser les ressources, à savoir, les promotions commerciales. On cherche à identifier et à caractériser les profils de clients plus susceptibles de transformer en présence d'une intervention marketing, à savoir, une promotion commerciale.

En marketing relationnel et fidélisation des clients, 2 principes de ciblage (en intervention marketing) se dégagent :

- la sélection des personnes à forte probabilité d'achat, si elles sont à intégrées dans une campagne de marketing ;
- la sélection des personnes à forte probabilité, si et seulement si elles sont intégrées dans une campagne marketing.

La différence entre ces deux principes est très subtile, et chacun d'eux est associé à une technique de ciblage :

- l'approche traditionnelle ;
- l'approche *uplift*.

La première approche cherche à prédire si le client va acquérir le produit. Elle cible les clients émettant des signaux de souscription : c'est un ciblage par scoring [9] de l'appétence à la souscription.
Cependant, cette approche a tendance à concentrer le ciblage sur les individus qui auraient transformé quelles que soient les circonstances. De fait, la promotion commerciale perd de son efficacité, et se présente plus comme un coût que comme une *plus-value*.
Ainsi, mesurer l'impact d'une opération commerciale s'intègre parfaitement dans une approche d'optimisation de l'usage des promotions commerciales. Cela permet de statuer si l'intervention marketing a eu un réel effet, et de distinguer les clients qui n'auraient pas acquis le produit sans cette intervention, de ceux qui l'auraient acquis quelles que soient les conditions.

C'est cette première cible que l'on tend à atteindre, les clients sur qui la présence d'une promotion commerciale va influer dans le sens de l'acquisition d'un produit. C'est le segment appelé « influençable » sur la Figure II.2.

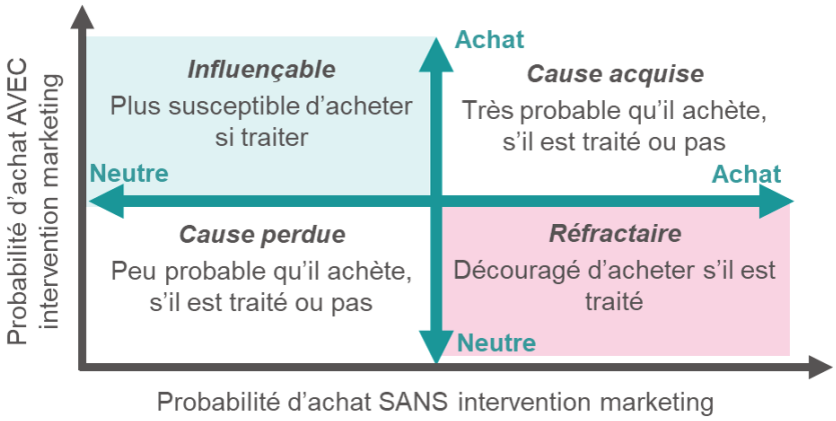


Figure II.2 – Segments clients selon leur réaction aux opérations commerciales

Les autres segments sont également à identifier car sur certaines cibles, les opérations commerciales peuvent avoir un effet neutre ou contre-productif.
En effet, certaines populations à forte propension à la souscription peuvent réagir de façon négative aux sollicitations commerciales, être découragées. Les opérations commerciales ont alors un effet contre-productif en abaissant l'appétence des clients : ce sont les « réfractaires ».
Les deux autres segments sont les « causes perdues », ceux qui n'auraient probablement pas acquis avec ou sans promotion, et les « causes acquises », à savoir, ceux qui auraient acquis en toutes circonstances, et qui de ce fait n'auraient pas forcément été les plus rentables.

L'identification de ces segments offre la possibilité d'adapter la démarche commerciale à chaque profil, et ainsi, de maximiser son retour sur investissement. En effet, la propension à la souscription et la valeur client ne sont plus les seuls critères à prendre en considération, il y a aussi la réaction du client face à une opération de fidélisation, on parle alors d'optimisation du ciblage.

9. En marketing, le scoring est une technique qui permet d'affecter à un client ou prospect, un score. Ce score traduit souvent la probabilité qu'un individu réponde à une sollicitation marketing ou appartienne à la cible recherchée. Il peut également exprimer la valeur potentielle d'un client ou prospect.

C'est à ce niveau que l'approche *uplift*, ou mesure de la sensibilité de l'action commerciale est intéressante, car elle permet d'identifier les groupes d'individus qui sont plus susceptibles de répondre positivement à une intervention marketing. Elle identifie les clients à forte valeur, ceux sur lesquels il faut concentrer les ressources consacrées à la fidélisation.
Pour cela, elle mesure l'impact de l'intervention commerciale, en calculant la différence entre la probabilité d'achat avec promotion commerciale versus celle sans promotion.
Plus la différence est élevée, plus l'impact est élevé, c'est-à-dire que l'intervention marketing augmente la probabilité d'achat. Autrement dit, des valeurs élevées seront généralement associées au segment « influençable », celui que l'on tend à atteindre.

C'est là, tout l'enjeu de ce mémoire qui se décompose en deux partie. La première vise à caractériser l'utilisation de la promotion de nos jours, notamment avec la pluralité des stratégies commerciales qui existent au sein du réseau. La deuxième partie cherche à déterminer des critères, issus des caractéristiques des profils de clients sensibles à la présence d'une promotion de façon à améliorer le processus de vente en les ciblant. Ainsi, on maximise le taux de transformation grâce à l'optimisation de l'allocation des promotions.

Chapitre III

La base de données

La constitution de la base de données est une étape très importante, si ce n'est la plus importante. Elle permet de découvrir, et de se familiariser avec l'ensemble des données, étape essentielle pour la construction des modèles ainsi que leurs interprétations.
Mais surtout elle permet de s'assurer de la cohérence, de l'exactitude, de l'homogénéité des données, ce qui est primordial pour la qualité des interprétations et de l'étude en elle-même. En effet, si les données d'entrée sont altérées, alors même le modèle le plus performant donnera des résultats incorrects, voir, faux.

Constituer une base de données répond à plusieurs questions, parmi lesquelles on peut citer :

- définir le besoin auquel la base de données doit répondre ;
- définir le périmètre d'étude, ainsi que la période d'observations ;
- identifier où se trouvent les informations, et comment les joindre si elles se trouvent dans des environnements (i.e. des tables) différents ;
- nettoyer la base de données, autrement dit, vérifier la qualité des données en supprimant les doublons, en traitant les données non normées, les valeurs manquantes… ;
- créer de nouvelles variables, retraiter celles déjà existantes, ou encore, ajouter des variables externes.

Après avoir déterminé le besoin auquel la base de données doit répondre, il est facile de définir le périmètre d'étude, et les variables qui seront utiles.

1 Périmètre de la base d'étude

1.1 Période d'observation

Notre étude s'étend sur une période de 3 ans, et s'intéresse aux projets réalisés entre le 1er janvier 2017 et le 31 décembre 2019, et aux affaires nouvelles (AN) réalisées entre le 1er janvier 2017 et le 30 avril 2020. Ces choix sont motivés pour différentes raisons.
La première, concerne la nécessité de vieillir la base de données pour prendre en compte le temps de liquidation des projets. Une étude a montré qu'un projet met au maximum 4 mois pour être transformé en contrat (cf. Figure III.1). Il est donc nécessaire de prendre en compte ce temps, et de comptabiliser les affaires nouvelles réalisées 4 mois après le 31 décembre 2019.

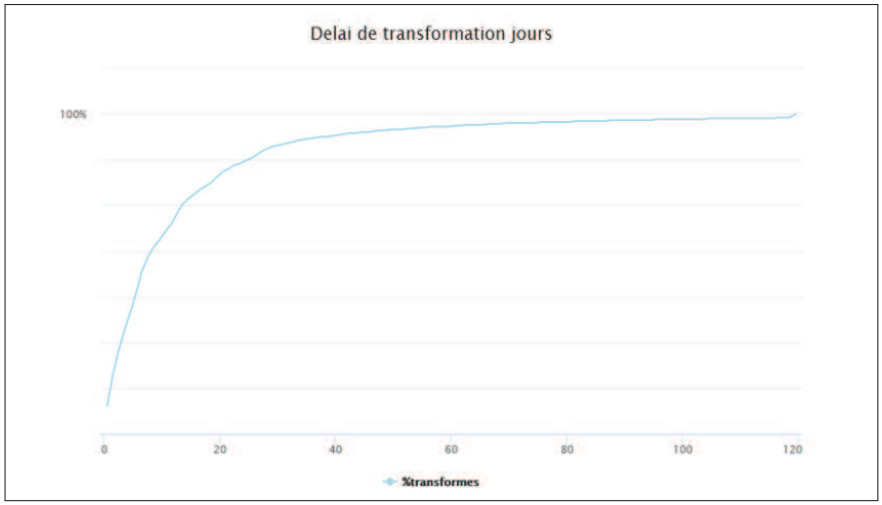


Figure III.1 – *Liquidation du taux de transformation*

Ensuite, plus l'historique est profond, plus on peut mesurer l'évolution temporelle et les tendances. Cependant, un historique trop important générera une base de données avec beaucoup d'observations et donnera des temps de traitement très long. Voilà pourquoi on se concentre sur un historique de projets, de 3 ans.

1.2 Périmètre de l'étude

Quand un projet est créé, les données qu'il contient vont alimenter et être stockées dans une base de données. Cependant, en fonction du type de projet, si c'est un devis ou une proposition, ce n'est pas la même base de données qui sera alimentée. On parle de la base *Devis*, s'il s'agit d'un devis, ou de la base *Propo*, s'il s'agit d'une proposition.
Les données se trouvant dans deux environnements différents, il était nécessaire de les réunir au sein d'une même base de données. Cependant, les informations contenues dans chacune des bases différaient quelque peu, que ce soit en termes de nom de variable, de format, ou même des variables exclusives à une base. Ces variables présentant des différences ont dû être identifiées afin d'être modifiées et adaptées quand cela est possible ou supprimées lorsqu'il s'agissait de variables non-communes aux deux bases. Cette étape d'uniformisation des tables est cruciale et constitue notre premier nettoyage ; elle permet de s'assurer de la cohérence des périmètres : devis et propositions.

1.2.1 Exclusions

Les bases de données étant multi-sources (*Devis*, *Propo*) et multi-canal (agences, internet), il a fallu exclure certains projets pour ne pas biaiser notre modélisation, notamment avec des projets en doublon ou des projets hors périmètre d'étude. Un prospect pouvant réaliser plusieurs projets pour un même bien à assurer, il était nécessaire de conserver la date de création du premier, associée aux informations du dernier projet réalisé afin d'éviter d'avoir un projet en double.
Le chaînage des projets est résumé sur la Figure III.2.

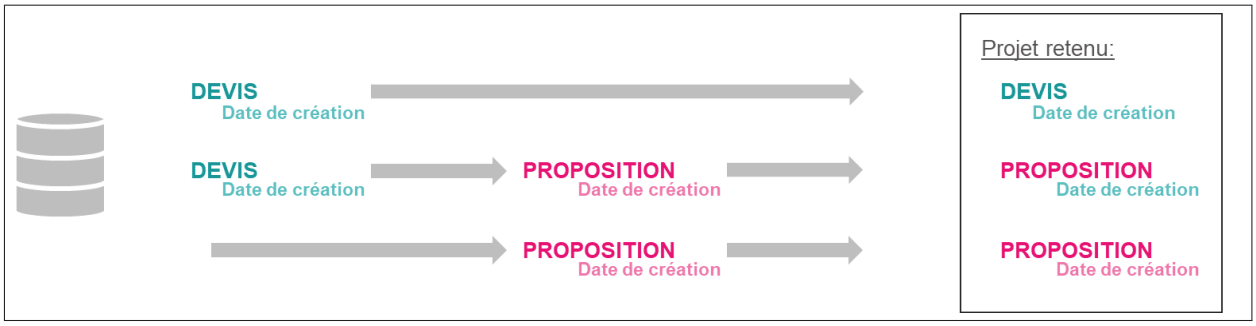


Figure III.2 – Chaînage des devis et propositions

Toujours dans l'idée d'avoir des périmètres comparables entre eux, il a fallu exclure certains canaux de distribution, ceux avec un fonctionnement bien spécifique et donc jugé trop atypiques. Il en est de même pour certaines formules. En effet, nous ne nous sommes pas intéressés aux formules dites en *run-off*[10], car en plus d'être peu représentatives, elles ne sont plus commercialisées. Elles ne sont donc plus concernées par l'allocation des promotions commerciales.

En plus de ces notions de compatibilité de périmètres, de sous-représentativité… Il faut prendre en compte l'application des résultats de notre étude *a posteriori*. En effet, si les observations et conclusions sont non applicables dans la réalité, les conserver pour la suite de l'étude n'a pas de sens. Par exemple, si la transformation est meilleure les jours de pluie, garder la variable *pluie* n'a aucun sens. Cette information est en dehors de notre champ d'action, elle est donc difficile à prendre en compte, inexploitable.

1.2.2 Regroupement des bases

Après avoir nettoyé, uniformisé les variables, les deux bases ont pu être jointes en une nouvelle base, qu'on appellera *base_etude*. Cependant, la jointure entre les bases *Devis* et *Propo* fut complexe, car aucune clé primaire officielle n'existe. Pour joindre de façon unique et sûre, il a fallu créer une clé, en concaténant plusieurs variables et en croisant les dates de création, et caractéristiques projets. En procédant ainsi, un léger biais est instauré.

1.3 La qualité des données

La qualité des données est un pré-requis à l'analyse et au pilotage car elle garantit la fiabilité des projets, à notre époque du *Big Data* et de l'émergence de la donnée non structurée. Elle garantit le respect des différentes réglementations existantes (Bâle II, RGPD…) et limite la perte d'informations.

1.3.1 Qu'est-ce qu'une donnée ? Comment définir sa qualité ?

D'après le dictionnaire Larousse, on définit une donnée informatique comme la « représentation conventionnelle d'une information permettant d'en faire le traitement automatique ».
Cette information peut être :

- quantitative (montant, âge, durée…) ;
- qualitative (nom, ville, formule…) ;
- technique (identifiant, code…).

10. *Run-off* ou liquidation de portefeuille, consiste à arrêter la souscription d'affaires nouvelles sur un portefeuille tout en gérant dans le temps, le traitement, les provisions techniques, les sinistres jusqu'à l'épuisement complet du portefeuille.

De manière générale, l'information est associée à une question :Quel est le montant de la cotisation ? Quel est le nom de l'assuré ? ...). Elle renseigne sur un état de l'objet étudié à un instant donné. Ce qui donne un caractère temporaire à la donnée. Une même question posée à deux instants différents, n'aura pas forcément la même réponse et donc la même information. Une donnée évolue (changement d'adresse, contrat avenanté…), elle devient obsolète.

Il est donc important de s'assurer de la qualité des données exploitées afin de s'assurer de la cohérence et fiabilité des résultats.
Il n'existe pas de définition universelle pour définir la qualité des données, cependant toutes s'accordent sur le fait que la qualité des données peut se décomposer en un certain nombre de critères qui dépendront de l'usage que l'utilisateur souhaite faire.
La qualité des données, et même d'une donnée ne doit pas être évaluée dans l'absolu, mais de façon relative, en fonction du besoin métier. Ces besoins étant en évolution constante, l'appréciation de la qualité ne peut pas être fixée ; elle est sujet à évolution.

On définira qu'une donnée est de qualité si elle est :

- **Accessible :** Une donnée de qualité doit être présente dans le système d'information et accessible à tout instant par les processus et utilisateurs qui l'exploitent.

- **Intègre :** La donnée ne doit pas porter de valeur aberrante, elle doit se maintenir dans la plage des valeurs acceptables, rester conforme aux normes et critères qui la définissent. Par exemple, une donnée associée à l'âge ne peut pas être négative, $âge \geqslant 0$.

- **Cohérence :** Si une donnée est présente en plusieurs endroits à la fois, elle doit toujours porter la même valeur à un instant donné. Le SI (Système d'Information) doit toujours refléter la même information, être synchronisé.

- **Exactitude :** La donnée doit refléter la réalité, être précise pour l'usage que l'on en attend.

- **Exhaustivité :** Y a-t-il des données manquantes, erronées ?

- **Unique :** Il ne doit pas exister de doublon, la donnée doit être unique.

Ces critères sont généralement vérifiés au moment du chargement des données, où les données dupliquées, erronées, incomplètes… sont corrigées.
Cependant pour s'assurer que les données restent de qualité, le meilleur moyen est de faire de la qualité de données, c'est-à-dire mettre en place une (« bonne ») stratégie de gestion des données. Cela se traduit par la mise en place de process qui veillent à la non-dégradation de l'information, notamment au risque d'obsolescence.

1.3.2 Origine de la dégradation de la donnée

Les origines de la « mauvaises qualité » des données sont diverses et variées, on peut entre autres citer :

- L'import et la migration : risque d'incompatibilité de format, de détérioration… générant un risque de non-cohérence, de doublon, de pertes…
- Le partage d'informations (qui augmente le risque de doublon et/ou de perte d'observation) : une information doit être unique, cependant si elle est commune à plusieurs services le risque de doublon ou de modification par l'un, augmente ; entraînant la possibilité d'avoir deux informations au lieu d'une.

- Des entrées par plusieurs utilisateurs : risque de non-uniformité et donc de divergence, surtout si c'est en saisie libre. Pour une même modalité de variable, il peut y avoir différentes saisies. Selon l'utilisateur, pour l'un, la civilité sera « Mme » et pour un autre, ce sera « Madame ». Généralement, pour éviter ces erreurs, on limite la saisie libre en créant des outils de pré-saisies tels que des listes déroulantes.
- L'erreur de saisie : souvent liée au manque d'expertise, à la monotonie d'une tâche ou à l'erreur humaine comme la faute de frappe ou la saisie dans le mauvais champ. Cependant, elle peut parfois être volontaire afin gagner du temps ou par simplicité.

Ces deux dernières causes d'erreur favorisent le manque de précision des données, et les informations manquantes.
Elles sont également à l'origine de la correction orthographique appliquée à l'une des variables de notre étude, la variable *orgcie* qui contient le nom de l'ancienne compagnie d'assurance du prospect. Mais avant de parler de la correction orthographique, on va rapidement aborder le traitement des doublons et des valeurs manquantes.

1.3.3 Traitement des doublons et des valeurs manquantes

Cet aspect sera abordé brièvement dans cette partie et ne sera pas développé par la suite, car il ne constitue pas un élément majoritaire du mémoire.

Un doublon, ce sont deux observations (ou plus) identiques au sein d'une même base de données. Cette répétition génère une redondance de l'information et est source d'erreur en instaurant un biais.
Il est donc nécessaire de supprimer les projets dupliqués. Pour cela, il faut identifier de façon unique chaque observation à l'aide d'une clé primaire [11]. Dans le cas de notre étude, la clé primaire est la concaténation de plusieurs variables.

La correction des valeurs manquantes est complexe, car il faut remédier à ce manque d'information sans altérer significativement le jeu.
Plusieurs méthodes existent, on peut par exemple citer :

- La suppression des observations présentant des *Na*. La limite de cette technique est d'écarter trop d'observation et de se retrouver avec un jeu de données très réduit.
- L'imputation de données. Cette technique consiste à remplacer la valeur manquante par une valeur artificielle, le plus souvent la modalité de référence, la moyenne ou le mode…

Ici, nous avons décidé que pour les variables qualitatives, si les valeurs manquantes représentent plus de 80% alors on ne conserve pas la variable. Pour les variables quantitatives, on a choisi de normaliser nos variables afin de réduire le problème d'échelle entre les variables et ainsi remplacer les *Na*. Cette méthode instaure cependant un biais : on perd la valeur d'origine.

1.3.4 Traitement des erreurs de saisie et entrées multiples - Correction orthographique

Une des variables présente dans notre base de données et qu'on souhaite exploiter dans nos modèles est en saisie libre, et à la main de l'agent qui réalise le projet.
L'entrée de cette variable est donc multiple et présente des milliers d'orthographes pour une seule et même modalité (i.e. compagnie d'assurances) et cela même sans tenir compte de la casse et de la ponctuation.

11. Une clé primaire permet d'identifier de façon unique chaque individu dans une table de base de données, elle ne peut donc pas contenir de valeur *Null*. Chaque table ne peut contenir qu'une seule clé primaire, constituée d'une ou plusieurs colonnes.

Il est donc légitime de penser que toutes ces orthographes ne sont pas correctes, et que les sources d'erreurs sont nombreuses et banales tels que :

- les erreurs de saisie, de typographie ou de transcription ;
- les données pouvant être épelées ou abrégées de manières différentes.

Et dans des cas plus larges, on peut rencontrer :

- des erreurs liées au format ;
- des erreurs liées au changement d'orthographe selon l'époque, la géolocalisation, c'est notamment vrai pour les prénoms ;
- des erreurs liées aux imports qui peuvent dégrader les données en transcrivant mal les caractères accentués, en modifiant certains formats, tronquant des données, en changeant le format (ex : les dates) lorsque ce dernier n'est pas universel et spécifié.

Il est donc nécessaire d'envisager une correction orthographique, afin d'améliorer la qualité de la donnée. Cette correction orthographique repose sur 3 axes :

- définition d'un mot erroné (un vert rempli d'eau) ;
- mot correctement orthographié, mais n'appartenant pas au lexique de la langue d'étude (*Veepee* est un site de e-commerce) ;
- notion de similarité, distance ou dissimilarité.

Cependant, toutes ces corrections sont intuitives pour un humain, il comprend la différence entre les mots vert et verre, entre Pacifica et Paifica. En revanche, c'est plus complexe pour une machine.

C'est pour cela, qu'on a mis en œuvre des méthodes de Recherche Approximative (RA) pour corriger la variable, à savoir la distance de Levenshtein et l'algorithme *Soundex*.
En effet, les principales sources d'erreurs rencontrées sont liées aux fautes de frappe et d'orthographe. Si on prend comme exemple la compagnie d'assurances Pacifica, on peut rencontrer l'orthographe exacte mais également les orthographes *PACIIFIA*, *PACIFIAC*, *PAICFICA*… ou des orthographes liées aux abréviations et autres désignations : *PCK*, *CR*, *CACR*… pour Pacifica, et plus largement Crédit Agricole.

La distance de Levenshtein et l'algorithme *Soundex* se veulent complémentaires, ces méthodes permettent d'appréhender la chaîne de caractères à corriger d'un point de vu orthographe « pur » et phonétiquement.
Mais avant de pouvoir les appliquer, il nous a fallu procéder à quelques étapes intermédiaires.

1.3.4.1 Extraction de la variable à corriger

Tout d'abord appliquer de la RA à l'intégralité de la base de données n'a aucun sens, et nécessite beaucoup de ressources (environ 4 millions de lignes à traiter).
Une extraction de la variable d'intérêt (celle à corriger), après suppression des doublons est entièrement suffisante. De fait, on passe à environ 7 000 lignes, soit 7 000 orthographes à corriger.

1.3.4.2 Uniformisation des chaînes de caractères

Une fois, l'extraction réalisée, la première étape est d'uniformiser les chaînes de caractères à corriger. Par uniformisation, on entend appliquer quelques règles telles que mettre tous dans la même casse (ici, en majuscule), supprimer les accents, les symboles (tiret, parenthèse, *underscore*…), remplacer les espaces multiples par des espaces simples…
De fait, ce sont les chaînes de caractères « brutes » que nous comparerons sans caractères « parasites ». Les chaînes de caractères *Crédit Agricole* et *Credit agricole* seront identifiées comme identiques, car les accents n'existeront plus, idem pour les minuscules.

1.3.4.3 Création d'un dictionnaire

La seconde étape consiste à définir un dictionnaire de modalités correctement orthographiées, auxquelles les modalités de la variable à corriger pourront être comparées. En effet, les méthodes que nous utilisons reposent sur la comparaison de chaînes de caractères deux à deux. Il est donc essentiel de constituer un dictionnaire de modalités «correctes ». Ces modalités se doivent également de respecter les règles cités ci-dessus et qui ont été appliquées à la variable à corriger.
Une fois les variables « uniformisées » et le dictionnaire constitué, on peut comparer nos chaînes mal orthographiées à celles du dictionnaire.

1.3.4.4 Représentation alphanumérique et distance de Levenshtein

La distance de Levenshtein, calcule le degré de ressemblance entre deux chaînes de caractères. Cette distance est donc très appropriée pour corriger les fautes de frappe.
Si on reprend notre exemple de *Pacifica*, on peut voir que pour une orthographe, on dénombre des dizaines d'orthographes [12] *PACIIFIA*, *PACIFIAC*, *PAICFICA*, *PCK*…
Intuitivement on comprend que la correction à apporter serait de corriger ces orthographes par *PACIFICA*, mais pour une machine c'est plus compliqué.
Ainsi, chaque observation de la variable à corriger est comparée à chaque élément du dictionnaire et la distance de Levenshtein est calculée pour chaque association, ainsi que le degré de similitude, qui on rappelle se calcule :

$$\frac{(|a| + |b|) - lev_{a,b}(i,j)}{|a| + |b|}$$

où |a| et |b| correspondent aux longueurs des chaînes de caractères a et b.

	ALLIANZ	*AXA*	*GAN ASSURANCE*	*GROUPAMA*	*MNT*	…	*PACIFICA*
PAIFICA	*6*	*5*	*11*	*7*	*7*	…	*1*
AXA	*5*	*0*	*11*	*6*	*3*	…	*6*
GROUPAAM	*7*	*7*	*10*	*2*	*8*	…	*8*
MAIF	*6*	*3*	*12*	*7*	*3*	…	*5*
...	...	...	...	...	...	...	...

On obtient ainsi une base de données avec les champs à corriger en ligne, et en colonnes les valeurs de référence (valeur du dictionnaire) et à chaque croisement la distance de Levenshtein.

L'étape suivante consiste à sélectionner pour chaque champ, l'association présentant la distance la plus faible et à calculer le pourcentage de similitude. Si plusieurs associations présentent la même distance minimale alors on sauvegarde les 10 premières associations associées à cette distance de Levenshtein minimale et on calcule le pourcentage de similitude.

Dans notre exemple, pour la modalité *PAIFICA*, on retient *PACIFICA* avec une distance de Levenshtein de 1 et un pourcentage de similitude de 93.3%. En revanche pour la modalité *MAIF*, si on ne considère que ces 7 références de dictionnaire, on retiendrai *AXA*, *MMA* et *MNT*, avec une distance de Levenshtein de 3 et un pourcentage de similitude de 57.2%. La base de données post-étape est semblable à :

	TOP10	*DIST*	*PCT*
AXA	*AXA*	*0*	*100.0*
GROUPAAM	*GROUPAMA*	*2*	*87.5*
MAIF	*AXA, MMA, MNT*	*3*	*57.2*

12. Et cela même après avoir mis tout en majuscule, supprimé les accents, les symboles, les espaces multiples…

On aurait pu se contenter de ces résultats et arrêter là, la RA. Cependant, comment savoir quelle suggestion choisir lorsqu'il y a plusieurs associations possibles ?
Pour départager ces cas-là, et pour affiner la RA, nous nous intéressons à la sonorité de la chaîne mal orthographiée.

1.3.4.5 Représentation phonétique - Algorithme *Soundex*

En effet, si on se base seulement sur la ressemblance entre deux chaînes de caractères, on aura tendance à corriger les fautes de frappe sans forcément tenir compte des fautes d'orthographe en tant que telle.

Si on prend la compagnie d'assurance *AMAGUIZ*, respectivement écrite *AMAGISE*. On trouve une distance de Levenshtein de 3. Si on garde cette orthographe et qu'on la compare à une autre compagnie, *AMALINE*, on trouve une distance de Levenshtein entre ces deux chaînes de caractères plus faible, 2. En regardant juste les chaînes de caractères on corrigerait *AMAGISE* par *AMALINE* avec un pourcentage de ressemblance de 85,7%.

Sur cet exemple, on constate que la RA sur chaîne alphanumérique n'aurait pas soumis la bonne suggestion. Donc, en plus de départager en cas de suggestions multiples, la représentation phonétique permet également d'éviter certaines erreurs.

La mise en œuvre est similaire à celle réalisée pour la distance de Levenshtein. L'étape en plus concerne la transcodification de chaque chaîne de caractères en code *Soundex*, que ce soit celles à corriger ou celles du dictionnaire.
Brièvement, le *Soundex* est un algorithme qui convertie une chaine en un code « phonétique » en suivant des règles reposants sur la consonance des lettres. Ces règles sont rappelées ci-dessous dans la Figure III.3.

Règles :

1. Mise en majuscule du mot
2. On conserve la première lettre du mot
3. Elimination des voyelles, du H et du W
4. Transcodification selon les règles suivantes:

Lettres	B P	C K Q	D T	L	M N	R	G J	S X Z	F V
Code	1	2	3	4	5	6	7	8	9

5. Élimination des lettre doubles
6. Conservation de 4 caractères du *Soundex*, à compléter par des 0 le cas échant

Figure III.3 – *Règles de l'algorithme Soundex (version langue française)*

Si on reprend le mot *AMAGUISE*, l'algorithme *Soundex* nous donner A578 comme code «phonétique».
Ce code est obtenu, en reprenant le procédé de l'algorithme de la manière suivante :

1. *Retranscription du mot en majuscule (déjà le cas)*
2. *Conservation de la première lettre ; A*
3. *Suppression des voyelles et lettres muettes ; AMGS*
4. *Transcodification des lettres restantes ; M $\longrightarrow$ 5, G $\longrightarrow$ 7, S $\longrightarrow$ 8.*
5. *Suppression de toutes les paires consécutives de chiffres dupliqués (étape non-nécessaire dans cet exemple)*
6. *Conservation des 4 premiers caractères du code Soundex, à compléter par des 0 si besoin ; A578*

Une fois, que chaque chaîne est transcodée on procède de manière analogue à la représentation alphanumérique, c'est-à-dire qu'on utilise la distance de Levenshtein. Chaque code *Soundex* (associé à une chaîne de caractères mal orthographiée) est comparé un à un à ceux du dictionnaire. Pour chaque association, on calcule la distance de Levenshtein, ainsi que le pourcentage de similitude, avant de sélectionner l'association présentant la distance la plus faible (ou les associations en cas de suggestions multiples).

En reprenant notre exemple avec *AMAGISE*, on trouve une ressemblance de 100% entre *AMAGISE* et *AMAGUIZ* contre 50% avec *AMALINE*. Ce qui est plus exact que la suggestion post représentation alphanumérique, qui nous suggérait AMALINE à 85,7%.

Maintenant qu'on a effectué une RA selon les deux représentations possibles d'une chaîne de caractères, il nous faut rassembler les suggestions émisent par ces deux méthodes, et les agréger.

Pour connaître quelle suggestion prendre, des règles sont établies.

- Si les deux approches suggèrent la même compagnie d'assurance, alors c'est elle qui est suggérée en tant que correction, et on conserve la moyenne des pourcentages de similitude.
- Si les deux approches proposent des suggestions différentes, alors on conserve la compagnie avec le pourcentage le plus élevé et on fait la moyenne des pourcentages de similitude.
- Si au moins une des deux approches suggèrent plusieurs compagnies, alors on conserve la suggestion qui est commune aux deux approches et on fait la moyenne des pourcentages de similitude.

Pour AMAGISE, on rappelle que la correction orthographique basée sur la représentation alphanumérique suggère AMALINE avec un pourcentage de similitude de 85,7%, tandis que la représentation phonétique suggère AMAGUIZ avec un pourcentage de similitude de 100%. La suggestion retenue serait AMAGUIZ, et le pourcentage de similitude associé serait la moyenne de 100% vs 85,7%, soit 92,9%.

Le schéma présenté en Figure III.4 résume la stratégie de correction orthographique appliquée à la base de données et permet d'illustrer avec un second exemple.

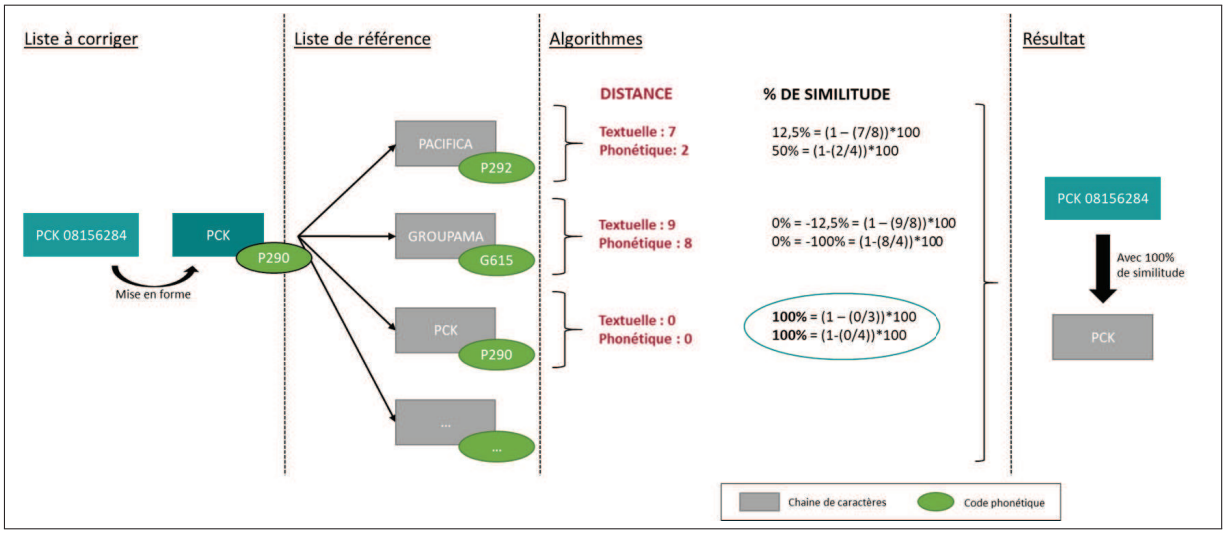


Figure III.4 – *Schéma illustrant la correction orthographique pour une observation*

Une fois ces règles appliquées, chaque chaîne de caractères à corriger se voit associer une suggestion de correction [13], ainsi que le pourcentage de similitude entre ces deux chaînes. Un pourcentage de 100 indique une correspondance parfaite ; c'est le cas quand la chaîne à corriger est déjà correctement orthographiée. A l'inverse, plus le pourcentage est faible, moins la ressemblance est avérée. Si on reprend nos exemples, on peut voir que la RA est juste et qu'elle a permis de suggérer une correction, mais ce n'est pas toujours le cas, entraînant ainsi un biais.

1.3.4.6 Vérification des résultats

Cette correction de variable par RA n'est pas toujours exacte, de nombreux biais existent comme la sélection d'une suggestion en cas d'égalité, ou encore le fait que le dictionnaire soit non exhaustif et ne couvre pas toutes les compagnies d'assurances.
De plus, il nous a fallu définir un seuil en deçà duquel le pourcentage de similitude n'est plus fiable. Pour les 7 000 champs à corriger, nous avons vérifié si la suggestion proposée par l'algorithme est correcte ou non. A partir des résultats obtenus, la part de correction exacte, de correction fausse et le taux d'erreur sont tracés. Les résultats sont présentés en Figure III.5, et le seuil est établi à 70%.

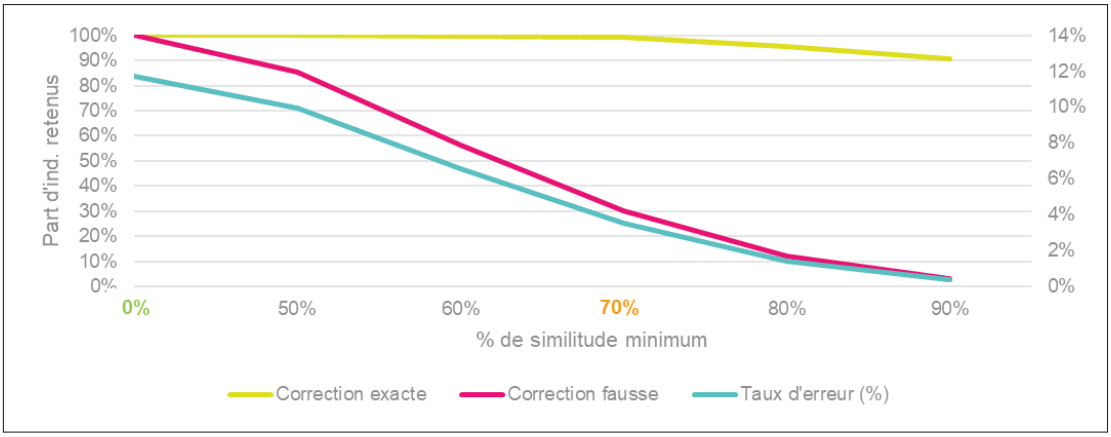


Figure III.5 – *Taux d'erreur après correction orthographique de la variable orgcie (ancienne compagnie)*

13. « suggestion de correction » est un abus de langage, car toutes les chaînes n'ont pas besoin d'être corrigées, dans ce cas les distances sont nulles, car la correspondance est parfaite.

D'après le graphe, on constate que l'algorithme nous donne un taux d'erreur de 11,7%. Cependant si on regarde la part d'individus avec une correction exacte ou fausse selon le pourcentage de similitude, on observe qu'avec un seuil de 70% de similitude minimum, on corrige 91% des observations avec un taux d'erreur de 3,5%. Parmi ces observations corrigées, on corrige 30% des individus, avec une « mauvaise » suggestion et 99% des « bons ».

Ainsi, pour les observations avec au moins 70% de similitude, on conserve la suggestion et pour les individus ne respectant pas ce seuil, on force la valeur à « *INCERTAINE* ».

1.3.4.7 Regroupement des modalités

Une fois la correction orthographique achevée, la nouvelle variable présente plus de 100 modalités, ce qui est difficilement interprétable en sortie de modèles. Pour résoudre ce problème, on décide de regrouper certaines modalités entre elles.
In fine, on obtient une nouvelle variable, *orgcie3*, valorisée par le nom de l'ancienne compagnie d'assurances corrigée et regroupée ; qu'on joint à la base de données afin d'y ajouter le nom de la compagnie d'assurances, suggéré par le processus de correction orthographique.

1.3.4.8 Les limites de la correction orthographique

Cependant, même si cette variable a été corrigée, il ne faut pas oublier qu'elle est à exploiter avec parcimonie, car elle reste une variable biaisée. Ce biais est issu principalement de trois raisons. La première est qu'il s'agit d'une variable déclarative, le prospect est libre de dire ce qu'il veut. Il peut donner une fausse information par erreur ou volontairement (s'il sait par exemple que sa compagnie actuelle est plus chère, il peut en dire une autre pour ne pas influencer l'agent). La seconde raison concerne la saisie, le gestionnaire peut saisir ce qu'il veut également. Il est libre de renseigner la variable. La part de *Na* dans cette variable est donc à interpréter avec prudence. Enfin, le dernier biais provient du processus de correction lui-même. En effet, on corrige la variable en s'appuyant sur des règles et un seuil qu'on a nous-même fixé.

2 Ajout de nouvelles variables

Afin de conceptualiser et modéliser correctement le taux de transformation, enrichir la base de données de nouvelles variables est essentiel. Plus les variables sont diverses et variées, plus elles couvrent des aspects différents de l'étude.
Cet enrichissement passe par la création de nouvelles variables, mais également par l'ajout de données clients, ainsi que des données externes.

2.1 Création de nouvelles variables

A partir des données disponibles dans notre base, de nouvelles variables sont créées, dont :

- ***nb_transfo***, une variable indicatrice qui renseigne sur l'état du projet, à savoir s'il a été transformé ou pas. Elle prend en valeur 1 si le projet a été transformé et 0 le cas échéant.

- ***nb_promo***, également une variable indicatrice qui indique si le projet a été fait avec une promotion commerciale. Elle prend 1 s'il y a eu une promotion, et 0 le cas contraire.

- ***valprom***, qui correspond à la valeur de la promotion. En effet, la valeur de la promotion n'est pas connue directement. Elle est contenue dans deux variables, l'une nous renseignant sur le type de la promotion, un montant ou un pourcentage, et la seconde sur la valeur associée. Comparer un montant à un pourcentage n'étant pas possible, il a donc

fallu assembler ces deux informations afin d'avoir une valeur de promotion comparable, quel que soit le type de promotion.

A partir de ces variables on peut entre autres calculer le taux de transformation, le taux de promotion, la valeur moyenne d'une promotion... des variables essentielles à la modélisation du taux de transformation.

Par anticipation de la seconde phase de l'étude, à savoir caractériser l'usage de la promotion commerciale, d'autres variables sont également créées. Ces dernières ne sont pas à intégrer à la modélisation du taux de transformation, mais elles constituent une étape dans la constitution de notre base finale et seront présentées dans le Chapitre IV.

Enfin des données internes et externes sont ajoutées.

2.2 Données internes, l'équipement client

En effet, quand un prospect fait un projet, et qu'il est déjà client, il possède un identifiant client (*inclient*), à partir duquel on peut rapprocher d'autres informations comme connaître les contrats qu'il a souscrit, est-ce qu'ils sont toujours actifs ?...

Toutes ces informations nous permettent d'appréhender l'équipement « avant projet » d'un client, de savoir s'il est multi-équipé ou pas, et sur quel produit. Cette approche de multi-équipement n'est pas nécessaire, mais elle constitue une hypothèse à vérifier dans notre étude. Est-ce que les prospects multi-équipés transforment mieux ? Notamment sur le multi-équipement habitation, c'est pour cela qu'on a distingué le multi-équipement habitation des autres produits, en créant une variable indicatrice, *multequip*, qui prend 1 en valeur, si le prospect possède déjà un contrat habitation chez Pacifica, sinon 0.

2.3 Données externes

Toujours dans l'optique de consolider notre modèle, on a ajouté des données externes, notamment pour caractériser l'environnement du bien assuré.
Idéalement, pour caractériser le bien assuré, il nous faut connaître l'adresse du lieu de stationnement, soit l'adresse du parking. Malheureusement, cette information n'étant pas suffisamment renseignée, il nous a donc fallu trouver une alternative. Nous sommes partis de l'hypothèse que l'adresse du lieu de résidence du prospect, était également celle du lieu de stationnement. A partir de cette adresse, et plus précisément du code INSEE associé nous avons pu ajouter,

2.3.1 Des données socio-économiques

Des données socio-économiques publiques et communiquées par l'INSEE, telles que le revenu moyen, le taux de chômage, la densité de population par commune…

2.3.2 Des données géocodées

Ces données géocodées répondent à un besoin, celui d'essayer de caractériser l'intensité concurrentielle. Quelle est la distance entre l'adresse du projet (qui rappelons le, correspond à l'adresse du bien assuré dans notre hypothèse) et l'agence Crédit Agricole la plus proche ? Est-ce que la proximité à une agence Crédit Agricole impacte la transformation ? Une faible distance laissera penser que la présence des agences est suffisante dans ce secteur, à l'inverse une distance importante supposera que Crédit Agricole n'est pas suffisamment présent, et par conséquent potentiellement moins concurrent.

Pour commencer, nous avons calculé la distance à vol d'oiseau entre l'adresse du projet et l'agence Crédit Agricole la plus proche. Pour ce faire une étape de géocodage fût nécessaire, afin de géocoder les projets d'une part et les agences de l'autre.

2.3.2.1 Géocodage

Le géocodage s'est déroulé en deux temps. D'abord nous avons collecté et stocké les adresses dans deux bases de données distinctes, une avec les projets et l'autre avec les agences, avant d'être géocodés sous R en appelant l'API [14] de la BAN (Base de Données Nationale). Le géocodage consiste, à donner la latitude et la longitude associées à une coordonnée, dans notre cas l'adresse de l'agence ou du projet.

2.3.2.2 Calcul des distances

Une fois la latitude et la longitude obtenue, on peut calculer la distance entre chaque projet et chaque agence référencée dans les bases de données. Enfin, pour faciliter la lecture *a posteriori*, les distances sont mises sous forme de quantiles. A l'issu de cette étape, nous avons une nouvelle variable : *dist_agence_crca* qu'on ajoute à notre base de données.

2.3.2.3 Vérification de la géocodification

Cependant, comme tout algorithme, il n'est pas infaillible. Il a donc fallu vérifier le géocodage. Pour se faire, il fallait s'assurer de la cohérence entre les coordonnées remontées (latitude et longitude) et l'adresse.
Intuitivement, on peut penser à une vérification manuelle. Sous-entendu, sélectionner quelques adresses aléatoirement, vérifier leur géocodage sur internet (avec une application comme *Google Maps*) et faire l'hypothèse que si 10% de la base est correcte alors on peut généraliser cette observation à toute la base.
Cette méthode bien que possible, n'est pas la plus optimale, d'autant qu'il serait question de vérifier plus de 400 000 adresses pour couvrir à peine 10% de nos données.

La solution abordée, est de calculer la similitude entre l'adresse d'entrée et celle de sortie, celle que l'algorithme remonte et qui correspond aux coordonnées.
En effet, une dissimilitude entre les deux adresses peut supposer une erreur de géocodage.
Pour illustrer, on utilise deux exemples fictifs.

Adresse entrée	*Latitude*	*Longitude*	*Adresse sortie*
37 bis Rue Jeanne d'Arc, 76000 Rouen, France	*49.4427305*	*1.090762*	*37 bis Rue Jeanne d'Arc, 76000 Rouen, France*
50 rue Gambetta, 92170 Vanves, France	*48.823829*	*2.2961675*	*5 rue Gambetta, 92170 Vanves, France*

Dans le premier exemple, on constate que l'adresse d'entrée et de sortie sont identiques, ce qui laisse supposer que les coordonnées affichées sont bien celles de l'adresse à géocoder. A l'inverse, pour le second exemple, les adresses d'entrée et de sortie sont différentes : les coordonnées indiquées par l'algorithme correspondent à l'adresse de sortie. Or c'est l'adresse d'entrée qui nous intéresse. On peut donc supposer que lorsque l'adresse de sortie est différente de celle d'entrée, il y a un risque de mauvais géocodage.
Dans le cas de notre l'exemple, l'erreur de géocodage est liée au fait que le numéro 50 de la rue Gambetta n'existe pas. L'algorithme, remonte alors l'adresse qui lui semble la plus probable et la géocode. Dans le cas d'une erreur de numéro, l'erreur de géolocalisation est minime et

14. L'API ou interface de programmation applicative (*Application Programming Interface* en anglais), peut être résumée à une solution informatique qui permet à des applications de communiquer entre elles et de s'échanger mutuellement des services ou des données.

n'impacterait presque pas le calcul des distances. En revanche, si l'erreur porte sur le nom ou le type de la voie (rue, avenue, boulevard…) ou de la ville, la géocodification ne sera pas la même, et les distances non plus.

Connaître la dissimilitude entre deux adresses peut être un bon indicateur pour mesurer la qualité du géocodage.
Pour calculer cette dissimilitude, ou similitude, cela revient à comparer deux chaînes de caractères et à comptabiliser les différences.
Comme dans la partie précédente, où l'on corrige une variable de type chaîne de caractères, on a utilisé la distance de Levenshtein pour mesurer la similitude entre nos adresses d'entrée et de sortie. Puis on a fixé des seuils, afin de définir la qualité. Si la distance est nulle, alors les adresses sont identiques. Si la distance est comprise entre 1 et 5 alors on peut supposer que les adresses sont semblables, et au-delà de 5 les adresses sont différentes.

Une fois tous ces différents traitements appliqués à la base de données, cette dernière contient environ 4 millions d'individus et 80 variables présentées en Annexe B.

3 Base d'apprentissage et de test

Enfin, nous divisons notre jeu de données en deux : une base d'apprentissage (environ 2/3 des données) et une base de test (le 1/3 restant) sur laquelle on va évaluer notre apprentissage. Concernant les paramètres optimaux de nos différents modèles, à savoir, *GBM*, *Random Forest*, régression pénalisée, ils seront déterminés par validation croisée (*cross validation* (CV)).

En effet, se contenter d'une division du jeu de données en deux, n'est pas suffisant. Les prédictions pourraient être hasardeuses et instables si l'on relançait la modélisation. Idéalement, il faudrait avoir une troisième base, une base de validation pour calibrer le modèle avant de procéder à la prédiction. Le problème, de ce partitionnement est qu'il nécessite un échantillon de grande taille. Une solution est de procéder à une validation croisée. Cette procédure consiste à diviser en k sous-échantillons (ou plis) le jeu de données. Chaque pli sera ensuite utilisé une fois comme validation tandis que les k - 1 plis restants forment l'ensemble d'apprentissage.

Chapitre IV

La stratégie promotionnelle de nos jours

Avant de pouvoir mesurer l'impact des promotions commerciales, il nous faut au préalable comprendre ce qu'est une promotion ? Comment est-elle utilisée actuellement en agence ? Y a-t-il des profils d'agences en termes d'utilisation de la promotion commerciale ? L'objectif de ce chapitre est donc de donner une image de l'utilisation des promotions commerciales de nos jours, dans le réseau Pacifica et plus particulièrement au sein des caisses régionales (CR).

De nos jours, l'étude des promotions commerciales est complexe, car elles sont à la main des agents, et chaque CR à sa propre politique commerciale. Cette multitude de pratiques commerciales est donc difficilement appréhendable et soulève plusieurs interrogations auxquelles nous allons tenter de répondre.

Afin de visualiser la situation actuelle, et définir ce qu'est une promotion commerciale au sein de Pacifica, un tableau de bord (TDB) et des ACP (Analyse en Composantes Principales) ont été développés.

1 Des disparités régionales

Depuis 3 ans, le taux de transformation du réseau est stable, contrairement au taux de promotion qui croît chaque année.
En revanche, si on descend à l'échelle des CR cela devient moins stable et des différences géographiques apparaissent. Certaines CR, présentent un taux de transformation largement supérieur au taux nationale tandis que d'autres, affichent un taux plus bas. Il en est de même pour le taux de promotion.
Ces différences entre les CR peuvent en partie être justifiées par une intensité concurrentielle différente entre les régions. Par exemple, elle est plus faible en Guadeloupe et Martinique qu'en métropole. Cependant, l'intensité concurrentielle ne peut pas tout justifier, les stratégies commerciales ainsi que le comportement client sont des facteurs à mesurer et à prendre en compte.

Pour pouvoir évaluer les stratégies commerciales, il faut au préalable les identifier et les définir.

2 Des indicateurs caractérisant la stratégie commerciale...

Précédemment, nous nous sommes appliqués à la constitution de la base de données utilisée pour la modélisation du taux de transformation, qui sera abordée dans le chapitre suivant. Cependant, certaines variables non-nécessaires au modèle, le sont pour l'étude axée sur les promotions commerciales, c'est pourquoi nous les intégrons à notre base de données créée pour cette partie.

Malgré l'absence d'uniformité dans l'utilisation des promotions commerciales entre les CR, les connaissances métiers nous ont permis de définir un ensemble de variables pouvant potentiellement nous aider à caractériser l'utilisation des promotions. Ces variables tendent à mettre en lumière des pratiques commerciales, l'environnement de souscription et de réalisation des projets. Plus généralement, elles vont nous permettre de répondre aux questions telles que : Par qui les promotions sont utilisées ? Quand ? Comment ?

Parmi ces variables, on peut citer :

- L'identification des vendeurs « habituels » et des vendeurs « occasionnels »
  L'une des hypothèses soulevées et relatives à l'usage des promotions commerciales concerne le statut des vendeurs.
  Est-ce que les vendeurs « occasionnels », c'est-à-dire les vendeurs non formés à la vente de produits d'assurance (on rappelle que les agents en agence sont des agents bancaires) usent plus régulièrement de promotions ? En vue de transformer plus de contrat et dissimuler leur manque d'expérience.
  L'identification des vendeurs s'est fait à partir de leurs résultats, et plus précisément à partir du nombre de projets réalisés. Si le vendeur fait partie des 20% d'agents à faire le plus de projets, alors on l'identifie comme un vendeur « habituel », sinon « occasionnel ».

- Les « TOP » vendeurs et promotions
  A partir de l'information précédente, vendeurs « habituels » versus « occasionnels », la création et l'affichage d'un top vendeur a pu être possible dans le TDB. Ce top affiche les n « meilleurs » vendeurs selon le nombre de projets réalisés, où n reste définissable par l'utilisateur. Le même principe a été appliqué sur les promotions, fournissant un top sur les promotions les plus utilisées.
  Malheureusement, ces « TOP » promotions ne sont pas comparables entre CR, car chaque CR possède ces propres promotions commerciales et donc des libellés différents, qui ne sont pas toujours explicites. Ainsi, même si les promotions couvrent les mêmes offres elles ont des libellés différents dans les bases, et sont donc inexploitables.

- La part des promotions « temps fort »
  Chaque année, en septembre – octobre, il est fait constat d'une hausse de l'usage des promotions. Cette période et ces promotions sont appelées « temps fort ».
  Cet indicateur tend à mesurer pour chaque agence, qu'elle est la part de ces promotions « temps fort » sur une année.

$$\frac{Nombre\ de\ promotions\ sur\ sept.\text{-}oct.}{Nombre\ de\ promotions\ sur\ l'année}$$

- La consommation du budget alloué aux promotions
  Chaque CR dispose d'une enveloppe qu'elle peut gérer comme elle l'entend. Cet indicateur tend à mesurer dans quelle mesure cette enveloppe est respectée ou dépassée.

$$\frac{Montant\ de\ promotion\ réellement\ utilisé}{Montant\ de\ l'enveloppe}$$

  Au-delà de 100% le budget alloué est dépassé, à l'inverse, en deçà, il reste du budget.

3 ... synthétisés dans un tableau de bord

Toutes ces nouvelles variables sont intégrées dans un TDB, et plus exactement dans un outil *Shiny*[15]. Cet outil présente plusieurs onglets, où chacun tend à caractériser un aspect de l'usage de la promotion commerciale :

- l'onglet **Général**, qui présente le taux de transformation, de promotion au global mais également selon la formule, la nature du vendeur, le CRM… ;
- l'onglet **Cartographie**, qui affiche le taux de transformation par CR ;
- l'onglet « **TOP** » qui affiche le « TOP » vendeurs, et « TOP » promotions ;
- l'onglet **Budget**, qui présente le montant moyen d'une promotion, ainsi que la part du budget consommé et si la CR est en dépassement budgétaire ou pas.

Pour chaque graphique, tableau que l'utilisateur veut afficher, il peut choisir l'année mais aussi filtrer sur la maille d'affichage qu'il veut regarder : choisir le réseau (Pacifica, LCL ou les deux), le canal (Agence, SNBAM ou les deux) ou encore la CR.

Pour illustrer ce TDB, une capture d'écran est présentée en Figure IV.1. Même si la maille d'affichage et l'axe des ordonnées sont masqués, on peut voir un taux de transformation stable sur 3 ans. En revanche le taux de promotion montre des « pics » au moment de la période « temps fort ». Un autre exemple de graphique est mis en Annexe C, on y retrouve le taux de transformation et le taux de promotion mais selon le critère multi-équipement.

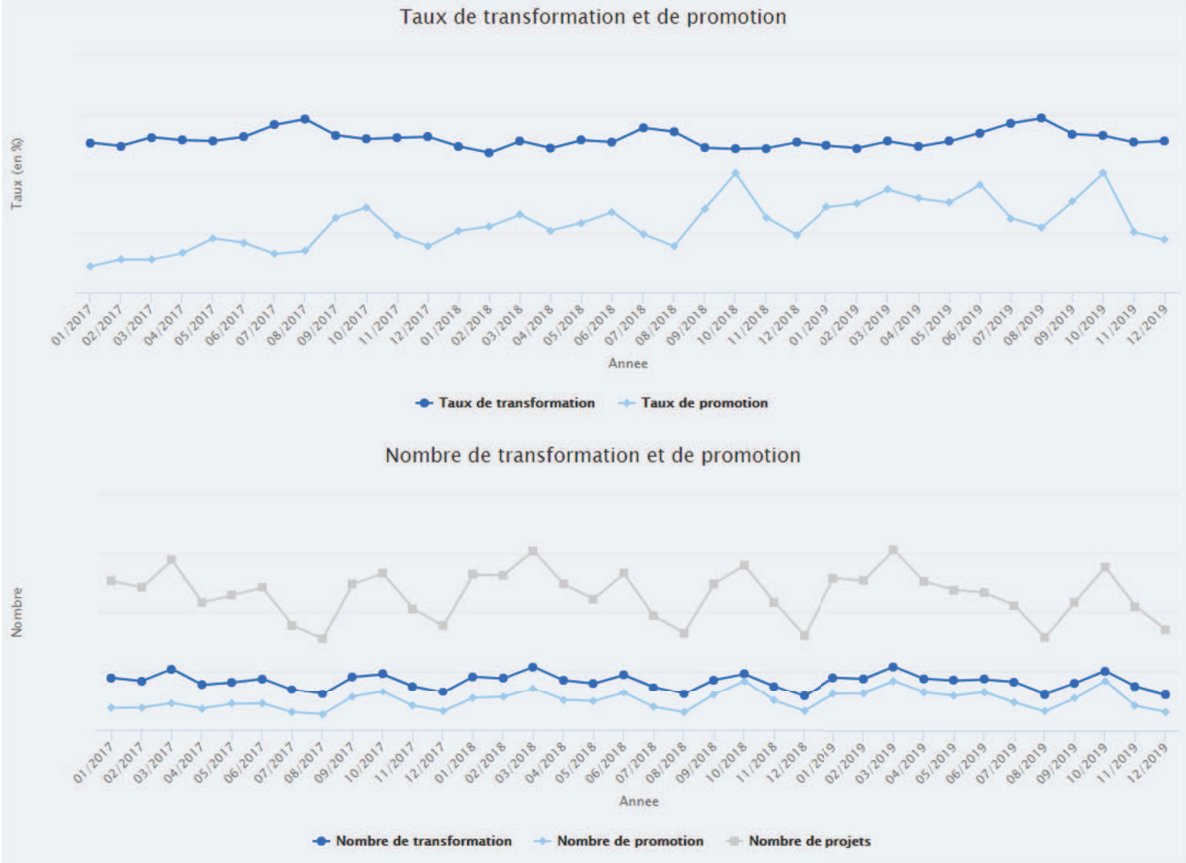


Figure IV.1 – Illustration du TDB : Taux de transformation, taux de promotion et volumétrie des projets

Ainsi, ce TDB permet de caractériser la situation actuelle, à travers différents indicateurs tels que le taux de transformation, le taux de promotion, le montant moyen de promotion ou encore

15. Outil de création d'applications interactives.

la part de dépassement budgétaire. Il nous permet, à nous utilisateur de définir ce qu'est une promotion, mais aussi une typologie comportementale pour chaque CR. Est-ce que cette CR use beaucoup de promotions ? Dépasse-t-elle son budget ? Est-ce que les pics de promotion sont accompagnés d'une hausse de la transformation ? etc

Toutes ces informations sont importantes pour notre étude, cependant il y a autant de typologies comportementales que de CR, soit 40. Étudier une à une chaque CR est impossible car extrêmement chronophage et difficilement pérennisable. Pour réduire ce nombre de typologies, un *clustering* et une ACP (Analyse en Composantes Principales) ont été mis en oeuvre afin de segmenter les CR selon leurs pratiques promotionnelles et identifier 2–3 profils de stratégies commerciales.

4 Segmentation des caisses régionales (CR)

Cette segmentation des CR permet de diviser les CR selon leurs pratiques commerciales. Pour cela, un *clustering* associé à une ACP sont réalisés.

Cette ACP, repose sur 6 variables :

- *tx_transfo*: le taux de transformation.
- *tx_promo*: le taux de promotion.
- *tx_promo_forf* : le taux de promotion « forfaitaire ». Parmi les projets avec une promotion, combien sont de type « forfait » (i.e 10€ de réduction).
- *promo_moy*: la valeur de la promotion moyenne.
- *part_dep_budg2*: la consommation du budget promotionnel. Cette variable calcule la part du budget consommé. Si la part vaut 0,8 alors la CR a utilisé 80% de son enveloppe, à l'inverse une valeur de 1,56 désigne une CR qui a dépassé son budget, et cela, de 56%.
- *ecart_promo_0910_rel*: l'écart relatif entre le taux de promotion sept.– oct. et celui de l'année. Un écart important correspond à une CR qui présente une hausse de l'utilisation des promotions en cette période de rentrée. Cette période est appelée « temps fort ».

Cette sélection de variables repose sur une connaissance métier, ainsi que sur les observations réalisées à partir du TDB et tend à compléter la recherche de typologie comportementale en répondant à certaines questions telles que : de quelle manière est utilisée la promotion ? A quel moment ?

Afin de s'assurer de la stabilité des résultats d'une année à l'autre, et de mesurer si d'une année à l'autre les pratiques peuvent être changeantes, on réalise une ACP par année, soit 3. Pour éviter la redondance, et par stabilité des résultats d'une année à l'autre, seuls les résultats de 2019 seront présentés.

4.1 Détermination du nombre optimal de *cluster*

Afin de connaître le nombre idéal de *clusters* à réaliser, une Classification Ascendante Hiérarchique (CAH) et la méthode des *k-means* sont réalisées.
La CAH aurait été suffisante pour connaître le nombre de *clusters* à réaliser, cependant le *k-means* permet de confirmer ou d'infirmer le résultat donné par la CAH.

Quel que soit l'année d'étude, le résultat du dendrogramme est le même, à savoir que 3 *clusters* sont suffisants pour segmenter nos CR. Ces 3 segments sont visibles sur le dendrogramme réalisé sur les données de l'année 2019, présentés en Figure IV.2. Les dendrogrammes de 2017 et 2018, concluent à la même information.

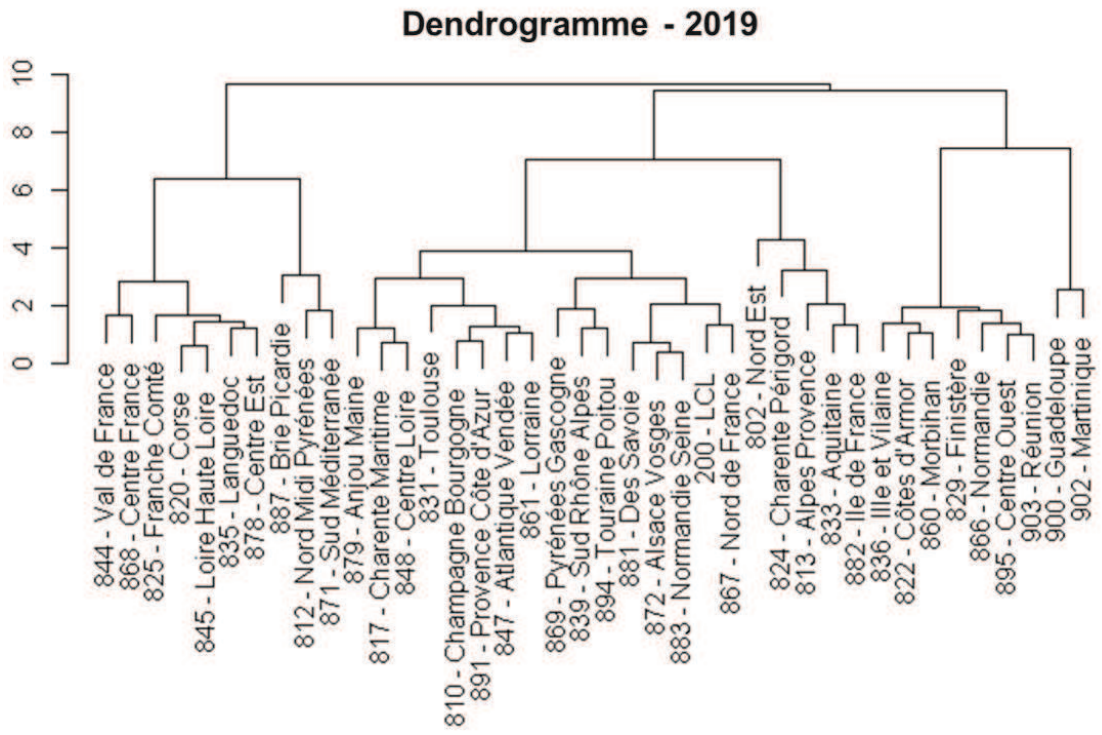


Figure IV.2 – Dendrogramme pour l'année 2019

Ces 3 *cluster* minimisent la variance *intraclasse* et maximisent celle *interclasse*. Ainsi, chaque individu appartenant à un même *cluster* se ressemblent, mais pas entre individus de *clusters* différents. Les résultats de la méthode des k-means sont affichés en Annexe D et donnent le même résultat, à savoir 3 *clusters*.

4.2 Étude à l'aide d'une analyse en composantes principales

Maintenant que le nombre de *clusters* optimal est connu, comment qualifier ces 3 *clusters*? Quelles stratégies commerciales les caractérisent ?
Dans le but de répondre à ces questions une ACP est réalisée, afin d'obtenir des variables orthogonales résumant au mieux l'information contenue dans nos 6 variables initiales.

4.2.1 Détermination du nombre d'axes à retenir

Le choix de l'ACP est légitime car il s'agit de 6 variables quantitatives, plus ou moins corrélées entre elles, comme l'illustre le corrélogramme en Figure IV.3.

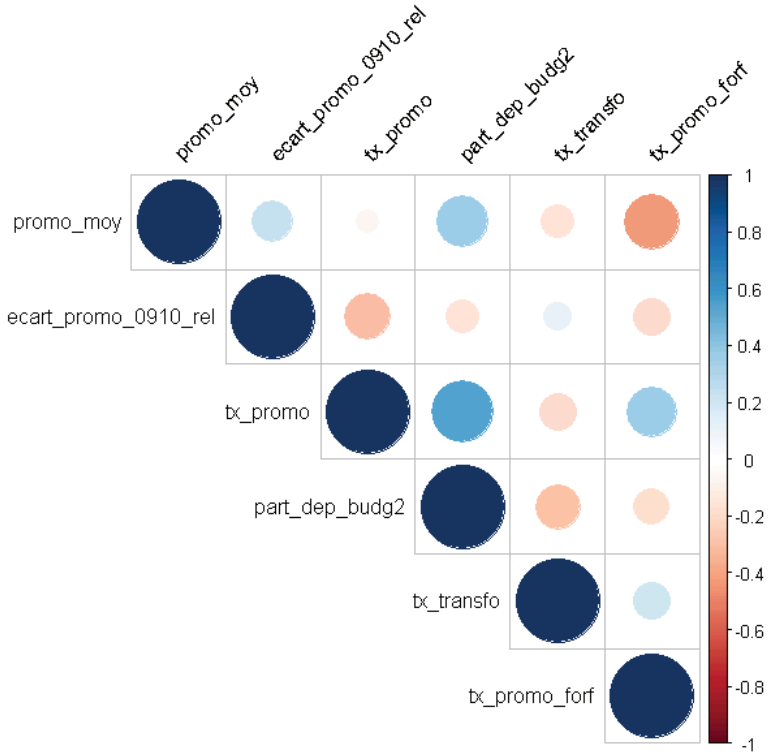


Figure IV.3 – Corrélogramme des variables pour l'année 2019

Toute l'information contenue dans les variables est à conserver, tout en limitant la redondance d'information liée à la corrélation des variables entre elles. C'est là tout l'intérêt d'une ACP, conserver un maximum d'inertie (c'est-à-dire un maximum d'informations et de dispersion) avec un minimum de facteurs.
Pour déterminer le nombre d'axes à conserver, on peut représenter au sein d'un histogramme, le pourcentage d'inertie associé à chaque axe et exploiter deux critères : le critère du coude et le critère de Kaiser.

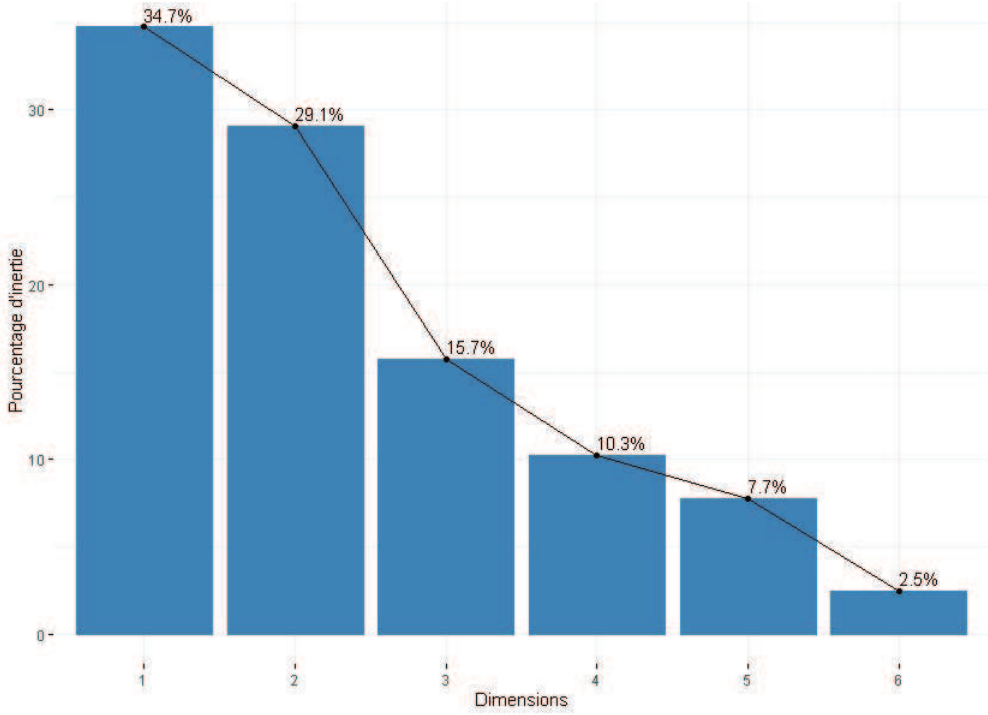


Figure IV.4 – Histogramme des valeurs propres pour l'année 2019

Les résultats obtenus sur l'année 2019 (cf. Figure IV.4) laissent apparaître un coude (i.e. décrochement avant une décroissance régulière) entre le 2eme et 3eme axe. D'après le critère du coude, cela sous-entend que les deux premiers axes sont à conserver.
Pour valider cette hypothèse, regarder le critère de Kaiser est possible. Ce dernier suggère de conserver les axes dont la valeur propre est supérieure à l'inertie moyenne, c'est-à-dire 1 sur le nombre d'axes, soit 17%. En ACP normé, on ne retiendrait que les axes dont les valeurs propres sont supérieures à 1, cela indique que la composante principale concernée représente plus de variance par rapport à une seule variable d'origine (lorsque les données sont standardisées).
Dans notre cas, seuls les deux premiers axes présentent une valeur propre supérieure à 17%. Ainsi, le critère du coude et celui de Kaiser recommandent tous deux de choisir 2 axes.

Au vu de cette analyse, nous décidons donc de conserver les deux premières composantes principales qui expliquent 63,8% de l'information contenue dans la base de données.
Maintenant que le nombre d'axes à retenir est connu, il nous faut les interpréter.

4.2.2 Interprétation des axes

L'interprétation des axes issus de l'ACP peut parfois être difficile, cependant, il existe plusieurs indicateurs permettant de les expliquer. Le cercle de corrélation des variables montre les relations entre toutes les variables. Plus la distance entre la variable et l'origine est élevée, meilleure est la qualité de représentation de la variable par l'ACP.
Cependant, en cas de nombreuses variables, le cercle de corrélation peut devenir illisible et compliquer d'interprétation. Examiner la contribution relative des variables à l'inertie d'un axe devient alors plus intéressant.

En étudiant le cercle de corrélation et la contribution relative des variables présentée en Figure IV.5, il apparaît que les variables qui participent le plus au premier axe (*Dim1*) sont le taux de promotion (*tx_promo*), la part de dépassement du budget (*part_dep_budg2*) et l'écart de promotion en septembre/octobre (*ecart_promo_0910_rel*). Leur interprétation se fait par rapport à l'axe des abscisses. Les deux premières variables sont associées positivement au premier axe, c'est-à-dire que plus on se décale sur la droite, plus les variables prendront des valeurs élevées. A l'inverse, pour l'écart de promotion en septembre/octobre, la corrélation est négative, les valeurs importantes seront rencontrées sur la gauche de l'axe.
Concernant le deuxième axe (*Dim2*), les variables qui contribuent le plus sont le montant de promotion moyen (*promo_moy*) et le taux de promotion forfaitaire (*tx_promo_forf*). Plus on se déplace positivement sur l'axe des ordonnées, plus le montant de promotion sera important et le taux de promotion forfaitaire baissera.

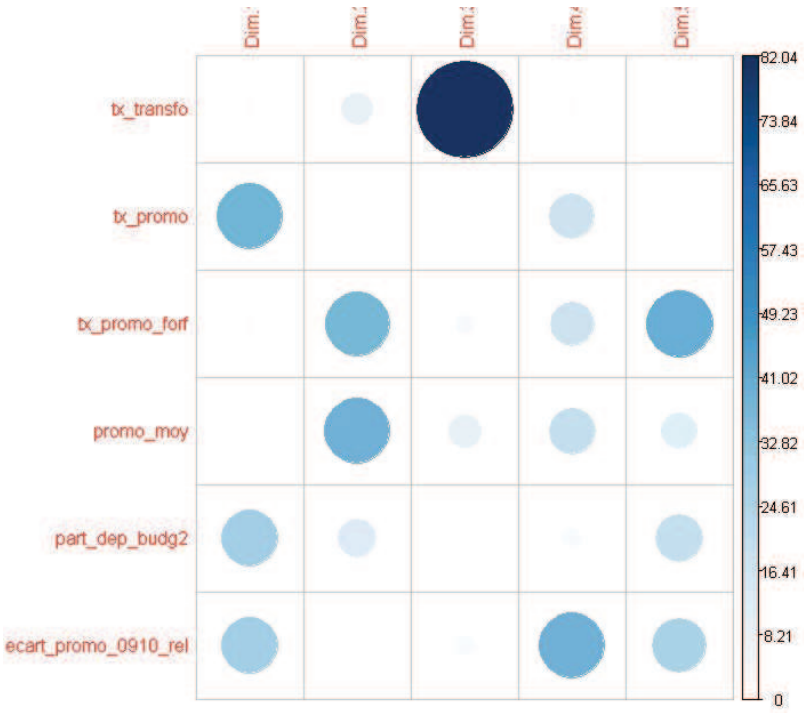


Figure IV.5 – Contributions relatives des variables aux axes pour l'année 2019

Enfin, pour connaître la qualité de représentation des variables sur le graphe de l'ACP, on peut regarder le *cos2*. Les résultats présentés en Figure IV.6, indiquent que les variables taux de promotion (*tx_promo*), la part de dépassement du budget (*part_dep_budg2*) et l'écart de promotion en septembre/octobre (*ecart_promo_0910_rel*) sont bien représentées par le premier axe, et de même pour le montant de promotion moyen (*promo_moy*) et le taux de promotion forfaitaire (*tx_promo_forf*) sur le second axe.

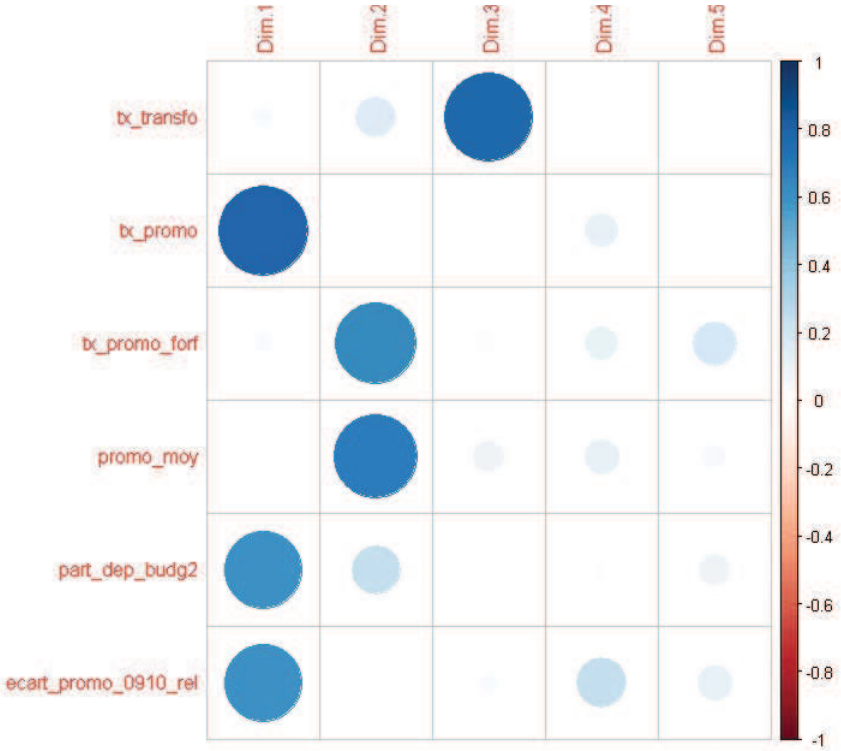


Figure IV.6 – Qualité de représentation des variables pour l'année 2019

Ainsi, nous pouvons conclure que les variables sont bien représentées par les axes et que le premier et second axe sont respectivement caractérisées par : le taux de promotion, la part de dépassement du budget et l'écart de promotion en septembre/octobre pour le premier, et par le montant de promotion moyen et le taux de promotion forfaitaire pour le second.
Néanmoins le taux de transformation (*tx_transfo*) est peu représenté par ces 2 axes, l'interprétation de cette variable sera donc à nuancer malgré son importance.

4.3 Identification de 3 typologies comportementales

L'analyse en composantes principales a permis de réduire le nombre de dimensions, en passant d'un espace de dimensions 6 à un espace de dimensions 2, et surtout de caractériser les 3 groupes obtenus par *clustering*.
En associant les résultats de l'ACP et ceux du *clustering* on obtient le graphe présenté en Annexe E qui superpose le graphique des variables et celui des individus colorés selon leur *cluster*.

Grâce à cette superposition et en exploitant le TDB, on identifie 3 grands *cluster* de CR caractérisables :

1. Les CR utilisant beaucoup de promotions, associé à un risque de dépassement du budget, ainsi qu'un montant de promotion moyen plus élevé que la moyenne nationale. De plus, il semblerait que la hausse de promotion n'est pas d'effet sur le taux de transformation.
2. Les CR utilisant peu de promotions, sauf en période « temps fort », avec un montant de promotion moyen plus faible que la moyenne nationale, et avec un taux de transformation « élevé ».
3. Les CR utilisant beaucoup de promotion de type « forfait », qui se traduit par un montant de promotion moyen plus faible que la moyenne nationale et un « bon » taux de transformation.

Avec ces deux approches, le TDB et l'ACP, 3 typologies comportementales caractérisées chacune par une stratégie commerciale se dégagent, permettant ainsi de caractériser la situation actuelle dans l'utilisation des promotions commerciales.
Cette étape de compréhension est nécessaire pour la modélisation du taux de transformation et mesurer l'impact des promotions commerciales.
Cependant, nous aurions peut-être aimé avoir plus de groupes afin d'affiner plus nos typologies comportementales. En effet, les 3 groupes précédemment définis restent larges, notamment le *cluster* n°1 qui semble rassembler un grand nombre de CR. Ils nous permettent tout de même d'avoir une première idée des stratégies commerciales, même s'ils restent trop larges.

5 Intégrations des *clusters* à notre base de données

Afin d'intégrer ces résultats à la modélisation du taux de transformation, un nouveau *clustering* et une nouvelle analyse en composantes principales ont été réalisé sur les données correspondant à la moyenne des 3 années.

Là aussi, le nombre de *clusters* à faire est de 3, et on retrouve les 3 mêmes grands groupes de CR (cf. Figure IV.7).

- Le *cluster* n°1, regroupant des CR utilisant beaucoup de promotions (dépassement du budget), un montant moyen de promotion élevé et un taux de transformation peu élevé.
- Le *cluster* n°2, avec des CR utilisant peu de promotions, sauf en période « temps fort », et avec un taux de transformation élevé.
- Le *cluster* n°3, contenant des CR utilisant beaucoup de promotion de type « forfait », avec un montant moyen de promotions bas et un « bon » taux de transformation.

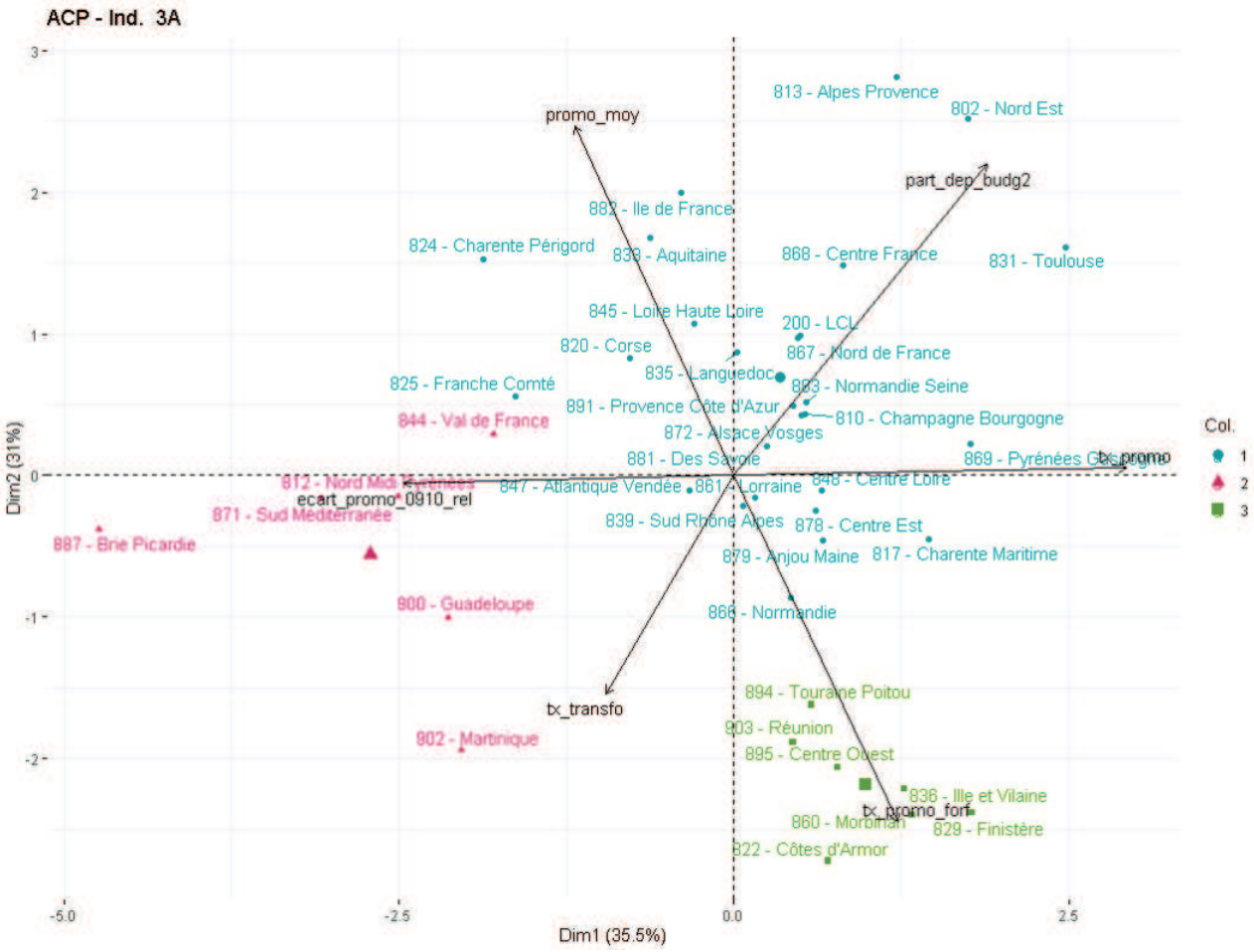


Figure IV.7 – Graphe des variables et des individus (moyenne des 3 années)

Ces résultats sont intégrés dans la variable *cluster3*, qui contient le segment auquel appartient chaque CR.

Chapitre V

Modélisation du taux de transformation et identification de critères optimaux

Dans cette partie, les résultats des différentes méthodes mises en œuvre dans ce mémoire sont présentés. Nous utiliserons les arbres de régression et les méthodes d'agrégation : *Gradient Boosting Machine* (GBM), *Random Forest*, ainsi que les modèles linéaires généralisés (GLM) et la régression pénalisée. Ces premiers vont nous permettre d'identifier les variables qui sont importantes, tandis que les GLM offrent l'avantage d'être facilement interprétables. Enfin, la régression pénalisée permet d'étudier les interactions entre variables et ainsi cibler les clients sensibles aux promotions.

L'une des contraintes des GLM, concerne la nécessité de disposer d'un échantillon de taille n suffisamment importante, et surtout que le nombre p de variables explicatives ne soit pas « trop » grand.
Pour pallier à cela, on effectue une sélection de variables grâce au GBM, avant de modéliser le taux de transformation par GLM. Puis on étudie les interactions à l'aide d'une régression pénalisée. Cette régression pénalisée doit permettre de mettre en évidence des interactions entre la présence d'une promotion et des caractéristiques clients (âge du conducteur, formule souscrite...), qui impactent de façon positive ou négative le taux de transformation.

1 Modélisation du taux de transformation par caisses régionales (CR)

1.1 Sélection de variables par *Gradient Boosting*

Le GBM est une approche de type *Machine Learning*, alternative au modèle GLM, et qui a pour objectif de modéliser le taux de transformation ainsi que de procéder à une sélection de variables (nécessaire pour notre GLM).

La mise en place d'un algorithme *Gradient Boosting* requiert le choix de paramètres optimaux dans le but de minimiser l'erreur.
Les paramètres à choisir sont extrêmement nombreux et peuvent interagir entre eux tout comme impacter la performance. Un arbitrage entre exploiter l'ensemble des données disponibles et le sur-apprentissage est donc nécessaire.

Il faut donc arbitrer entre :

- La profondeur des arbres, *max_depth*
  Un arbre petit sera plus résistant au sur-apprentissage, mais plus sensible au sous-apprentissage. C'est l'inverse pour les arbres dits grands.

- Le nombre d'arbres, *ntrees*
  Le risque de sur-apprentissage diminue avec le nombre d'arbres, cependant plus il y en a, plus le temps de calcul augmente.

- Le taux d'apprentissage, ν
  Une valeur trop faible sera associée à une lenteur de convergence, à l'inverse trop élevée, risque d'oscillations et de sur-apprentissage. Si la valeur diminue, le nombre d'arbres doit être augmenté.

- Le taux d'échantillonnage des individus, β (*Stochastic gradient boosting*)
  Pour une valeur de 1, l'algorithme utilise toutes les observations ; une valeur inférieure réduit la dépendance de l'échantillon, et résistance au sur-apprentissage.

- Le nombre minimum d'observations par nœud terminaux, *min_rows*.

L'optimisation des paramètres s'est fait sous *R*, en utilisant la recherche par quadrillage (*Grid Search*). Le principe de cette méthode est d'initialiser une grille contenant tous les paramètres du modèle à optimiser, et une liste de valeurs possibles pour chacun d'eux. Pour chaque combinaison, un modèle est entraîné et son score calculé. A la fin, seul les paramètres du meilleur modèle sont conservés et utilisés dans le modèle final, qu'on valide par *cross validation* (CV).
Sur le graphe de gauche de la Figure V.1, on peut voir que la *LogLoss* se stabilise à partir de quelques centaines d'arbres. Mais pour s'assurer de la stabilité du modèle, nous retenons $ntrees = 1500$.
Le graphe des variables importantes (graphique à droite de la Figure V.1), liste les variables ayant le plus d'influence dans la construction des arbres. Cette énumération permet de sélectionner les variables.

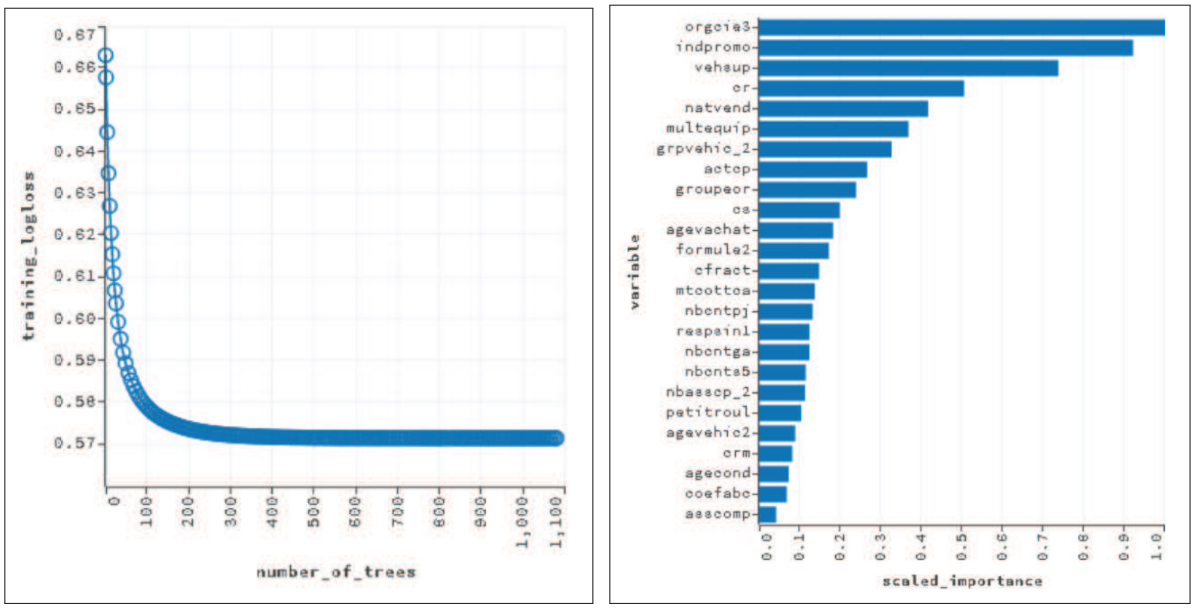


Figure V.1 – *Erreur LogLoss en fonction du nombre d'arbres et importances des variables*

Une variable apparaît comme très nettement significative, celle correspondant à l'ancienne compagnie d'assurance du prospect (*orgcie3*). Elle est suivie en matière d'importance par la

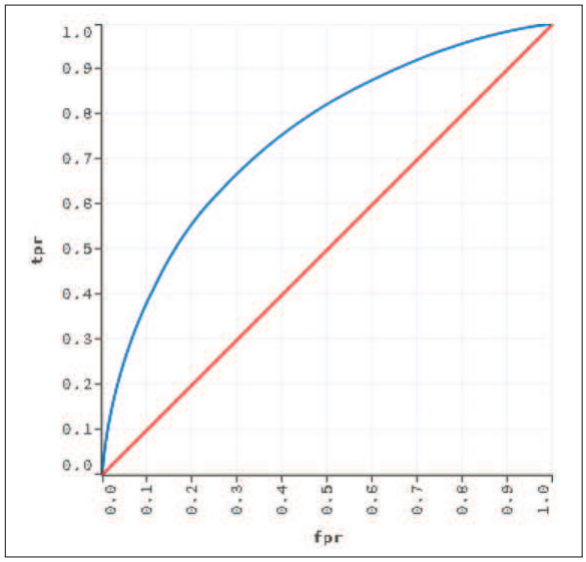
variable indicatrice de présence d'une promotion (*indpromo*), la variable « qualité » du véhicule (*vehsup*) et le numéro de la caisse régionale (*cr*).

La présence de la variable *indpromo* est rassurante, et légitime l'idée que la promotion impacte la transformation, reste à savoir dans quel sens. La troisième variable *vehsup*, apporte une information supplémentaire pour caractériser le taux de transformation. En effet, cette variable présente 4 modalités renseignant sur la « qualité » du véhicule : si le véhicule est déjà en possession et assuré (*vehsup = 1*), s'il s'agit d'un véhicule de remplacement (*vehsup = 2*), si le véhicule est acheté en supplément d'un autre (*vehsup = 3*) ou s'il s'agit d'un premier véhicule (*vehsup = 4*). Dans le premier cas, on est en conquête « pure », il s'agit du « vrai » taux de transformation. Le prospect est déjà assuré auprès d'un concurrent, on ne répond donc pas à un besoin comme dans le deuxième et dernier cas, qui correspondent eux aussi à de la conquête mais plus accessible. Enfin pour le troisième cas, il s'agit d'un prospect déjà en portefeuille, la conquête peut être plus aisée.

Avant de pouvoir exploiter les résultats de notre GBM, à savoir utiliser la sélection de variables dans notre GLM, il nous faut vérifier la qualité du modèle.

1.2 Qualité du modèle

L'évaluation de la performance du modèle garantie son pouvoir prédictif. Pour se faire, trois indicateurs de performance ont été exploités : la courbe ROC et l'AUC (l'aire sous la courbe ROC), ainsi que la *precision* et le *recall* présentés en Figure V.2. Plus ces indicateurs sont importants, meilleur est le modèle. Cependant, ils sont à regarder ensemble, et trouver un équilibre peut parfois être nécessaire.
Les métriques obtenues sur l'échantillon d'apprentissage et l'échantillon de test sont comparables, cela s'explique par le fait que la prédiction faite à partir de l'échantillon de test reflète bien celle obtenue lors de la phase d'apprentissage.



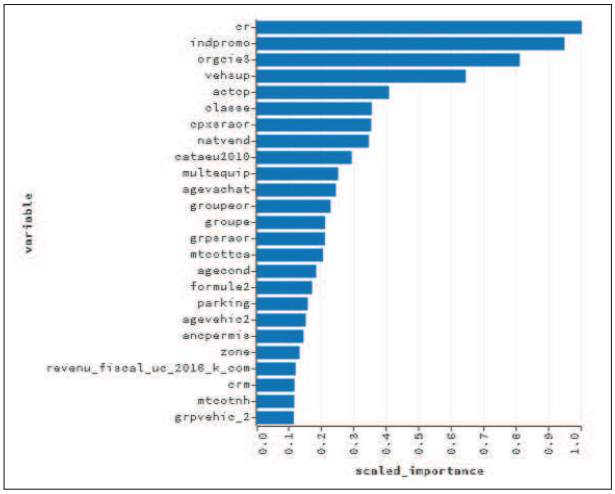
AUC	Recall	Precision
0,749	0,765	0,530

Figure V.2 – *Mesure de précision : Courbe ROC, AUC, recall et precision du modèle GBM*

La *precision* obtenue est de 0,53, ce qui n'est pas très élevée, et peut se traduire de la manière suivante : 47% (1 - 53%) des projets identifiés comme une future affaire nouvelle (AN) sont des erreurs, ce qui semble important. En revanche, si on regarde le *recall*, on est rassuré, car il indique que 77% des projets qui seront transformés sont identifiés par le modèle. Et ce sont ces projets que l'on vise, même s'il ne faut pas oublier la *precision*. Enfin l'AUC obtenu est de 0,75 et la F_{mesure} vaut 0,62 et rend compte de la qualité de la classification en fonction des classes.

Toutes ces mesures permettent d'attester de la qualité du modèle, et d'accepter ce dernier.

Pour challenger et valider la sélection de variables par GBM, un modèle *Random Forest* (RF) est mis en place. Les variables les plus significatives, ainsi que les indicateurs de performance associés sont présentés ci-dessous en Figure V.3.



AUC	Recall	Precision
0,743	0,760	0,528

Figure V.3 – *Variables importantes et mesure de précision : AUC, recall et precision du modèle RF*

D'après les différents indicateurs sélectionnés pour jauger de la qualité du modèle, les modèles GBM et RF sont très proches en termes de qualité. En revanche, quelques différences existent en matière de sélection de variables. Même si certaines variables sont identifiées comme significatives par les deux modèles, leur ordre d'apparition et donc leur importance varie. C'est sur cela que nous basons notre sélection de modèle, en préférant le modèle qui se rapproche le plus des connaissances métier et des dires d'expert, ainsi nous favorisons les résultats du GBM.
Pour confirmer cette analyse, on peut calculer les écarts dans la précision prédictive des modèles. En comparant la probabilité d'attrition à la valeur réelle binaire (Oui / Non), les résidus.

Si on regarde la Figure V.4, on peut voir sur les boîtes à moustache (graphique de droite) que les résidus absolus médians sont légèrement plus bas pour le modèle RF.
Cependant le modèle RF présente un nombre plus élevé de résidus dans la queue de la distribution résiduelle (graphe de gauche) ce qui peut laisser à penser qu'il peut y avoir un nombre plus élevé de grands résidus par rapport au modèle GBM.
Ainsi, même si le modèle GBM est moins performant lorsque l'on considère les résidus absolus, il reste assez proche du modèle RF et est plus performant en queue de distribution. On reste donc sur la sélection de variables fournie par le GBM.

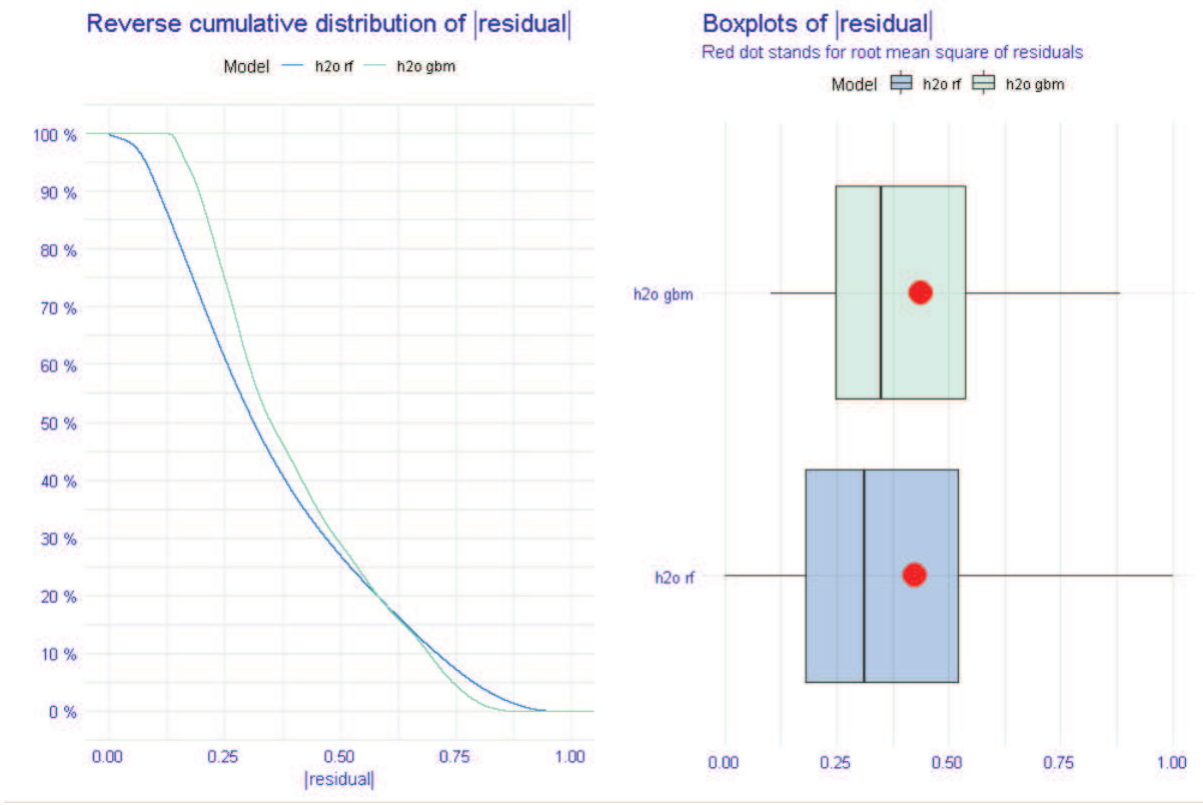


Figure V.4 – *Distribution des résidus en valeur absolue (à gauche), et boîte à moustache des résidus en valeurs absolues (à droite)*

1.3 Interprétations globale et locale

L'une des contraintes des modèles tel que le GBM, ou le *Random Forest* c'est leur effet « boîte noire », qui limite la compréhension des résultats. Différentes techniques de *Machine Learning* existent pour pallier à ce manque de transparence et connaître dans quelle mesure les variables identifiées comme « importantes » interviennent dans le modèle. Ainsi, pour avoir une première lecture de l'influence des variables dans le modèle, on peut tracer les graphes de dépendance partielle et les valeurs de Shapley .

1.3.1 Le graphe de dépendance partielle pour une interprétation globale

Le graphe de dépendance partielle montre l'effet marginal d'une variable sur la prédiction de la variable d'intérêt, dans notre cas, la variable indicatrice transformation (Oui/Non). Il permet de visualiser l'impact d'une variable en fonction des différentes valeurs qu'elle peut prendre en moyennant l'influence de toutes les autres variables.
Un autre avantage des graphes de dépendance partielle, est qu'ils permettent de vérifier la cohérence du modèle. En confrontant les résultats du modèle aux connaissances métiers, on peut vérifier leur pertinence. Par exemple, si dans une compagnie d'assurances il est connu que les formules « d'entrée de gamme » transforment mieux que les formules dites « haut de gamme », un graphe de dépendance partielle sous-entendant l'inverse remettrait en cause la cohérence du modèle.

Les graphes de dépendance partielle de certaines variables identifiées par le GBM : l'indicatrice de la présence d'une promotion (*indpromo*), la « qualité » du véhicule (*vehsup*) et l'âge du véhicule à l'achat (*agevachat*) sont présentés en Figure V.5 (la DP correspond au trait bleu). Sur ces mêmes graphes, les courbes d'espérance conditionnelle individuelle (*ICE, Individual Conditionnal Expectations*) sont tracées. Chaque tracé correspond à une observation, cette dernière varie en fonction de la valeur de la variable qu'on souhaite étudier et le trait bleu correspond à la moyenne de toutes ces observations, la dépendance partielle.

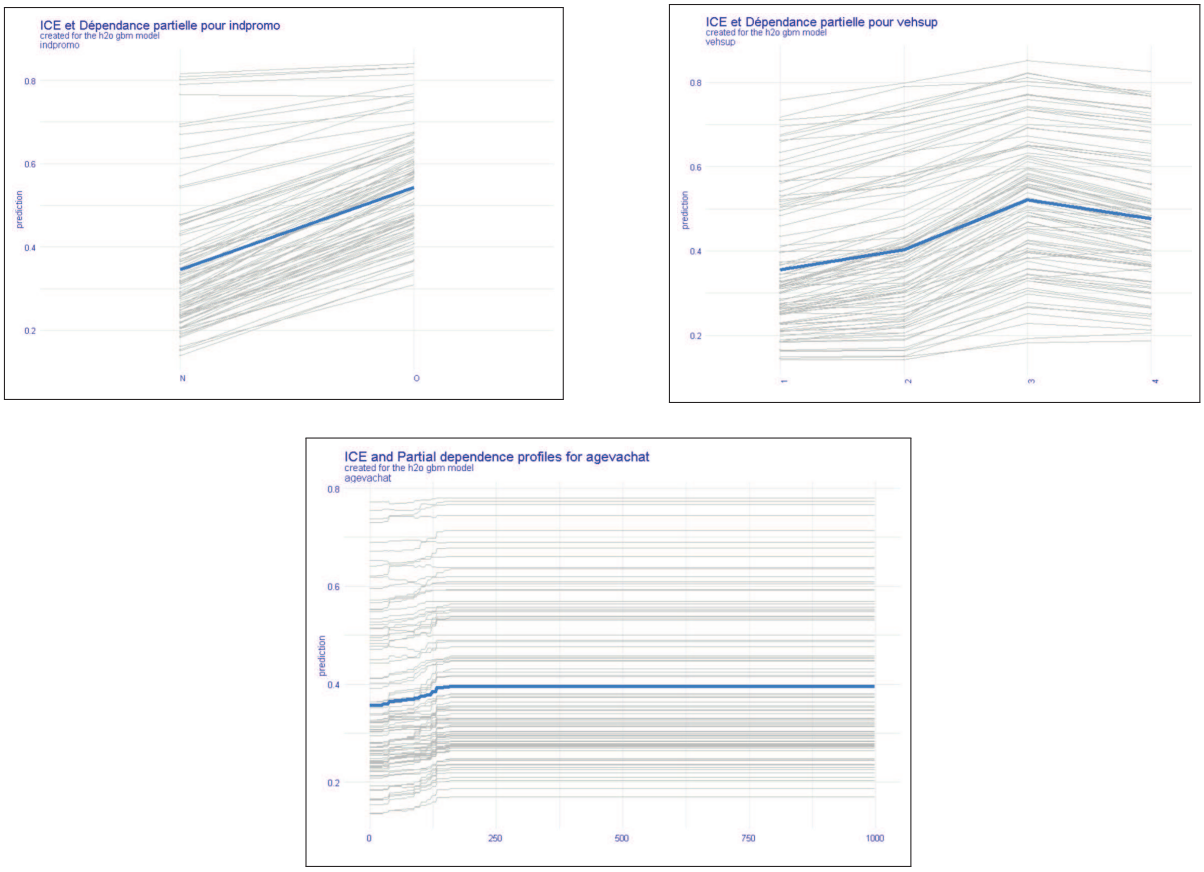


Figure V.5 – *Graphes de dépendance partielle et ICE des variables indpromo (en haut à gauche), vehsup (en haut à droite) et agevachat (en bas)*

D'après le graphique en haut à droite, les projets avec une promotion ont une plus grande chance d'être transformé. Cette conclusion rassure et légitime une nouvelle fois notre étude car une information contraire aurait remis en question l'idée d'étudier les promotions commerciales et la démarche d'optimisation de leur allocation, si ces dernières n'entraient pas en jeu dans l'acte de vente d'un contrat.
Pour les deux graphiques du bas, on peut voir que les projets couvrant des véhicules achetés en supplément d'un premier véhicule déjà assuré (modalité 3) ont également plus de chance d'être transformés. Et les clients semblent plus susceptibles de transformer avec l'âge (valeur élevée d'*agecond*).

1.3.2 Une interprétation locale grâce aux valeurs de Shapley *(SHAP Values*

Les valeurs de Shapley constituent une seconde approche pour interpréter la prédiction d'un modèle. Pour un individu donné, les valeurs de Shapley permettent de connaître la contribution de chaque variable à la prédiction, on parle alors d'interprétation locale.
Ainsi, il est possible de connaître pour un individu, si sa prédiction est inférieure (respectivement supérieure) à la prédiction moyenne et surtout qu'elles sont les variables qui ont influencé la probabilité de prédiction vers 0 (resp. 1).
On regarde une interprétation sur la Figure V.6 : un individu présentant une probabilité élevée de transformer son projet.

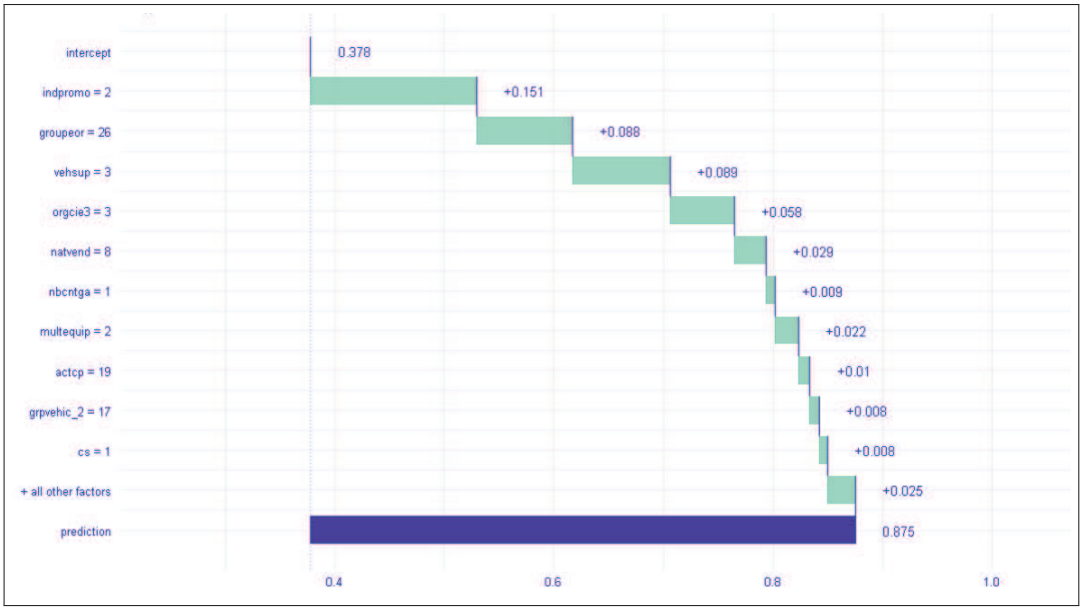


Figure V.6 – *Contributions de chaque variable dans la prédiction du modèle GBM pour un individu*

Une fois encore, la présence d'une promotion (*indpromo*) impacte la probabilité de transformer, elle contribue à presque 15 points dans sa prédiction. De même que la présence d'un conducteur secondaire (*cs=1*) et le multi-équipement contribueront positivement à la prédiction de l'assuré.

Enfin, ce graphe d'interprétation locale (car chaque observation obtient son propre ensemble de valeurs) peut être généralisé à l'ensemble des individus fournissant ainsi une explication des variables, et du sens à leur contribution à la variable cible, l'indicatrice transformation (Oui/Non) (*nb_transfo*). Cependant, le coût et le temps de calcul des valeurs de Shapley dans une interprétation globale sont importants, et cela, même sur un échantillon de taille relativement faible. De fait, cette fonctionnalité des valeurs de Shapley n'est pas mise en place dans cette étude, même si elle aurait apporté la contribution de chaque prédicteur à la variable cible et conforter une nouvelle fois la sélection des variables.

1.4 Modèle linéaire généralisé

La modélisation GLM constitue un élément majeur dans l'étude du taux de transformation. Nous utilisons cette technique pour modéliser l'appétence d'un client à la transformation, information non-négligeable dans le cas d'une refonte de produit, et plus spécifiquement en segmentation tarifaire.

La sélection de variables réalisée en premier volet de ce chapitre a permis de procéder à une première sélection des variables jugées comme étant les plus influentes. On peut entre autres citer : l'indicatrice de promotion (*indpromo*), la « qualité » du véhicule (*vehsup*), l'indicatrice de multi-équipement (*multequip*), ou encore l'âge du conducteur (*agecond*).
Ces variables sont intégrées à notre modèle GLM, avant ajustement du modèle. Pour l'interprétation, nous interprétons avec le lien *log* avant de repasser au *logit* pour la prédiction, par ailleurs les axes ont été masqués sur les graphiques présentés dans cette partie par souci de confidentialité.

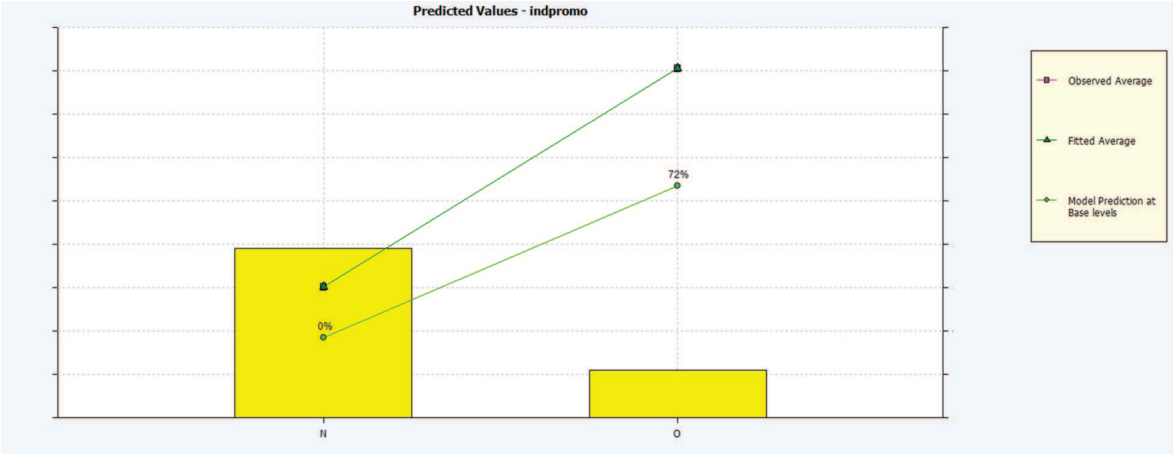


Figure V.7 – Modélisation du taux de transformation - Valeurs prédites *indpromo* - GLM

Après vérification de la qualité de l'ajustement, l'une des premières variables à regarder, est une nouvelle fois *indpromo*. En regardant sa prédiction (cf. Figure V.7), on peut voir qu'elle confirme le fait que les promotions commerciales jouent bien un rôle dans la transformation.

Par ailleurs, le modèle prédit bien la variable *indpromo*, car on capte bien la structure (= composition) de chaque modalité. Et même si modéliser un problème de classification avec la fonction de lien logarithme est inapproprié car incorrect d'un point de vue théorique. En pratique, elle permet de pouvoir faire une première lecture, en donnant un aperçu de l'effet pur par construction d'une variable.
Ici, l'écart relatif lié à la variable *indpromo* prenant la valeur « Oui », vaut $exp(\beta_{indpromo=Oui})-1$, ce qui revient à l'effet pur de la variable.
Ainsi, la présence d'une promotion (*indpromo* = « Oui ») augmente de 72% le taux de transformation.

Toujours sur l'étude de la promotion et pour confirmer une de nos hypothèses de départ, à savoir, qu'un usage excessif des promotions amoindri leur effet, on peut regarder le taux de promotion (*tx_promo*). Cette hypothèse est vérifiée, car on constate une pénalisation de la fréquence d'utilisation de la promotion, +10 points d'usage engendre un effet pur de -6% (cf Annexe F).

Cette lecture est possible pour chaque variable du modèle, et a permis une première interprétation des variables et du rôle qu'elles jouent dans la modélisation du taux de transformation. Sur la Figure V.8 ci-dessous, on présente deux autres variables : le montant de la cotisation TTC annuelle (*mtcottca*) et la « qualité » du véhicule (*vehsup*).
Pour cette première variable, on observe une sensibilité au prix stable sur les premiers prix, qui décroît à partir de P1€ (-1% tous les x€ jusqu'à P2€ ; avant de s'accentuer, -2% sur les cotisations élevées ($\geq$ P2€)). Tandis que pour *vehsup*, on constate une meilleure transformation sur les véhicules achetés en supplément d'un autre (modalité *3*), et une transformation moindre sur les véhicules déjà en possession et assurés (modalité *1*). Derrière cette interprétation se cache la notion de conquête, il est généralement plus difficile de faire transformer un nouveau client (modalité *1*) qu'un client déjà en portefeuille et qui vient couvrir un véhicule supplémentaire.

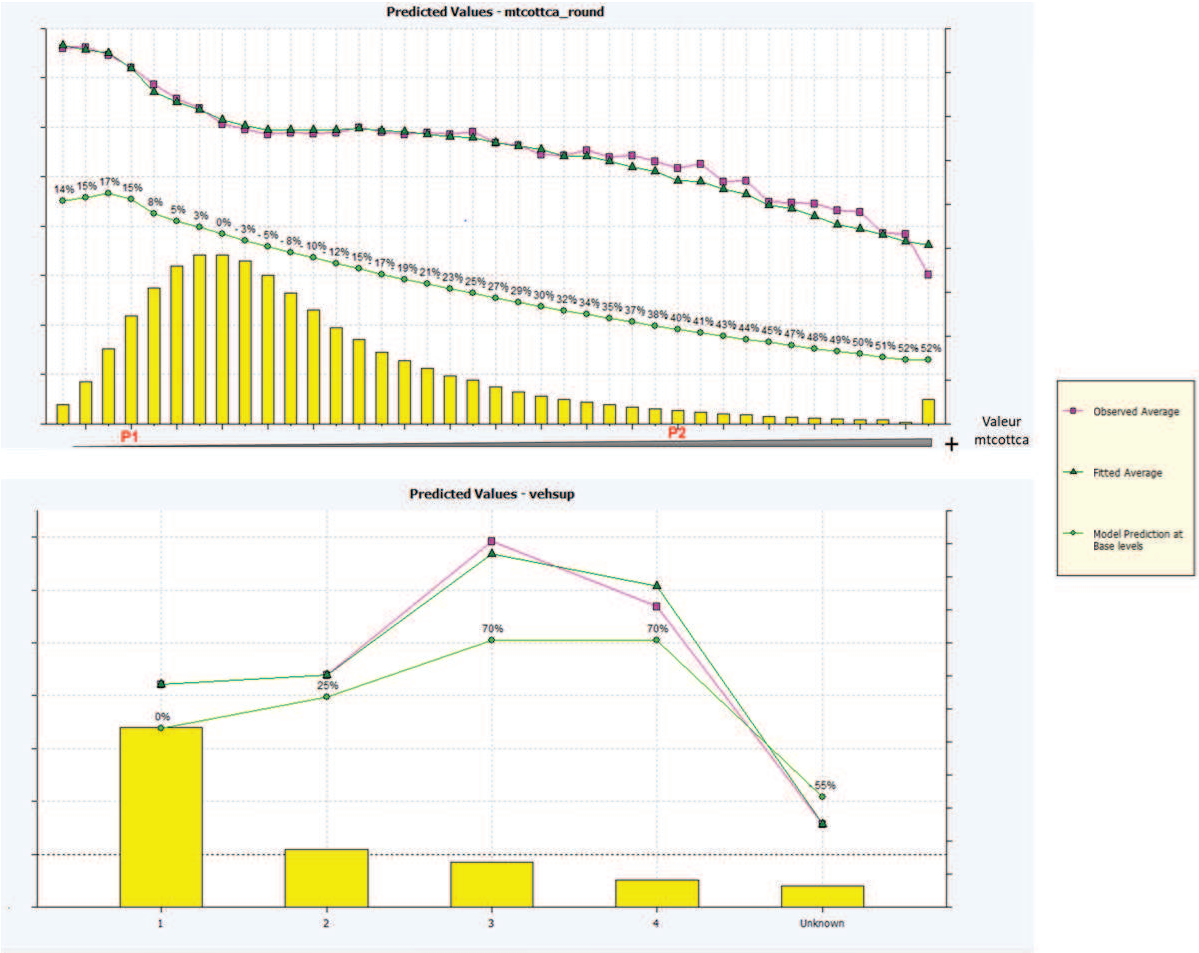


Figure V.8 – Modélisation du taux de transformation - Valeurs prédites des variables *mtcottca* et *vehsup* - GLM

Cependant malgré le fait que l'on puisse modéliser le taux de transformation avec toutes les variables sélectionnées (post GBM). Le GLM présente des limites dont la non-intégration des interactions statistiques par défaut, ces dernières sont à spécifier manuellement. Pour identifier ces interactions, on a utilisé la régression pénalisée avec la méthode *Lasso*.

1.5 *Lasso*: une méthode pour identifier les interactions

Cette méthode nous permet de sélectionner un sous-ensemble des variables explicatives à partir d'un ensemble de grande taille. Notre ensemble de départ contient toutes les interactions possibles avec l'indicatrice de la promotion commerciale et la cotisation, et nous n'en retiendrons que les plus significatives.
Cette sélection de variables va nous permettre de valider celle issue du GBM et d'identifier les interactions.
Si l'interaction entre deux variables (dont la variable indicatrice de la promotion) augmente la probabilité de transformation, alors il est probable que la seconde variable composant l'interaction constitue un critère d'utilisation de la promotion. A l'inverse, un impact négatif sur la probabilité de transformer, traduira un critère où la promotion est dissuasive, et donc à éviter de cibler.

A partir des variables identifiées comme significatives dans le GBM, et après recoupement avec les « connaissances métiers », 9 variables sont retenues et étudiées en régression pénalisée :

- l'âge du conducteur (*agecond*) ;
- l'âge du véhicule à l'achat (*agevachat*) ;

- l'ancienneté du permis (*ancpermis*) ;
- le CRM [16], coefficient de réduction-majoration (*crm*)
- la formule (*formule2*) ;
- l'indicateur de multi-équipement NH [17] (*multequip*) ;
- la nature du vendeur (*natvend*) ;
- l'ancienne compagnie d'assurance (*orgcie3*) ;
- la « qualité » du véhicule (*vehsup*).

Dans un premier temps, on regarde les variables qui interagissent avec le montant de cotisation TTC annuel (*mtcottca*), ainsi, nous faisons une petite sensibilité au prix. De cette première modélisation, il ressort que seule la formule « *Tous Risques Intégral* » interagit négativement avec la cotisation. Ainsi, la sensibilité au prix est accentuée sur cette formule, la probabilité de souscription est plus faible, ce qui est logique étant donné que c'est la formule la plus complète, la plus onéreuse.
Dans un second temps, chacune de ces 9 variables a été croisé avec le montant de cotisation TTC annuel (*mtcottca*) et la valeur de la promotion (*valprom2*) au sein d'un même modèle.
Cependant, en modélisant les interactions avec la valeur de la promotion, nous sommes dans l'incapacité de préconiser des critères concrets. En effet, nous avons seulement un ordre de grandeur, « plus la valeur de la promotion est importante, plus elle a d'effet », sans être capable de déterminer un montant à partir duquel la promotion devient « importante ».
C'est pour cela qu'un troisième modèle est mis en place. Il s'axe plus sur la présence d'une promotion (*indpromo*) que sur sa valeur, car il est plus simple de préconiser des critères avec ou sans promotion. Les résultats de ce modèle étudiant les interactions entre les variables retenues, *mtcottca* et *indpromo* sont illustrés en Figure V.9.

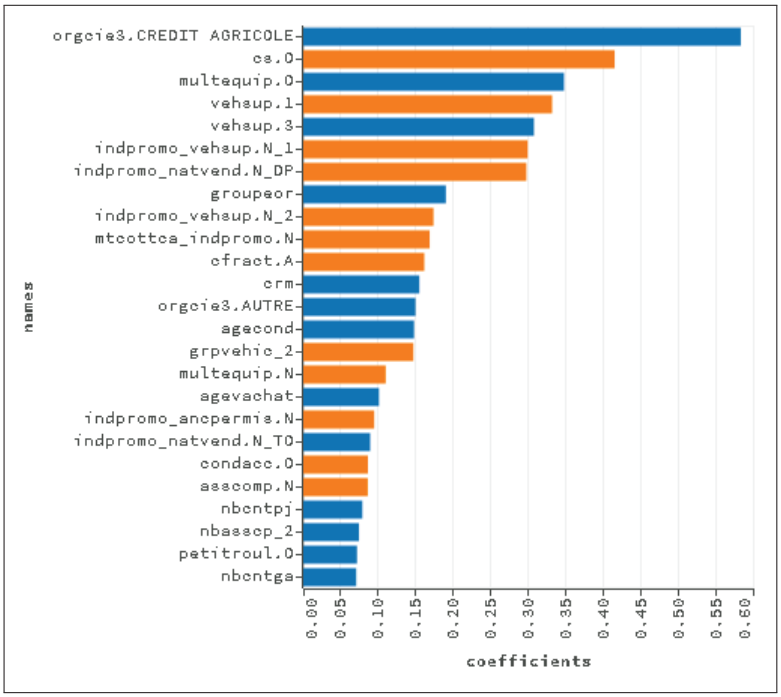


Figure V.9 – *Variables importantes en régression pénalisée Lasso par CR (le bleu est associé à un coefficient positif, et l'orange à un coefficient négatif)*

16. CRM, ou coefficient de réduction-majoration (Bonus/Malus), désigne une méthode de pondération de l'appréciation du risque. En fonction du nombre d'accidents responsables ou co-responsables commis par l'assuré, la cotisation augmente ou diminue à chaque échéance annuelle. En France, ce système de réduction-majoration est réglementé.

17. NH désigne l'assurance habitation

Au regard de la Figure V.9 [18], il semblerait que l'absence de promotions impacte plus, et de façon négative, les projets couvrant un véhicule déjà en possession et assuré (*vehsup.1*), ou un véhicule acheté en remplacement d'un précédent (*vehsup.2*), ainsi que les projets réalisés par un vendeur réseau (*natvend.DP*), de même que les conducteurs présentant une ancienneté de permis (*ancpermis*). A l'inverse, même en absence de promotion, si le vendeur est un téléopérateur (*natvend.TO*), le prospect sera quand même plus sensible à transformer.
De plus, l'une des interactions qui remonte comme significative et impactant négativement la probabilité de transformer c'est l'interaction du montant de la cotisation et l'absence de promotion (i.e. *mttcottca_indpromo.N*). Cette interaction est rassurante et logique. En effet, les projets associés à des cotisations élevées ont moins de chance d'être transformés, c'est la sensibilité au prix. Il est donc normal qu'un projet avec une cotisation élevée et sans promotion ait moins de chance d'être transformé.
A noter, que comme pour le GBM, la qualité du modèle est vérifiée et validée avec les mesures de précision précédemment utilisées, comme l'AUC qui se valorise à 0,76.

Afin d'aller plus loin et de continuer à affiner la compréhension des stratégies commerciales, un GBM et une régression pénalisée sont à nouveau mis en œuvre. L'absence de stratégie commerciale uniforme et nationale entraîne une difficulté supplémentaire à l'étude, car chaque caisse régionale (CR) a sa propre stratégie commerciale. On ne peut donc pas comparer les caisses entre elles, ni faire une analyse et des recommandations spécifique et unique à chaque caisse (trop coûteux en termes de temps). Ainsi, zoomer sur les *clusters* identifiés dans le volet précédent constitue une bonne alternative, car on raisonne en groupe de CR présentant des similitudes (soit 3 groupes) plutôt qu'en CR distinctes. Mais avant de se focaliser sur un seul *cluster*, une régression pénalisée intermédiaire est réalisée.
Il s'agit de la même que celles utilisée précédemment, mais avec un changement de variable, une substitution entre deux variables : celle contenant les *clusters* avec celle des CR. Ainsi, les résultats ne devraient pas changer ou alors à la marge (sauf s'il y a des corrélations) et aucune nouvelle interaction ne devrait apparaître, car les données restent les mêmes. Néanmoins, cette régression permet de vérifier les 3 *clusters* identifiés, car étant donné qu'ils ont des approches stratégiques et des résultats de transformation différents, leurs impacts sur la modélisation sera plus ou moins significatif. De cette façon, le *cluster* caractérisé par des caisses régionales réalisant peu de promotion, à l'exception de la période « temps fort » [19] et avec un taux de transformation élevé apparaît comme le plus significatif dans l'explication de la transformation. Cette présence conforte la qualité de nos *clusters*, et renforce l'idée selon laquelle la démarche commerciale mise en place par ce groupe de caisses régionales est efficace.

La dernière modélisation est réalisée sur le *cluster* n°1, qui se caractérise par un taux de transformation dégradé malgré un usage de promotions commerciales excessif, deux notions qui semblent contradictoires et qui caractérisent l'inverse de l'effet escompté. L'identification de caractéristiques spécifiques à se *cluster* permettrait donc d'affiner le ciblage des opérations commerciales, en mettant en lumière les interactions impactant négativement la probabilité de transformer. Par conséquent, on optimiserait l'usage des promotions en limitant leur utilisation sur ces critères.

18. Lecture des ordonnées :
En cas de non-interaction, seul le nom de la variable suivi de la modalité testée est affichée (*var.modalité*).
En cas, d'interaction c'est le nom des deux variables qui est affiché suivi des modalités associées (*var1.var2_modalité1.modalité2*).
A noter, si la variable est quantitative, seul le nom de la variable est affiché.

19. La période « temps fort » correspond aux mois de septembre - octobre, où on constate chaque année une hausse de l'usage des promotions

2 Focus sur le *cluster* n°1

Par souci de redondance, seuls les résultats du GBM et de la régression pénalisée *Lasso* sont présentés, le procédé ainsi que la mise en place sont strictement identique à la section précédente.
Avec le GBM, on regarde si les variables significatives précédemment identifiées sont toujours présentes ou si de nouvelles variables sont apparues, apportant possiblement un début d'explication au taux de transformation observé dans ce groupe de CR.
Cette dernière modélisation tend à mesurer l'impact de la promotion au sein d'un *cluster*, le n°1, en comparant les β obtenus en sortie de *Lasso* à ceux du premier modèle (i.e. celui avec toutes les CR).
L'une des hypothèses que l'on souhaite vérifier, concerne le faible impact de la promotion sur le taux de transformation. Nous nous attendons à avoir des β en valeur absolue plus petits que ceux obtenus avec le modèle par CR. La seconde hypothèse, concerne l'effet de prix. L'idée est qu'être dans les prix du marché serait un accélérateur du taux de transformation. Dans pareille situation, la variable associée au montant de cotisation TTC annuel (*mtcottca*) apparaîtrait plus significative dans les variables importantes et l'indicatrice de promotion (*indpromo*) serait déclassée. Les β associés à chacune de ces variables seraient également à comparer entre modèles. Dans le cas du *cluster* n°1, l'hypothèse est que les prix proposés par les CR qui le composent sont au-dessus du marché, justifiant un taux de transformation bas.

Seulement, que ce soit la première ou la seconde hypothèse, elles ne semblent pas vérifiées. La comparaison entre les β des deux modèles indique qu'il y a très peu d'écart contrairement à ce qu'on se serait attendu. Sur la Figure V.10 [20] on constate que la variable *mtcottca* n'apparaît pas plus significative, de même pour la variable *indpromo*. Cette dernière n'a pas été déclassée, elle reste importante dans l'explication de la transformation d'un contrat. Ces observations ne vont donc pas dans le sens de notre hypothèse de sur-tarification.

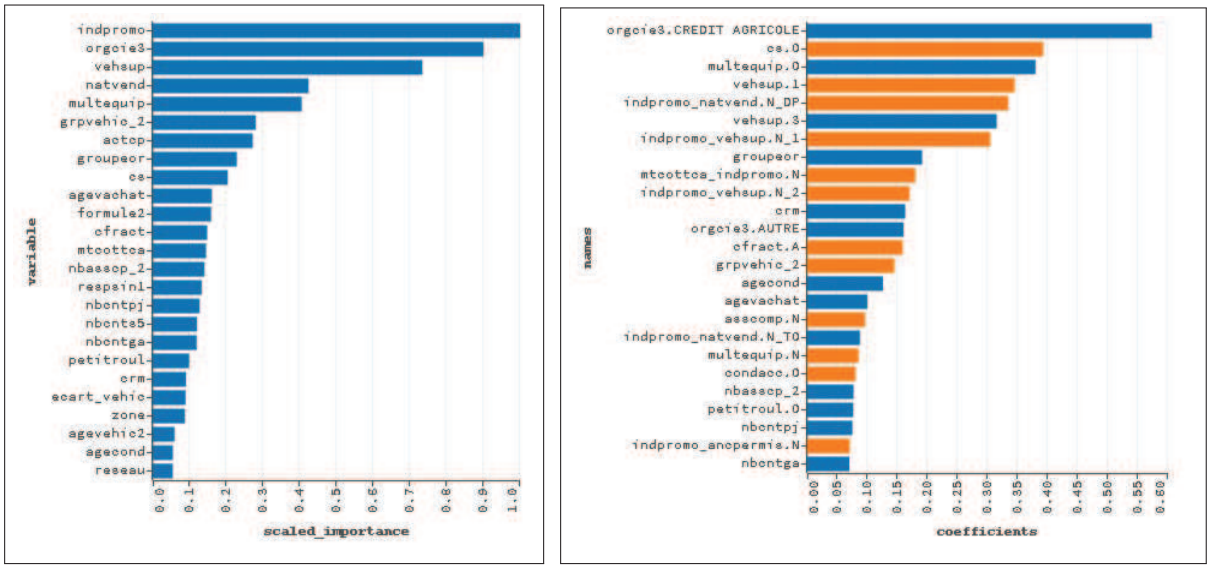


Figure V.10 – *Graphe des variables importantes en GBM (à gauche), en Lasso (à droite) sur les données des CR du cluster n°1*

Cette absence de résultats significativement différents entre le modèle centré sur le *cluster* n°1 et celui avec toutes les CR interroge. Il laisse à penser que les 3 stratégies commerciales identifiées ne sont peut-être pas si discriminantes (ce qui contredit notre observation faite ci-dessus avec la régression pénalisée « substitution des variables CR et *cluster* »), et/ou que

20. Pour la lecture du graphe de droite, le bleu est associé à un coefficient positif, et l'orange à un coefficient négatif

les promotions commerciales impactent de la même façon.
Nonobstant, ce manque de résultats s'explique par le fait que le *cluster* n°1 correspond à 67% des CR, soit 81% des projets de notre étude. Il est donc probable que la surexposition des CR de ce *cluster* dans notre modèle avec toutes les CR, influence les résultats et instaure un biais. Regarder les deux autres *clusters* serait donc légitime, afin de mettre en évidence de nouvelles interactions, et donc de nouveaux profils clients sensibles aux promotions.

Malgré ce défaut de résultats significativement différents, de premières conclusions peuvent être faites sur les quelques interactions identifiées, permettant ainsi d'esquisser un premier comportement client face à la transformation.

3 Identification des critères optimaux d'utilisation de la promotion

3.1 Résultats

Grâce à tous ces modèles pénalisés (*Lasso*), nous pouvons esquisser un certain profil comportemental, et ainsi identifier certains critères sensibles à la présence de promotion.

De cette façon, nous pouvons conclure que la sensibilité au prix est accentuée :

- sur les formules tous risques intégral (formule la plus complète, et donc la plus onéreuse);
- lorsque que le véhicule assuré est acheté pour remplacement (vehsup 2) ;
- en absence de promotion.

L'absence de promotion impactera plus :

- lorsqu'il s'agit du premier véhicule et qu'il est déjà assuré ;
- lorsque que le véhicule assuré est acheté pour remplacement ;
- les conducteurs présentant une ancienneté de permis ;
- les projets réalisés par un vendeur réseau (NATVEND = « DP »).

A l'inverse, seront plus sensible à la promotion :

- les conducteurs âgés ;
- la présence de multi-équipement (NH) ;
- un CRM « élevé ».

Ainsi, la promotion commerciale semble influencer positivement la transformation lorsque :

- le véhicule est déjà en possession et assuré ;
- le conducteur dispose d'une ancienneté de permis ;
- le conducteur est âgé ;
- le projet est réalisé par un vendeur réseau.

Ces premières interactions laissent apparaître une ébauche de profils clients, potentiellement plus sensible à la transformation en présence d'opérations marketing. L'identification de ces profils et plus exactement, l'élaboration de critères d'usage de la promotions : « Si vous êtes dans cette stratégie commerciale, il est conseillé de faire… » était l'objectif de cette modélisation du comportement client, afin d'optimiser le ciblage et d'avoir un meilleur retour sur investissement, et donc une meilleure transformation.
Néanmoins, ces premiers critères, basés sur les interactions sont peu nombreux et faibles. Ils sont à approfondir, à enrichir et surtout à évaluer. En effet, ces derniers se basent uniquement

sur l'appétence à la transformation et occultent la réaction du prospect face aux promotions commerciales.

Ainsi, les critères d'usage identifiés, à date, sont insuffisants et partiellement exploitables. La recherche de critères optimaux doit être poursuivie, afin de pouvoir identifier plus de profils de prospects sensibles à la présence d'une promotion commerciale, et de les asseoir avec, tout en prenant en compte la maturité de la caisse régionale.

3.2 Axes d'amélioration

Pour approfondir ces critères, et parfaire l'optimisation de l'usage des promotions commerciales, une étape de *Market Basket Analysis* (MBA) peut être mise en place.
Le MBA a pour vocation de mettre en évidence et d'identifier des combinaisons de facteurs caractéristiques de profils de prospects sensibles aux promotions commerciales et présentant une forte probabilité de transformer.
Cette recherche de combinaisons se veut complémentaire et plus fines que celles obtenues avec l'analyse des interactions (*Lasso*).

In fine, ces combinaisons de facteurs devraient être à la base des critères d'usage. Mais avant de valider ces combinaisons, il faudra mesurer l'impact de la promotion commerciale.
Dans quelle mesure la promotion joue-t-elle un rôle dans les combinaisons identifiées ? Est-ce que les combinaisons identifiées comme favorables à la transformation grâce à la présence d'une promotion le sont vraiment ? Comment être sûr que c'est bien la présence de la promotion commerciale qui entraîne la souscription et pas les autres facteurs qui constituent la combinaison.
Si on prend l'exemple de la combinaison de facteurs suivante : jeunes conducteurs âgés de 18 à 25 ans, avec une ancienneté de permis inférieure à 5 ans et en possession de leur premier véhicule. Comment être sûr que les prospects répondant à ce profil, n'auraient pas souscrit même sans promotion ?

Pour mesurer l'impact réel de la promotion et valider les combinaisons identifiées, une approche *uplift* est mise en place, et s'inscrit dans la continuité de cette étude sur le comportement client.
Cette approche de ciblage va permettre :

- d'affiner le ciblage des prospects, notamment en identifiant les segments contenant les prospects réellement sensibles aux promotions commerciales, les « influençables » et ainsi optimiser la marge en ciblant les segments les plus rentables ;
- de réduire l'hétérogénéité du taux de transformation entre les différentes populations identifiées afin de rester compétitif pour l'ensemble des segments.

Les conclusions, nous permettrons de préconiser une hausse de l'usage des promotions, et à l'accentuer sur certaines cibles telles que les formules *Tous risques*, les véhicules de remplacement ou déjà en possession et assurés et une certaine catégorie d'anciens assureurs...
Par ailleurs, après intégration de ces éléments, notamment de la hausse de l'usage des promotions, le taux de transformation simulé augmente, est plus homogène et permet de regagner des segments tels que les «petits *CRM* » [21] .

21. Derrière ce terme de « petits *CRM* », on désigne les assurés avec un coefficient de réduction-majoration peu élevé ; les « bons » conducteurs, ceux présentant un « faible » historique de sinistres ou nul.

Conclusions

Dans un environnement très concurrentiel, où le consommateur est de plus en plus exigeant et mobile, l'étude du taux de transformation d'une compagnie d'assurance constitue un outil très efficace pour mesurer sa performance et sa présence sur le marché.

L'enjeu de ce mémoire était de mesurer l'impact des promotions commerciales sur le taux de transformation, en modélisant le comportement client à la souscription, dans l'intention d'identifier des profils de consommateur plus sensibles à la promotion et ainsi formuler des critères d'utilisation de promotion optimaux. Pour ce faire, différentes étapes ont été effectuées.

Dans un premier temps, il était essentiel de comprendre ce qu'est une promotion commerciale et comment elle est utilisée de nos jours au sein du réseau Pacifica. Cette première phase exploratoire a permis de distinguer trois stratégies commerciales pratiquées par les caisses régionales.

- Une utilisation excessive de promotions, sans effet avéré sur le taux de transformation.
- Une utilisation rare, exceptée en période « temps fort » impactant favorablement le taux de transformation.
- Une utilisation plus parcimonieuse des promotions, avec une appétence pour les promotions de type « forfait » et présentant un effet positif sur le taux de transformation.

Une fois cet environnement promotionnel connu et maîtrisé, nous modélisons le taux de transformation avec la méthode GBM qui nous servira à identifier les variables les plus importantes, en vue de les intégrer à la modélisation du taux de transformation par GLM. Parmi ces variables les plus significatives, on trouve le nom de l'ancienne compagnie d'assurance du prospect, l'indicatrice de promotion, l'indicatrice de véhicule supplémentaire au foyer, la caisse régionale, la nature du vendeur, la présence de multi-équipement. . . .

Grâce aux graphes de dépendance partielle, il fut possible d'avoir une première piste de lecture, une première interprétation de ces variables importantes. De cette façon, nous avons pu identifier le multi-équipement comme un facteur favorable à la transformation. Plus un prospect est équipé, plus la transformation est facile. En revanche, si le véhicule est déjà en possession et déjà assuré, la transformation est plus compliquée ; nous sommes en conquête « pure ».

Enfin, les interactions étant plus difficiles à saisir en GBM, la régression pénalisée mit en œuvre par la suite a permis de déterminer des profils plus sensibles à la présence de promotions. Ainsi, l'absence de promotions à un effet négatif plus marqué lorsque le bien à assurer est déjà en possession et assuré, ou qu'il est acheté en remplacement d'un précédent. Les conducteurs avec une ancienneté de permis, ou réalisant leur projet en agence seront également plus sensibles à l'absence de promotion.

Dans ce contexte et étant donné un budget de promotion commerciale limité, établir des critères d'utilisation de la promotion optimaux prend tout son sens, afin de cibler certains consommateurs selon leurs caractéristiques, et les asseoir avec une promotion. En optimisant ainsi l'allocation des promotions, on maximise le taux de transformation. A date, les promotions commerciales semblent influencer positivement la transformation des consommateurs multi-équipés, des consommateurs avec de l'ancienneté de permis mais également certaines « qualité » du véhicule supplémentaire, notamment quand le véhicule est déjà en possession et assuré par ailleurs.

Cependant définir des profils de consommateurs et des critères optimaux peut être un travail fastidieux, confronté à certaines limites. La difficulté de chaîner de façon sûr les devis, propositions et affaires nouvelles entre eux constituent un biais dans notre étude. L'absence d'uniformisation des pratiques commerciales entre caisses régionales a également constitué un frein, en limitant la comparaison directe entre caisses régionales. A date de ce mémoire, nous ne sommes pas parvenus à identifier distinctement tous les critères optimaux d'usage de la promotion, et à formaliser des recommandations.
Ainsi, les pistes d'améliorations restent multiples, qu'elles soient liées à une meilleure appréhension de la multitude de stratégies commerciales, à l'étoffement des critères optimaux par *Market Basket Analysis* ou à l'évaluation de ces derniers... En effet, il serait intéressant de procéder à une approche *uplift* pour mesurer l'impact réel des promotions sur les critères identifiés afin de les valider. Enfin, notre étude s'axe seulement sur la transformation sans prendre en compte la notion de rentabilité ; calculer la marge associée à chaque segment client permettrait d'identifier les segments les plus rentables et ainsi de les privilégier lors de l'allocation des promotions.

Liste des sigles

AFC : Analyse Factorielle des Correspondances
ACM : Analyse des Correspondances Multiples
ACP : Analyse en Composantes Principales
AN: Affaire Nouvelle
AUC : Aire sous la courbe (*Area Under the Curve* en anglais)
CAH : Classification Ascendante Hiérarchique
CR : Caisse régionale
CRM : Coefficient de Réduction-Majoration
CV : Validation croisée (*Cross Validation* en anglais)
DP : Dépendance partielle
EMV : Estimation du Maximum de Vraisemblance
FN : Faux négatif
FP : Faux positif
GBM : *Gradient Boosting Maching*
GLM : Modèle linéaire généralisé (*Generalized Linear Model* en anglais)
ICE : Espérance conditionnelle individuelle (*Individual Conditional Expectation* en anglais)
INSEE : Institut National de la Statistique et des Études Économiques
KPI : Indicateur de performance (*Key Performance Indicator* en anglais)
MBA : analyse du « panier de consommation » (*Market Basket Analysis* en anglais)
OR : *Odds Ratios*
RA : Recherche Approximative
RF : Forêt aléatoire (*Random Forest* en anglais)
ROC : *Receiver Operating Characteristic*
TFP : Taux de Faux Positif
TVP : Taux de Vrai Positif
VN : Vrai négatif
VP : Vrai positif

Lexique

Conquête : Démarche qui consiste à trouver de nouvelles parts de marché, en attirant et en captant de nouveaux clients. Cette recherche de nouveaux clients passe par l'élaboration d'une bonne stratégie, dans laquelle on peut retrouver l'usage de promotions commerciales. En assurance, on distingue deux types de nouveaux clients : ceux déjà présent en portefeuille car possédant un autre produit d'assurance (exemple, un client souscrivant un nouveau produit, une assurance automobile, alors qu'il est déjà client car équipé d'une assurance habitation), et les clients réellement nouveaux, qui ne détiennent aucun produit d'assurance et sont absents de tous les portefeuilles d'assurance.

Caisse régionale (CR) : Réseau d'agences distributrices lié à un maillage géographique.

CRM, ou coefficient de réduction-majoration (Bonus/Malus) : Désigne une méthode de pondération de l'appréciation du risque. En fonction du nombre d'accidents responsables ou coresponsables commis par l'assuré, la cotisation augmente ou diminue à chaque échéance annuelle. En France, ce système de réduction-majoration est réglementé.
Plus le CRM est élevé, plus le risque est important.

Critère d'optimisation : Préconisations sur l'usage de la promotion, afin de cibler les clients les plus sensibles aux offres commerciales et maximiser leurs effets.

Multi-équipement : Terme désignant un client déjà en portefeuille et possédant déjà un contrat d'assurance sur un autre produit. A l'inverse, si le client ne détient aucun autre équipement que le contrat qu'il vient de souscrire alors on dit qu'il est mono-équipé.

Période « temps fort » : Chaque année, en septembre – octobre, il est fait constat d'une hausse des promotions. Cette période et ces promotions sont appelées « temps fort ».
Cet indicateur tend à mesurer pour chaque agence, qu'elle est la part de ces promotions « temps fort » sur une année.

Projet : Désigne indifféremment un devis ou une proposition.

Promotion commerciale : Une promotion commerciale est une technique de communication organisée de façon ponctuelle pour encourager les ventes d'un produit ou d'un service sur le court terme.

Prospect : Client potentiel

Run-off ou liquidation de portefeuille : Consiste à arrêter la souscription d'affaires nouvelles sur un portefeuille tout en gérant dans le temps, le traitement, les provisions techniques, les sinistres jusqu'à l'épuisement complet du portefeuille.

Scoring : En marketing, le scoring est une technique qui permet d'affecter à un client ou prospect, un score. Ce score traduit souvent la probabilité qu'un individu réponde à une sollicitation marketing ou appartienne à la cible recherchée. Il peut également exprimer la valeur potentielle d'un client ou prospect.

Stratégie commerciale : Ensemble de méthodes ayant pour but d'optimiser la croissance de l'activité, l'usage de promotions commerciales fait partie de ces méthodes.

Taux de transformation, ou taux de concrétisation, de souscription, de conversion : Est un indicateur très utilisé pour mesurer l'efficacité et la rentabilité d'un produit. L'analyse de ce taux permet un suivi, ainsi qu'une optimisation de la commercialisation du produit.
Il s'exprime comme le rapport entre le nombre de nouveaux clients et le nombre de prospects. Autrement dit, le nombre d'individus ayant réalisé un projet versus le nombre de clients ayant souscrit un contrat.

$$Taux\ de\ transformation = \frac{Nombre\ de\ projets\ transformés\ en\ contrat}{Nombre\ de\ projets\ réalisés}$$

Typologie comportementale : Classification d'individus selon un ensemble de caractères, afin de constituer des grandes familles (=typologie) d'individus présentant des similitudes. Dans le cadre de ce mémoire, on classifie les CR selon leurs pratiques commerciales afin de constituer des familles de CR affichant des ressemblances.

Validation croisée ou *Cross validation* : Consiste à diviser en k sous-échantillons (ou plis) notre jeu de données. Chaque pli sera ensuite utilisé une fois comme validation tandis que les k - 1 plis restants forment l'ensemble d'apprentissage. Cette étape de validation croisée est une alternative à l'usage d'une base de validation, qui permet de pallier aux prédictions qui pourraient rester hasardeuses et instables si on relançait la modélisation, et si on utilisait seulement une base d'apprentissage et de test. Ainsi, cette base de validation permet de calibrer le modèle avant de procéder à la prédiction, mais elle nécessite un échantillon de grande taille, d'où la validation croisée.

Annexes

Annexe A : Démonstrations

A.1 GLM : Démonstration de l'équation (I.1)

Si Y est distribuée selon une loi appartenant à une famille exponentielle, alors $\mathbb{E}(Y) = b'(\theta)$ et $Var(Y) = \phi b''(\theta)$

Preuve: Comme f appartient à une famille exponentielle, elle est de la forme

$$f(y) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

- Pour $\mathbb{E}(Y) = b'(\theta)$:

  $$\frac{\partial f}{\partial \theta} = f(y).\frac{y - b'(\theta)}{a(\phi)}$$

  On intègre de chaque côté par rapport à y.
  D'une part

  $$\int_y \frac{\partial f}{\partial \theta} dy = \frac{\partial}{\partial \theta} \int_y f dy = 0$$

  De l'autre

  $$\int_y f(y).\frac{y - b'(\theta)}{a(\phi)} = \frac{\mathbb{E}(Y) - b'(\theta)}{a(\phi)}$$

  Ce qui donne bien

  $$\frac{\mathbb{E}(Y) - b'(\theta)}{a(\phi)} = 0 \Longleftrightarrow \mathbb{E}(Y) = b'(\theta)$$

- Pour $Var(Y) = \phi b''(\theta)$:
  On reprend l'équation précédente, et on a

  $$\frac{\partial^2 f}{\partial \theta^2} = f(y).\left(\frac{y - b'(\theta)}{a(\phi)}\right)^2 - f(y).\frac{b''(\theta)}{a(\phi)}$$

  On intègre de chaque côté, soit

  $$\int_y \frac{\partial^2 f}{\partial \theta^2} dy = \frac{\partial^2}{\partial \theta^2} \int_y f dy = 0$$

Et

$$\int_y \left(f(y).\left(\frac{y-b'(\theta)}{a(\phi)}\right)^2 - f(y).\frac{b''(\theta)}{a(\phi)} \right) = \int_y .f(y).\frac{(y-\mathbb{E}(Y))^2}{a(\phi)^2}dy - \frac{b''(\theta)}{a(\phi)}$$
$$= \frac{Var(Y)}{a(\phi)^2} - \frac{b''(\theta)}{a(\phi)}$$

Ainsi, on trouve bien

$$Var(Y) = b''(\theta).a(\phi)\int_y \frac{\partial^2 f}{\partial\theta^2}dy = \frac{\partial^2}{\partial\theta^2}\int_y f dy = 0$$

A.2 GLM : Démonstration de l'équation (I.2 et I.3)

Soit Y_i (i = 1, ..., n) des variables aléatoires indépendantes et distribuées selon une même loi de la famille exponentielle dépendant d'un paramètre $\theta \in \mathbb{R}^d$. Soit $y_1, ..., y_n$ des observations de ces variables aléatoires.

La fonction de vraisemblance pour cet échantillon $(y_1, ..., y_n)$ dans le cas de modèle exponentiel, s'écrit

$$\mathcal{L}(\theta,\phi) = \mathcal{L}(\theta,\phi,y_1,...,y_n) = \prod_{i=1}^{n} f(y_i,\theta,\phi)$$

et on définit la log-vraisemblance comme

$$\begin{aligned} l(\theta,\phi) &= log\mathcal{L}(\theta,\phi,y_1,...,y_n) \\ &= \sum i = 1^n log\left(f(y_i,\theta,\phi)\right) \\ &= \sum i = 1^n \left(\frac{y_i\theta_i - b(\theta)}{\phi} - c(y_i,\phi)\right) \end{aligned}$$

Pour estimer les paramètres β il suffit de deviner la log-vraisemblance par rapport à β et d'écrire les conditions du premier ordre.
On rappelle

- $X = x_{i,j}$, la matrice où sont rangées les p variables explicatives
- $\beta = (\beta_1, ..., \beta_p)$ le vecteur des paramètres dans $mathbbR^p$
- et $\eta = X\beta$ le prédicteur linéaire.

La fonction de lien g est supposée monotone différentiable telle que: $\eta_i = g(\mu_i) = \beta x_i'$ et on a $g(\mu_i) = \theta_i$ si c'est la fonction de lien canonique.

Enfin d'après $\mathbb{E}(Y_i) = b'(\theta)$ et $Var(Y_i) = b''(\theta)\phi$, il en resort que μ_i dépend uniquement de θ_i, et comme $g(\eta_i) = \beta x_i'$, on a θ qui dépend de β. C'est pour cela qu'on peut écrire $\mathcal{L}(\beta)$.

Pour i et j donnés, on a:

$$\frac{\partial l_i}{\partial\beta_j} = \frac{\partial ln\mathcal{L}(i)}{\partial\beta_j} = \frac{\partial ln\mathcal{L}_i}{\partial\theta_i}.\frac{\partial\theta_i}{\partial\beta_j}$$

D'une part

$$\frac{\partial ln\mathcal{L}_i}{\partial\beta_j} = \frac{y_i - b'(\theta_i)}{\phi} = \frac{y_i-\mu_i}{\phi}$$

D'autre part, comme $\eta_i = x_i\beta$ on peut écrire

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i}.\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i}.x_{i,j}$$

où $x_{i,j}$ la j^{ime} coordonnées de X_i

Enfin,

$$\frac{\partial \theta_i}{\partial \eta_i} = \frac{\partial \theta_i}{\partial \mu_i}.\frac{\partial \mu_i}{\partial \eta_i}$$

où,

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1} = \left(\frac{Var(Y)}{\phi}\right)^{-1}$$

car Var(Y) =$\phi.b''(\theta) = \phi\frac{\partial \mu_i}{\partial \theta_i}$ et $\frac{\partial \mu_i}{\partial \eta_i}$ dépend de la fonction lien $\eta_i = g(\mu_i)$

On obtient ainsi les équations suivantes

$$\begin{aligned}
\sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi}.x_{i,j}.\frac{\phi}{Var(Y_i)}.\frac{\partial \mu_i}{\partial \eta_i} \\
&= \sum_{i=1}^{n} \frac{y_i - \mu_i}{Var(Y_i)}.x_{i,j}.\frac{\partial \mu_i}{\partial \eta_i} \\
&= 0 \qquad\qquad \text{pour tout j = 1, ..., p}
\end{aligned}$$

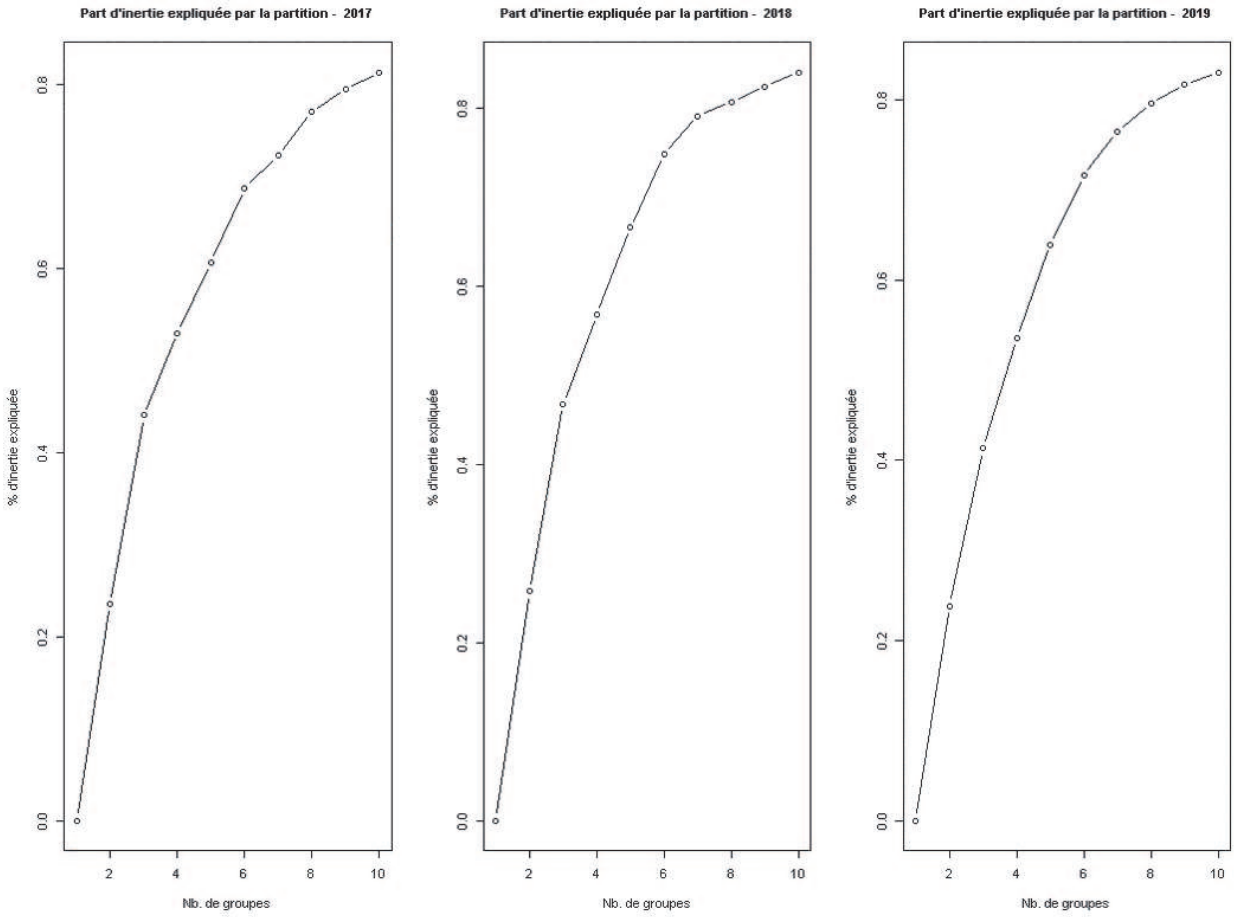
Annexe B : Liste des variables

NOM	LIBELLÉ
***a**ctcp*	Code activité du conducteur principal
agecond	Âge du conducteur
agetrfcp2	Âge tarifaire du conducteur principal, en nombre de mois
agevachat	Âge du véhicule à l'achat, exprimé en mois
agevehic2	Âge du véhicule
ancpermis	Ancienneté du permis
asscomp	Option assistance complète
***b**ancassurance*	Indicateur de score du client
***c**apdec*	Indicateur extension capital décote souscrite
cataeu2010	Zonage en Aires Urbaines
ca_lcl_100k_qtl_com	Nombre d'agences CA-LCL par commune (pour 100K hab) (QTL)
cfract	Code fractionnement
chomage_t42018_ze2010_qtl_com	Taux chômage (zone emploi T4-2018) par commune (QTL)
classe	Code classe du véhicule
co2emis	CO2 émis par le véhicule
coefabc	Coefficient ABC
coefcom	Coefficient commercial appliqué
condac	Indicateur formation en conduite accompagnée
cpxsraor	Classe de prix d'origine
cr	Référence externe de la caisse régionale
crm	Coefficient de réduction-majoration attribué
cs	Présence de conducteur secondaire non assuré par ailleurs
***d**ens_pop_round_com*	Densité de la population par commune (QTL)
detentvh	Durée de détention du véhicule précédent
dist_min_ca_lcl_qtl	Distance minimale à l'agence CA_LCL la plus proche (QTL)
***e**cart_vehic*	Écart entre l'ancien et le nouveau véhiculier (*vehiculier_2* et *grpvehic_2*)
energie	Code énergie
espinsee	Espace INSEE
***f**ormule2*	Code de la formule commerciale du projet
***g**enre*	Genre du véhicule sur la carte grise (véh. particulier, utilitaire, camping car)
groupe	Code groupe du véhicule
groupeor	Groupe d'origine du véhicule assuré
grpsraor	Groupe d'origine SRA
grpvehic_2	Nouveau groupe de véhicule en fonction de la formule
***i**nd4x4*	Indicateur véhicule 4X4
indbdg	Indicateur d'option « bris de glace » souscrite
indeff	Indicateur de présence de la garantie « effets et bagages »
indpann	Indicateur de garantie « panne auto » ou « panne auto plus » souscrite ; ancienne formule
indpannplus	Indicateur d'option de garantie « panne auto » ou « panne auto plus » souscrite ; nouvelle formule
indpromo	Indicateur de présence d'une promotion commerciale
***l**ongueur*	Longueur du véhicule

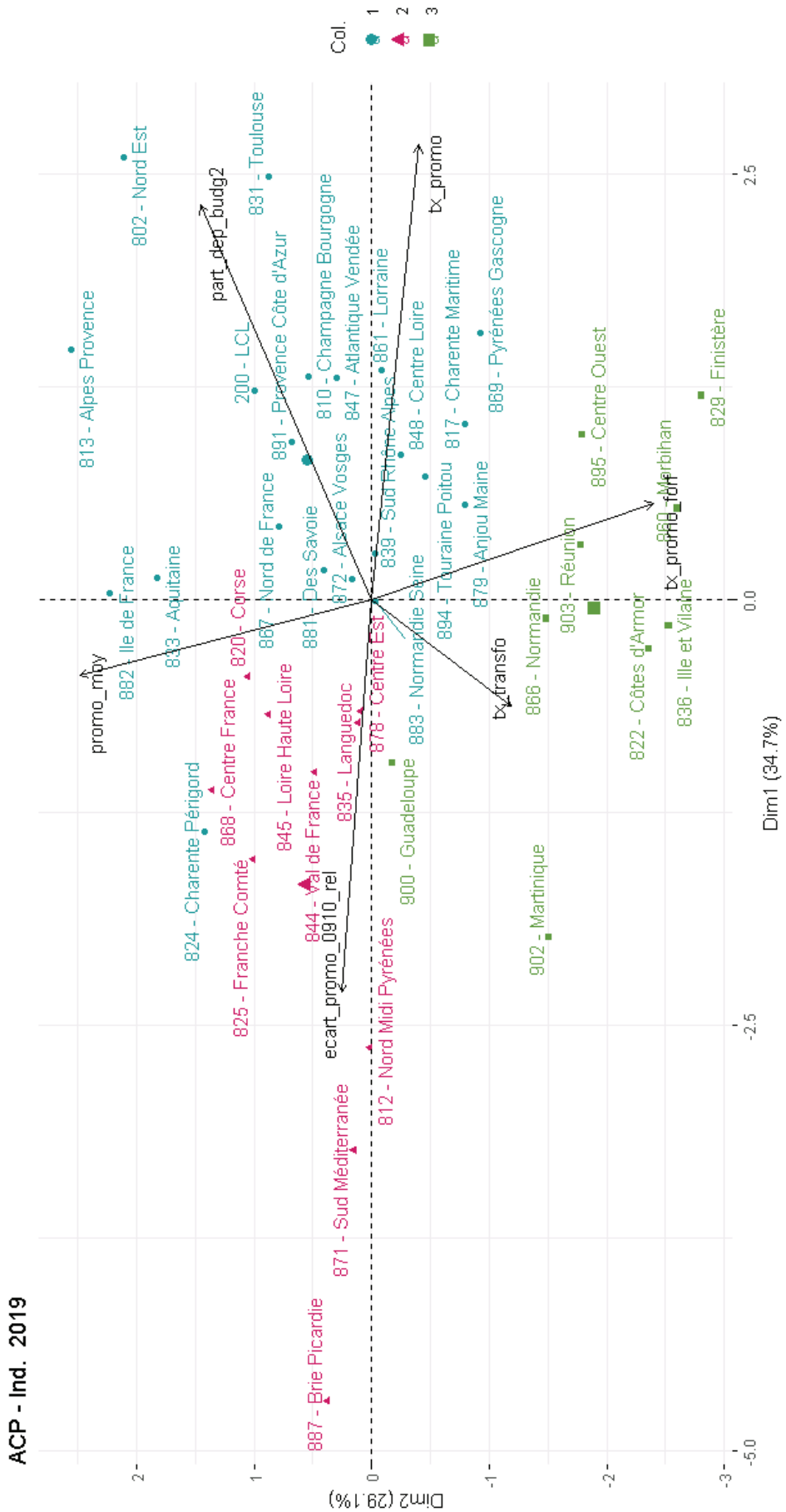
NOM	LIBELLÉ
***m**ed_niv_vie_2015_k_com*	Médiane niveau de vie 2015 par commune (en k)
mtcotnh	Montant de cotisation d'un contrat potentiel en NH
mtcottca	Montant de cotisation TTC annuel
multequip	Indicateur d'une réduction pour multi équipement A4 NH
***n**atvend*	Nature du vendeur
nbasscp_2	Nombre de mois d'assurance du conducteur
nbcnta2	Nombre de contrats potentiels en A2 pour le client
nbcntav	Nombre de contrats potentiels en AV pour le client
nbcntga	Nombre de contrats produit GA souscrits
nbcntnh	Nombre de contrats potentiels en NH pour le client
nbcntpj	Nombre de contrats potentiels en PJ pour le client
nbcnts4	Nombre de contrats potentiels en S4 pour le client
nbcnts5	Nombre de contrats potentiels en S5 pour le client
nbcntvu	Nombre de contrats produit VU souscrits
nbenfant	Nombre d'enfants du client
nbsin	Nombre de sinistres dans la compagnie antérieure
***o**rgcie3*	Ancienne compagnie d'assurances du prospect (après correction et regroupement de la variable *orgcie*)
***p**arking*	Mode de stationnement la nuit
petitroul	Option faible kilométrage
poidspuissance	Rapport entre les variables *poidsvid* et *puisdin*
poidsvid	Poids à vide du véhicule
promo_euros	Valeur de la promotion commerciale en euros
puisdin	Puissance du véhicule
***r**eseau*	Réseau (Crédit Agricole ou LCL)
respsin1	Code responsabilité des sinistres
revenu_fiscal_uc_2016_k_com	Revenu fiscal par commune (en k)
***s**ecnov*	Présence d'un conducteur secondaire novice
sexe	Code sexe du client
sexecp	Code sexe du conducteur principal
***t**ypers*	Type de personne (morale ou physique)
typfranc	Type de franchise
typprom	Type de remboursement. Spécifie si la promotion est de type « forfait » (ex: 10€ offert) ou de type « coefficient » (ex: 10% offert)
***v**alprom*	Valeur du coefficient ou du montant selon le type de remboursement
vehiculier_2	Véhiculier en vigueur au moment du projet
vehsup	Indicateur de véhicule supplémentaire au foyer (premier véh., véh. déjà en possession et assuré, véh. de remplacement...)
vitesse	Vitesse maximum du véhicule
***z**one*	Code zone tarifaire

Annexe C : Graphiques présents dans le TDB, affichant le taux de transformation et de promotion selon le critère multi-équipement

Annexe D : Résultats de la méthode des k-means

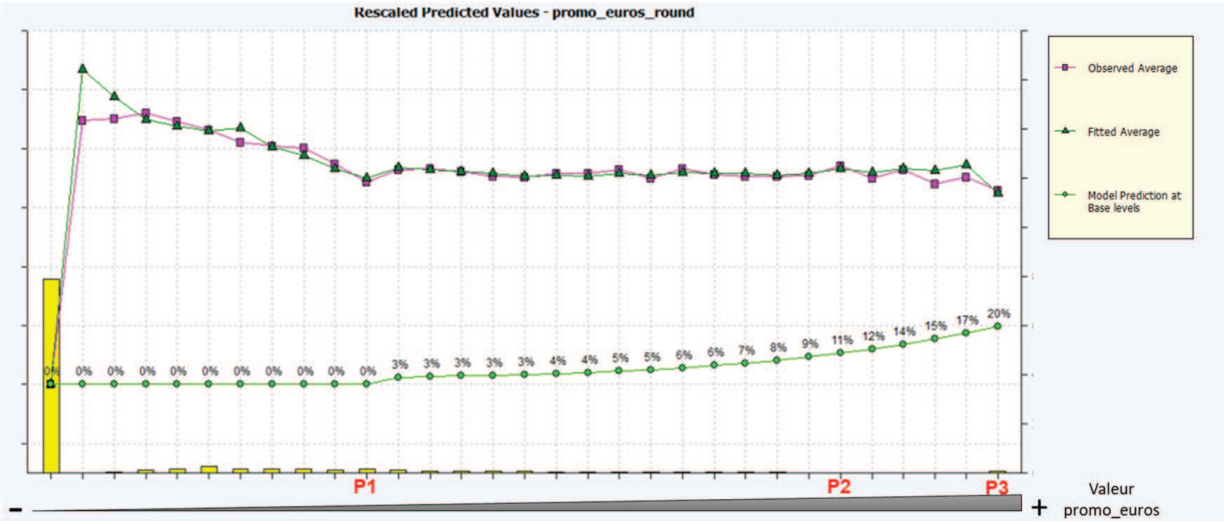


Annexe E : ACP - Graphe des variables et des individus pour l'année 2019

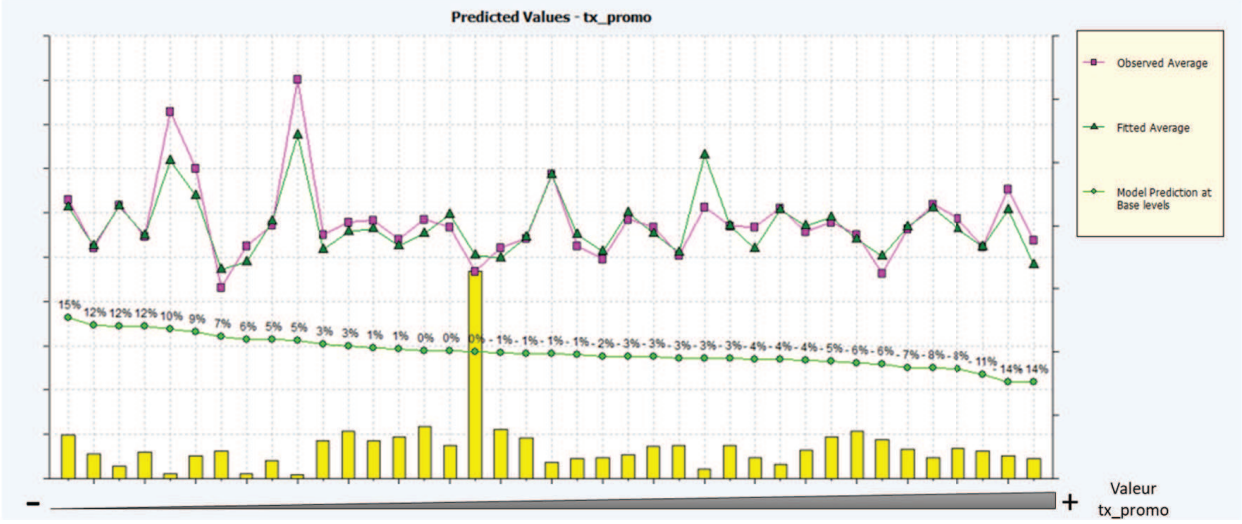


Annexe F : Résultats de la modélisation du taux de transformation par GLM

F.1 Valeurs prédites de la variable *promo_euros*



F.2 Valeurs prédites de la variable *tx_promo*



Afin de conserver la confidentialité des données, les axes ont été masqués. On peut cependant voir un effet de la promotion équivalent sur les promotions inférieures à un certain montant, P1€, et accéléré entre P2€ et P3€. Et un effet couplé à une pénalisation de la fréquence d'utilisation de la promotion commerciale, +10 points d'usage génère -6% d'effet.

Bibliographie

Mémoires

1. **CHARGUÉRAUD Alice** - *Le taux de transformation en automobile : comparaison de différentes méthodes d'apprentissage.*
2. **GUILLOT Antoine** - *Apprentissage statistique en tarification non-vie : quel avantage opérationnel ?*
3. **JERRARI Omar** - *Modélisation dynamique du coût des inondations historiques en France.*
4. **LUCCHINO Arthur** - *Optimisation du ciblage client dans un centre d'appel*, 2015.
5. **PARIENTE Jennifer** - *Modélisation du risque géographique en assurance habitation*, 2017.
6. **SOIX Emilie** - *Estimation du ratio de solvabilité à l'aide de méthodes d'apprentissage statistique supervisé.*

Articles et publications

1. **Automobile Club Association (ACA)** - *Budget de l'Automobiliste*, mars 2018.
2. **Comité des Constructeurs Francais d'Automobiles (CCFA)** - *L'industrie automobile française - Analyse et statistiques 2019.*
3. **DIEBOLD Jean-Blaize**, Directeur Marketing Europe du Sud, Pitney Bowes Software - *Optimisation du ciblage des opérations de fidélisation.*
4. **GUILLOT Antoine** - *Détection des interactions en tarification non-vie.* Bulletin français d'actuariat, janvier - juin 2019.
5. **EIDELMAN Alexis** - *La valeur de Shapley - Comment individualiser le résultat d'un groupe.* INSEE - Direction des statistiques démographiques et sociales, janvier 2012.
6. **TREMBLAY Charles** - *Interprétabilité des modèles - Méthode des valeurs de Shapley; Formation Base.* Société de conseil Kobia, janvier 2020.

Livres

1. **FRIEDMAN Jerome, HASTIE Trevor, TIBSHIRANI Robert** - *The Elements of Statistical Learning; Data Mining, Inference, and Prediction.* Second Edition (Springer Series in Statistics), Février 2009.
2. **MOLNAR Christoph** - *Interpretable Machine Learning - A Guide for Making Black Box Model Explainable. e-book.*

Supports de cours

1. **GONZALEZ Pierre-Louis** - *L'Analyse en Composantes Principales (A.C.P).*
2. **HUSSON François** - *Classification ascendante hiérarchique (CAH).* Laboratoire de mathématiques appliquées, Agrocampus Rennes.
3. **LECROQ Thierry** - *Distance entre mots.* Université de Rouen.

4. **RAKOTOMALALA Ricco** - *Gradient Boosting - Technique ensembliste pour l'analyse prédictive. Introduction explicite d'une fonction de coût.* Université Lyon 2.
5. **THOMAS Maud** - *Econométrie de l'assurance non vie - Tarification a priori.* ISUP, 2019.
6. **Université Le Mans** - *Les Soundex.* Site internet.

Sites internet

1. **1min30** - https://www.1min30.com/evenementiel/marketing-promotionnel-6082
   Consultation de l'article *Le marketing promotionnel ou promotion des ventes*
2. **Astera** - http://jeux-et-mathematiques.davalan.org/lang/algo/lev/index.html
   Consultation de l'article *Gestion de la qualité des données: qu'est-ce que cela signifie et pourquoi c'est important?*
3. **Datacamp** - datacamp.com/community/tutorials/market-basket-analysis-r
   Consultation de l'article *Market Basket Analysis using R*
4. **Interpretable Machine Learning** - https://christophm.github.io/interpretable-ml-book.html
   Consultation de plusieurs sujets: les graphes de dépendance partielle, les courbes d'espérance conditionnelle, les valeurs de Shapley
5. **Jeux et Mathématiques** - http://jeux-et-mathematiques.davalan.org/lang/algo/lev/index.html
   Consultation de l'article *Distance de Levenshtein*
6. **Le blog du dirigeant** - https://www.leblogdudirigeant.com/le-ciblage/
   Consultation de l'article *Qu'est-Ce Que Le Ciblage Marketing ?*
7. **Towards data science** - https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce
   Consultation de l'article *A Gentle Introduction on Market Basket Analysis — Association Rules*

Table des matières