HEC MONTRÉAL

Élaboration d'un modèle de profitabilité

par

Nissay Lim

Sciences de la gestion
(Intelligence d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M.Sc.)

13 mars 2014

Sommaire

Les entreprises du secteur énergétique offrent de plus en plus une gamme complète de produits et services pour satisfaire leur clientèle. Il est important pour elles de bien déterminer la profitabilité reliée à chacun de leurs clients. L'objectif de la présente étude est d'élaborer un modèle de profitabilité pour la clientèle de ABC.

Pour ce faire, nous avons étudié trois modèles classiques de régression : linéaire multiple, gamma et logistique avec et sans interactions. Nous avons également créé un nouveau modèle qui est la combinaison de la régression logistique et de la régression linéaire multiple. Au total, neuf modèles ont été évalués pour expliquer la profitabilité client. La performance des modèles est jugée selon les différents critères de sélection qui leur sont propres.

Le modèle de la régression linéaire multiple avec interactions a été retenu car il s'agit du modèle qui a donné la plus petite erreur de prévision. De plus, c'est un modèle qui est facilement interprétable. Ce modèle permet d'expliquer à 13,45% la profitabilité moyenne du client de ABC. Il permet également de départager les clients profitables des clients non profitables dans 90% des cas.

Suite à ces analyses, nous recommandons trois actions précises à ABC. La première est d'augmenter le prix de ses produits. La deuxième est de facturer les interventions ou d'introduire un concept de franchise après un certain nombre d'interventions pour assurer la profitabilité. La dernière action est d'exclure certains équipements de son programme de protection.

Table des matières

Liste des tableaux

Tables des figures

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué, directement ou indirectement, à l'aboutissement de ce mémoire. Leur collaboration et leurs encouragements ont été des facteurs déterminants pour mener à terme ce projet.

Merci à Marc Fredette, mon directeur de mémoire, de m'avoir orientée et éclairée par ses précieux conseils et recommandations tout au long de la rédaction de ce mémoire.

Merci à la direction de l'entreprise ABC de m'avoir donné l'opportunité de faire un projet de mémoire avec leurs données. Leur confiance et leur support financier ont contribué positivement à l'accomplissement de ce mémoire. Je veux particulièrement remercier le directeur marketing de m'avoir accompagnée lors des différentes phases du projet.

Finalement, un merci tout spécial à ma famille et à mes amies pour leurs encouragements et leur support moral.

Chapitre 1

Introduction

ABC est une entreprise de service qui œuvre dans le secteur énergétique. Elle assure une présence dans le marché résidentiel, commercial et institutionnel. Elle offre une gamme complète de produits et services pour satisfaire sa clientèle en ce qui concerne l'installation, l'entretien et la protection des équipements de chauffage.

Parmi les activités de ABC, celle qui nous intéresse pour l'étude est la protection des équipements de chauffage chez la clientèle résidentielle. En d'autres termes, ABC offre une forme d'assurance à sa clientèle appelée communément « programme de protection ». Selon le dictionnaire Le Petit Robert[1], une assurance se définit comme suit : « *Contrat par lequel un assureur garantit à l'assuré, moyennant une prime ou une cotisation, le paiement d'une somme convenue en cas de réalisation d'un risque déterminé* ». Contrairement aux autres assureurs qui fixent habituellement la tarification selon le type de produits/services assurés ainsi que l'historique du client, ABC tient uniquement compte de l'équipement pour établir son prix. Cette politique particulière fait

[1] Le nouveau Petit Robert, dictionnaire alphabétique et analogique de la langue française – version électronique. Nouvelle édition du Petit Robert de Paul Robert, 2012. http://www.lerobert.com

en sorte que certains clients sont très profitables et d'autres le sont moins. À un autre extrême, assurer certains clients coûte plus d'argent à l'entreprise que s'ils n'étaient pas clients. Une telle situation ne permet pas à ABC de maximiser ses profits.

La présente étude porte sur l'élaboration d'un modèle de profitabilité d'un client de ABC. Notre contribution consiste à analyser la base de données fournie par l'entreprise et de faire ressortir les caractéristiques qui expliqueraient la profitabilité. L'objectif poursuivi est de déterminer quels sont les facteurs propres à la profitabilité et à la non profitabilité afin d'aider ABC à mieux cerner sa clientèle et d'établir un prix adéquat en fonction de la profitabilité qu'elle aura fixée au préalable.

Dans le chapitre 2, nous recensons les écrits qui traitent du concept de la profitabilité clientèle. Nous abordons l'importance de la profitabilité clientèle pour les entreprises. Nous analysons les différentes définitions de ce concept selon le point de vue de plusieurs chercheurs. Et nous présentons quelques modèles de profitabilité développés par les chercheurs.

Dans le chapitre 3, nous expliquons les démarches en lien avec la préparation et l'exploitation des données dont nous disposons afin d'être en mesure de les manipuler adéquatement. Nous dressons également un portrait de la clientèle de ABC. Dans cette section, nous survolons les caractéristiques qui font en sorte qu'un client est profitable ou ne l'est pas. De plus, pour approfondir un peu plus cette notion de façon exploratoire, nous effectuons des tests de moyenne sur les différentes variables.

Dans le chapitre 4, nous abordons trois modèles classiques de régression: la régression linéaire multiple, la régression gamma et la régression logistique. Nous présentons également un quatrième modèle qui est la combinaison de la régression logistique et de la régression linéaire. Dans le but de sélectionner le meilleur choix de modèle, nous discutons des différents critères de sélection

pour chacun des modèles abordés. Pour terminer, nous traitons des limites de la recherche et de la façon que ABC pourrait tirer davantage de ses données en modifiant quelques informations de son système d'informations.

Finalement, nous concluons en résumant sur ce qui a été effectué dans le cadre de ce mémoire. Et nous indiquons quelques pistes de recherche que ABC pourrait bénéficier pour accroître la profitabilité de sa clientèle résidentielle.

Chapitre 2

Revue de la littérature

Nous présentons dans ce chapitre les études reliées à l'analyse de la profitabilité de la clientèle. D'abord, nous traitons de l'importance de la profitabilité de la clientèle pour les entreprises. À la section 2, nous définissons ce qu'est la profitabilité client (« customer profitability ») pour les divers chercheurs et à la section 3, nous présentons différents modèles de profitabilité. Enfin, certaines applications de ces modèles sont illustrées par des cas à la section 4.

2.1 Introduction : profitabilité client

Les entreprises évoluent dans un environnement d'affaires de plus en plus concurrentiel. Il devient difficile pour elles de demeurer compétitives sur le marché. Au cours de la dernière décennie, la gestion de la relation client (« customer relationship management ») prend de plus en plus d'importance. Cet outil stratégique de premier plan est de comprendre la profitabilité des clients et de retenir ceux qui sont les plus profitables (Lee et Park, 2005; Kim et al., 2006; Benoit et den Poel, 2009). Cette importance accordée à la profitabilité au cours de ces dernières années est liée à la prise de conscience

qu'il y a des différences entre les revenus et les coûts parmi les clients et les différents segments de clients (McManus, 2007).

Les chercheurs traitent la notion de profit au niveau de l'entreprise, d'une unité d'affaires, du produit ou de la marque (Anderson, Formell et Lehmann, 1994). La mesure de la profitabilité au niveau du client est moins étudiée. Il est souvent difficile d'obtenir des informations précises sur les comportements d'achat du client comme des données précises sur le client et les coûts marketing au cours d'une période (Mulhern, 1999). Blattberg et Deigthon (1991) ne partagent pas le même point de vue. Pour eux, l'analyse de la profitabilité est désormais accessible, car les bases de données clients contiennent l'historique des transactions du client. Aujourd'hui, grâce aux outils technologiques performants, la profitabilité client prend donc de l'ampleur (Gleaves et al. 2008).

2.2 Concept de profitabilité et ses termes

De plus en plus d'études portent une attention sur la question spécifique de la profitabilité client et de la valeur à vie du client (« customer lifetime value ») (McManus et Guilding, 2008). Cependant, une confusion existe dans la terminologie. À travers ses recherches, Mulhern (1999) a identifié sept termes reliés à la profitabilité client : valeur à vie (« lifetime value »), valeur à vie du client (« customer lifetime value »), valeur du client (« customer valuation »), évaluation de la valeur du client (« customer lifetime valuation »), évaluation de la relation client (« customer relationship valuation »), capital client (« customer equity ») et profitabilité client (« customer profitability »). Il note que ces termes sont interchangeables. Cette confusion mène des chercheurs à clarifier le point sur la définition des termes. Notons le cas de Pfeifer, Haskins et Conroy (2005) qui soulèvent la confusion grossière de deux des plus importants concepts en marketing: profitabilité client et valeur à vie du client.

Gleaves et al. (2008) constatent que les notions de profitabilité client, valeur à vie du client et capital client ne sont pas définis clairement et sont même contradictoires.

Face à cette confusion des termes, faisons un survol des différentes définitions utilisées par les chercheurs au fil du temps et soulignons les différences et les similitudes de chacune d'elles. Le tableau 2-1 dresse le sommaire de ces différentes définitions.

Chercheur(s)	Valeur à vie (« lifetime value »)	Profitabilité client (« customer profitability »)	Capital client (« customer equity »)
Blattberg et Deigthon (1996)			La contribution future de chaque client en actualisant les contributions à la valeur actuelle nette et en additionnant toutes les valeurs actuelles nettes des contributions.
Berger et Nars (1998)	La valeur à vie est le profit. Le profit se définit par la différence entre les revenus et les coûts (d'attraction, de vente et de services) actualisés à la valeur actuelle au cours de la relation.		La différence entre les revenus et les coûts (d'attraction, de vente, de services et d'acquisition) actualisés à la valeur actuelle au cours de la relation. Différence avec la valeur à vie : l'ajout des coûts d'acquisition.
Verhoef et Donkers (2001)	Les deux dimensions pour évaluer la valeur à vie : valeur courante et valeur potentielle. La valeur courante est la mesure des profits antérieurs. La valeur potentielle est la mesure des ventes additionnelles.		

Chercheur(s)	Valeur à vie (« lifetime value »)	Profitabilité client (« customer profitability »)	Capital client (« customer equity »)
Jain et Singh (2002)	La valeur à vie est le profit net. Le profit net est la différence entre les revenus et les coûts (d'attraction, de vente, de services, d'acquisition et de rétention) actualisés à la valeur actuelle au cours de la relation. Différence avec Berger et Nars (1998) : l'ajout des coûts d'acquisition et de rétention.		
Hwang, Jung et Suh (2004)	Les trois dimensions pour évaluer la valeur à vie : valeur courante, valeur potentielle et loyauté afin de considérer la défection du client. La loyauté est la mesure de rétention. Différence avec Verhoef et Donkers (2001) : l'ajout d'une troisième dimension, la loyauté.		

Chercheur(s)	Valeur à vie (« lifetime value »)	Profitabilité client (« customer profitability »)	Capital client (« customer equity »)
Gupta et Lehmann (2005)	La valeur actuelle de tous les profits courants et futurs générés par le client au cours de la relation.		
Pfeifer, Haskins et Conroy (2005)	La valeur actuelle des flux de trésoreries au cours de la relation. * Les auteurs emploient des termes financiers tels que flux de trésoreries en opposition à profit net. Les flux de trésoreries réfèrent à ce qui peut être actualisé. Les profits considèrent les coûts contrairement aux flux de trésoreries (ex : dépréciation d'une flotte de camions). * Leur définition est large, car ils laissent l'utilisateur décider quels sont les flux de trésoreries à considérer.	La différence entre les revenus générés et les coûts associés au cours de la relation. * Leur définition est large, car ils n'imposent pas lesquels des coûts sont à considérer ou à exclure indépendamment qu'ils impliquent ou non les flux de trésoreries.	

Chercheur(s)	Valeur à vie (« lifetime value »)	Profitabilité client (« customer profitability »)	Capital client (« customer equity »)
Lee et Park (2005)		Un indicateur de mesure du client en ce qui attrait aux revenus qu'il rapporte moins les coûts générés au cours d'une période. Pour considérer la profitabilité d'un segment de clients, ils considèrent le niveau de satisfaction, l'intention de racheter et la bouche à oreilles (obtenus par le sondage) ainsi que la perte ou le profit généré par le client (avec les données comptables).	
Kim et al. (2006)	Les trois dimensions pour évaluer la valeur à vie : valeur courante, valeur potentielle et loyauté afin de considérer la défection du client. Différence avec Verhoef et Donkers (2001) : l'ajout d'une troisième dimension, la loyauté. Similitude avec Hwang, Jung et Suh (2004) : la même définition.		

Tableau 2-1 : Sommaire des différentes définitions du concept de profitabilité

Ainsi, différents chercheurs interprètent à leur façon le concept de la profitabilité client, de la valeur à vie du client et du capital client, ce qui peut amener deux études à des résultats divergents. Jain et Singh (2002) éclaircissent la situation en les regroupant en trois points principaux :

1. Le développement de modèles de la valeur à vie pour chaque client. Les modèles incluent les sources de revenus, les coûts d'acquisition, de rétention et d'autres coûts marketing.
2. L'analyse de base du client. C'est l'analyse de l'information et les calculs de probabilité des futures transactions.
3. L'application du concept sur la gestion des décisions marketing et les effets sur la profitabilité.

De leur côté, McManus et Guilding (2008) classent la profitabilité client en deux catégories. La première est une mesure similaire à la notion de profit employée en comptabilité. Il s'agit d'une mesure historique par rapport à une période antérieure. La deuxième concerne la profitabilité future du client.

À la lumière des différentes définitions utilisées par les chercheurs pour un même concept, nous constatons à quel point il est important de bien définir les termes pour éviter toute confusion. Lors de l'élaboration de notre modèle dans le chapitre 4, nous allons définir clairement ce qu'est la profitabilité client.

2.3 Modèles

Tel que souligné précédemment, de nombreuses confusions existent dans la terminologie de la profitabilité client. Dans la section qui suit, nous allons montrer les différents modèles employés pour calculer cette profitabilité au fil du temps.

La façon traditionnelle d'établir la profitabilité client est de segmenter les clients selon le profil démographique, les attitudes et les attributs psychographiques du client (Griffin, 2003) ou par la règle 80/20. Les résultats de la segmentation sont simplistes et peu précise. La règle de 80/20 stipule que 80% des profits sont générés par les 20% des clients les plus profitables. À l'inverse, 80% des coûts sont attribuables aux 20% des clients les moins profitables (Duboff, 1992; Gloy, Akridge et Preckel, 1997).

Aujourd'hui, il existe des outils à la fine pointe de la technologie qui permettent aux systèmes d'informations de collecter des données sur les transactions et les comportements d'achats des clients (Lee et Park, 2005). Ces données permettent de mesurer la profitabilité client.

2.3.1 Modèles de base

Jain et Singh (2002) rapportent que le modèle de base de la valeur à vie du client est facile à utiliser, car il suppose que les flux de trésorerie sont les mêmes à chaque période. Le modèle est applicable aux clients actuels et ne tient pas compte des coûts d'acquisition. Le modèle se définit ainsi :

$$CLV = \sum_{t=1}^{T} \frac{(R_t - C_t)}{(1+d)^{t-0,5}},$$

où

t : la période des flux de trésoreries des transactions du client,
R_t : le revenu du client à la période t,
C_t : le total des coûts pour générer des revenus à la période t,
d : le taux prédéterminé d'actualisation,
T : le nombre total de périodes.

Le modèle de la valeur à vie du client de Berger et Nars (1998) se traduit comme suit :

$$CLV_{i,x} = \sum_{t=0}^{T} \frac{\pi_{i,x+t}}{(1+d)^t},$$

où

i : le client i,
t : le client i à la période t,
d : le taux prédéterminé d'actualisation,
x : la période à partir de laquelle nous voulons calculer le CLV,
T : le nombre total de périodes.

Dans les industries multiservices, le profit est défini :

$$\pi_{i,t} = \sum_{j=1}^{J} v_{ij,x} * u_{ij,x} * \pi_{j,x},$$

où

i : le client i,
j : le service j,
J : le nombre total des différents services vendus,
$v_{ij,x}$: l'indicateur si le client i achète le service j à la période t,
$u_{j,x}$: le montant du service acheté,
$\pi_{j,x}$: le profit moyen pour le service j.

2.3.2 Modèles Pareto/NBD

Une autre approche du calcul de la valeur à vie du client est celle introduite par Schmittlein, Morrison et Colombo (1987), le modèle Pareto/NBD. Le modèle calcule la probabilité que le client est encore actif. Il est applicable lorsqu'il n'y a pas de relation contractuelle entre l'entreprise et le client. Trois comportements d'achat antérieurs sont requis pour chaque client :

1. « cohorte » : la date d'entrée du client jusqu'à ce jour,
2. « fréquence » : la fréquence des achats,
3. « récence » : la période entre la date d'entrée et le dernier achat.

Le modèle Pareto/NBD de Schmittlein, Morrison et Colombo (1987) est le suivant :

$$P(actif / r,s,\alpha,g,\beta,t,t) = \left\{1 + \frac{s}{r+g+s} * \left[\begin{array}{l} \left(\frac{\alpha+T}{\alpha+t}\right)^{r+g}\left(\frac{\beta+T}{\alpha+t}\right)^{s} F(a_1 b_1; c_1; z_1(t)) \\ -\left(\frac{\beta+T}{\alpha+T}\right)^{s} F(a_1, b_1; c_1; z_1(T)) \end{array}\right]\right\}^{-1},$$

où

$a_1 = r + g + s$,

$b_1 = s + 1$,

$c_1 = r + g + s + 1$,

$z_1(t) = \alpha - \beta / \alpha + t$,

r,s,α , β : les paramètres du modèles à déterminer par des tests préliminaires (où $\alpha > \beta$),

r : le taux d'achat,

s : le taux de clients inactifs,

t : la période écoulée depuis la transaction la plus récente,

T : la période écoulée depuis la première transaction,

$F(a_1, b_1; c_1; z)$: la fonction hypergéométrique Gauss,

g : le nombre d'achats effectués par le client.

À travers ce modèle, les auteurs démontrent qu'il est possible de connaître le nombre de clients que l'entreprise possède, la croissance de la clientèle au cours de la dernière année, lesquels des clients sont probablement actifs ou inactifs. Le modèle fournit le nombre de transactions que l'entreprise peut s'attendre l'année suivante pour un client spécifique et l'ensemble de la clientèle.

Reinartz et Kumar (2000) suggèrent une extension du modèle de Pareto/NBD en incorporant la probabilité que le client est actif ou inactif en dichotomie. L'application est dans un contexte où il n'y a pas de contrat.

Glady, Baesens et Croux (2009) proposent une modification de l'approche Pareto/NBD pour prédire la valeur à vie du client, le modèle Pareto/Dépendant. Ils prouvent qu'il y a une dépendance entre le nombre de transactions et le profit moyen par transaction. Cette dépendance augmente la précision de la prédiction. La valeur monétaire du client dépend du nombre de transactions qu'il effectue. La valeur peut être différente selon le cas de chacun des clients. Un client avec une grande probabilité de transactions futures peut engendrer une grande valeur monétaire, mais l'inverse est aussi possible.

$$CLV_{i,T} = \sum_{t=1}^{T} \frac{x_{i,hi+t}\, m_{i,h_{i}+t} - x_{i,h_{i+t-1}}\, m_{i,h_i+t-1}}{(1+d)^t},$$

où

i : le client i,

t : le client i à la période t,

d : le taux prédéterminé d'actualisation,

x_i : le nombre de transactions effectuées par le client i à ce jour,

m_i : le profit moyen par transaction du client i à ce jour.

h : la période à laquelle le client *i* a effectué sa première transaction,
T : le nombre total de périodes.

2.3.3 Régression quantile

Benoît et den Poel (2009) recommandent d'analyser la valeur à vie du client par la régression quantile. Ils sont les premiers à utiliser ce modèle dans le contexte de modélisation de la valeur à vie du client. Les auteurs se réfèrent à Koenker et Basset (1978) et Koenker (2005) qui affirment que la régression quantile est une extension de la régression moyenne. En régression linéaire, la méthode des moindres carrés donnent des estimations de la moyenne conditionnelle. Quant à la régression quantile, elle estime les quantiles conditionnels de la variable réponse, en se basant sur la médiane. Ce point permet de faire la nuance entre la relation des variables explicatives et un groupe de covariables, puisqu'il permet à l'utilisateur d'étudier la relation entre les covariables et les différents quantiles de la variable réponse. Cette approche se fait dans un contexte dans lequel un contrat existe.

2.3.4 Autres modèles

Hwang, Jung et Suh (2004)

Hwang, Jung et Suh (2004) proposent la valeur à vie en considérant les profits antérieurs, les bénéfices potentiels et la défection du client. Pour eux, la valeur courante donne un point de vue financier. La valeur potentielle permet de saisir les opportunités de vente croisée. La loyauté permet d'estimer la durabilité des deux points précédents. Leur modèle s'établit selon cette équation :

$$LTV_i = \underbrace{\sum_{t_i=0}^{n_i} \pi_p(t_i)(1+d)^{n_i-t_i}}_{\text{Profits antérieurs}} + \underbrace{\sum_{t_i=n_i+1}^{n_i+e(i)+1} \frac{\pi_f(t_i)+B(t_i)}{(1+d)^{t_i-n_i}}}_{\text{Futurs flux de trésoreries attendus}},$$

où

i : le client i,

t : le client i à la période t,

d : le taux prédéterminé d'actualisation,

n_i : le nombre total de périodes que l'entreprise a servi le client i jusqu'à ce jour,

$(e)_i$: le nombre de périodes que l'entreprise s'attend à servir le client i dans le futur,

$\pi_p(t_i)$: les profits antérieurs du client i à la période t_i,

$\pi_f(t_i)$: les profits futurs du client i à la période t_i,

$B(t_i)$: les bénéfices potentiels du client i à la période t_i.

Lee et Park (2005)

Lee et Park (2005) proposent un système de sondage sur la profitabilité client (« survey-based profitable customers segmentation system »). Il s'agit d'un système basé sur le data mining et des agents technologiques. Ces derniers conçoivent et exécutent (en ligne, par courriel…) un sondage sur la satisfaction du client. Ce système mène à des procédures prédéfinies pour la segmentation de la clientèle qui est profitable. Il y a trois types d'agents dans le sondage :

1. Gestion de sondage (« Survey Management »),
2. Segmentation de la clientèle profitable (« Profitable customer segmentation »),
3. Assistant de l'utilisateur (« User assistant »).

L'agent gestion de sondage prend en charge la conception et l'exécution du sondage de satisfaction. La première étape consiste à trouver les clients qui ont le plus grand rendement par rapport aux coûts, ceux-ci sont classés dans un groupe appelé groupe de clients d'efficacité supérieure (« Higher Efficiency Customer Group »). L'agent segmentation de la clientèle profitable recourt à l'analyse d'enveloppement de données (DEA) (« Data Envelopment Analysis ») pour cette étape. L'enveloppement de données est une approche de programmation linéaire pour analyser les entrées et les sorties, appelées unités de décision. Il mesure la performance relative des unités de décisions en attribuant un poids aux unités qui donnent la meilleure efficacité. Ensuite, la carte auto-organisatrice (« Self-Organizing Map »), une méthode d'apprentissage non supervisée, est employée par l'agent segmentation de la clientèle profitable pour retirer les clients non profitables qui se sont retrouvés dans le groupe de clients d'efficacité supérieure. La carte auto-organisatrice permet de cartographier la répartition de données dans un espace à grande dimension (Kohonen, 1989).

Par la suite, l'agent segmentation de la clientèle profitable fait appel au C4.5 (Quinlan, 1993) et à la carte auto-organisatrice pour déterminer l'ordre des clients non profitables qui pourront potentiellement devenir profitables. Le C4.5 est un algorithme utilisé pour générer des arbres de décision. L'objectif des arbres est d'estimer la valeur de la variable cible, soit la profitabilité, à partir d'un ensemble de variables explicatives. Le C4.5 permet à l'agent segmentation de détecter quels sont les facteurs qui discriminent les clients profitables des clients non profitables. Finalement, l'agent assistant de l'utilisateur agit à titre d'intermédiaire avec l'utilisateur.

Verhoef et Donkers (2001)

Verhoef et Donkers (2001) proposent de considérer les valeurs sociodémographiques et le comportement d'achat du client dans leur modèle de prédiction de la valeur potentielle du client. Ils utilisent trois modèles : modèle probit multivarié (« multivariate probit model »), modèle probit univarié et le modèle de régression.

1. Modèle probit multivarié :

$$VP_i = \sum_{k=1}^{K} \Pr ob(y_{ik})\pi_k ,$$

où

VP_i : la valeur potentielle du client *i*,
y_{ik} : la détention du portfolio *k* par le client i (où y=1 : détention et y=0 : pas de détention),
k : le portefolio *k* (produit et service),
K : le nombre total de portefolios *k*,
π_k : le profit du portfolio *k*.

2. Modèle probit univarié :

Dans ce modèle, il existe une interdépendance entre les décisions d'achat. Ainsi, la probabilité de détenir le service j dépend du fait de détenir le service k.

$$VP_i = \sum_{j=1}^{J} \Pr ob\left(y_{ij} = 1\right)\pi_j ,$$

où

$$\Pr ob(y_{ji} = 1) = \Pr ob\left(\varepsilon_{ij} > \beta_j W_i - \sum_{k=1}^{J} \gamma_{jk} Z_{ik} \right),$$

VP_i : la valeur potentielle du client *i*,

j : le service j,

J : le nombre total de services j,

y : la détention du service j par le client i (où y=1 : détention et y=0 : pas de détention),

ϵ : l'erreur des résidus,

γ : le paramètre du modèle à déterminer par des tests préliminaires,

W : l'indicateur sociodémographique,

Z : la détention observée du service k par le client i.

3. Modèle de régression :

$$VP_i = \beta W_i + \sum_{k=1}^{J} \gamma_k Z_k + \varepsilon_i ,$$

où

VP_i : la valeur potentielle du client *i*,

J : le nombre total de services j,

ϵ : l'erreur des résidus,

k : le portefolio *k* (produit et service),

W : l'indicateur sociodémographique,

γ : le paramètre du modèle à déterminer par des tests préliminaires,

Z : la détention observée du portefolio k du client i.

2.4 Applications

Dans cette section, nous présentons quelques cas de modèles présentés à la section 2.3 qui sont appliqués sur des données d'entreprises.

2.4.1 Modèle de Kim et al. (2006)

Kim et al. (2006) reprennent le modèle de la valeur à vie développé par Hwang, Jung et Suh (2004). Ils appliquent ce modèle sur les données d'une compagnie de télécommunication sans-fil en Corée. Les données recueillies sont sur une période de six mois. Elles sont classées en deux : le profil sociodémographique et les informations relatives à l'utilisation sur le service du mobile.

Leur travail s'effectue en trois phases tel que représenté dans la figure 2-1. La phase 1 explique les étapes qui mènent à la définition de la valeur du client et de la mise en place des stratégies marketing. La phase 2 évalue la valeur du client selon les trois dimensions (la valeur courante, la valeur potentielle et la loyauté). Et la phase 3 consiste à analyser chacun des segments selon les trois dimensions et recommande les stratégies marketing appropriées.

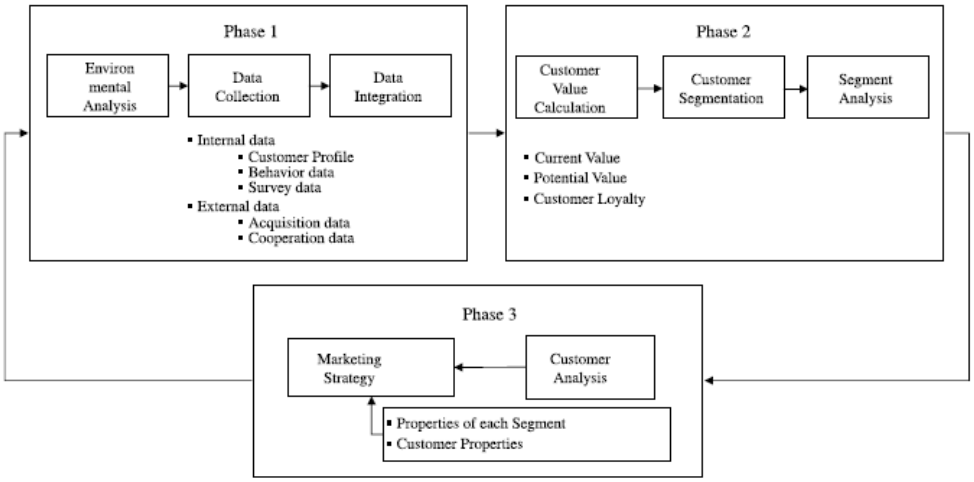

Figure 2-1 : Cadre pour la segmentation de la clientèle basée sur la valeur à vie du client

Une fois les calculs effectués pour les trois dimensions de chaque client, les auteurs les ont disposés dans un espace tridimensionnel afin de voir la distribution des clients. Chaque axe représente la valeur courante, la valeur potentielle et la loyauté. La figure 2-2 illustre le résultat de la segmentation. L'espace tridimensionnel donne le portrait global de la clientèle, mais n'est pas assez complexe pour établir des stratégies marketing. Comme il est difficile de déterminer laquelle des dimensions définit le mieux le client, les segments sont étudiés selon le score accordé aux trois dimensions. Le tableau 2-2 identifie les caractéristiques des segments selon le score des dimensions. Pour obtenir les caractéristiques de ces segments, les auteurs ont utilisé l'arbre de décision pour discriminer les variables.

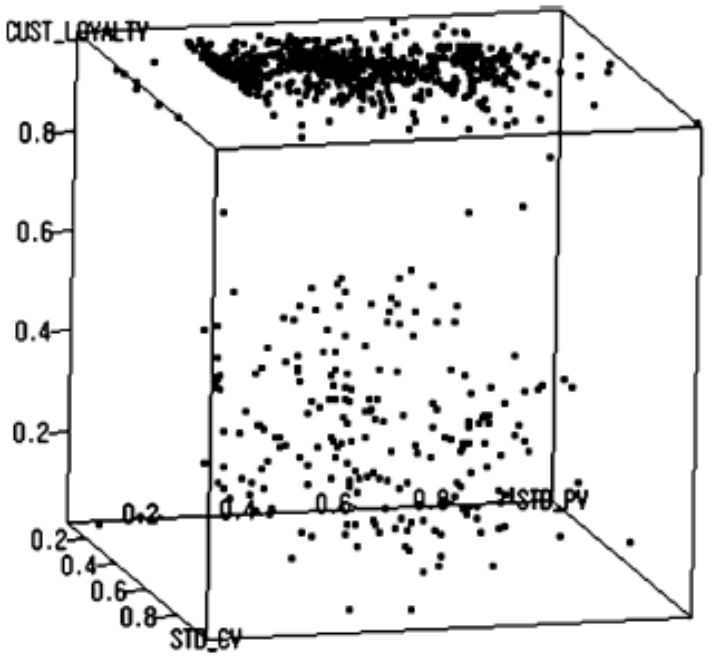

Figure 2-2 : Résultat de la segmentation clientèle selon les trois dimensions

	Segment 1	Segment 2	Segment 3
Dimension / score	Faible loyauté	Valeur courante élevée	Valeur potentielle élevée
Variables discriminantes	1 - Mode de paiement pour les frais d'adhésion. 2 - Durée du contrat. 3 - Plan de paiement.	1 - Écart-type du coût moyen d'usage. 2 - Durée d'interruption d'un service. 3 - Changement de forfait. 4 - Réception du mobile.	1 - Âge du client. 2 - Type d'appareils.
Description selon les variables	1 - 90% des clients ne paient pas les frais d'adhésion. 2 - Durée de contrat inférieur à 18 mois. 3 - Pas de plan de paiement.	1 - Écart-type du coût moyen d'usage > 20862, interruption du service >8 jours, changement de forfait <1. 2 - Écart-type du coût moyen d'usage < 20862, pas de demande de changement de forfait ni de meilleure réception du mobile.	1 - 27,5 ans et moins habitant dans la région de Séoul et Incheon. 2 - 27,5 ans et plus avec un type d'appareil spécifique et dont la durée de contrat est courte.

Tableau 2-2 : Caractéristiques des trois segments selon le score des dimensions

À la lumière des résultats obtenus, les auteurs recommandent cinq stratégies que nous rapportons dans le tableau 2-3.

Stratégies	Description
1 - Charger des frais à l'adhésion	Charger des frais lors de l'adhésion est un moyen de garder le client à continuer d'utiliser les services de l'entreprise. Ne pas charger de frais implique que ces frais sont transférés sur d'autres coûts, ce qui peut mener le client à se tourner vers la concurrence.
2 - Surclasser les appareils	Pour les clients dont la valeur courante et la valeur potentielle sont élevées : offrir un programme de récompenses qui permet d'échanger les points contre un surclassement d'appareil. Ces clients sont plus susceptibles d'ajouter des options supplémentaires à leur forfait.
3 - Attirer les clients en fonction de leur âge et occupation	Créer des plans de service attrayants. Le plan de service optionnel est populaire auprès des jeunes.
4 - Offrir de meilleurs services pour les clients loyaux	Offrir des options gratuites et des rabais sur la facture aux clients loyaux, ceci les convaincra de demeurer avec l'entreprise grâce aux bénéfices conçus spécialement pour eux.
5 - Renforcer l'image de marque	Développer une image de marque forte selon la clientèle visée. L'image de marque autant que le service offert sont des atouts de plus en plus importants pour les clients.

Tableau 2-3 : Stratégies marketing

2.4.2 Modèle de Lee et Park (2005)

Lee et Park (2005) implantent leur système de sondage sur la profitabilité client avec les données de la compagnie T Motor. L'agent gestion de sondage exécute un sondage par courriel pour lequel 491 clients y ont répondu. Le sondage comporte 24 questions abordant la qualité du produit, le service à la clientèle, l'intention de racheter etc. L'agent segmentation de la clientèle profitable exécute le DEA pour sélectionner les clients qui ont la plus grande profitabilité, ils sont appelés le groupe de clients d'efficacité supérieure. Dans l'étude, ce groupe doit avoir un score supérieur à 0,95 lors du DEA, ce qui

représente un total de 88 clients. Ensuite, l'agent segmentation de la clientèle profitable fait deux fois le filtrage (« screening ») afin de n'obtenir qu'un groupe de clients profitables (GCP) (« profitable customer group »). Finalement, l'agent segmentation de la clientèle profitable génère des règles de décisions à travers le C4.5 afin de déterminer quelles sont les règles communes de ces clients profitables. Dans le cas de l'étude, nous retrouvons trois règles sociodémographiques tel qu'indiqué dans le tableau 2-4. Ces règles sont importantes car elles permettent d'identifier les clients non profitables qui sont susceptibles de se retrouver dans le groupe de clients profitables. Pour ce faire, il s'agit d'identifier lesquels des onze facteurs de l'étude (design, couleur, confort, bruit du moteur, prix, service après-vente…) ont classé les clients appartenant au groupe de clients profitables. Il y en a quatre : le confort de la conduite, le bruit du moteur, l'économie d'essence et le design. Ensuite, les clients qui ne font pas partie du groupe de clients profitables passent à travers la carte auto-organisatrice qui va les classifier. Seuls ceux ayant un score supérieur à 0,9 sont retenus pour au moins un des facteurs discriminants. Il y a alors trois clients qui peuvent se retrouver dans le groupe de clients profitables tel que représenté dans le tableau 2-5. L'agent segmentation de la clientèle profitable décide de la priorité de ces trois clients selon leur résultat. Celui qui a le résultat le plus élevé a une plus grande probabilité que les autres de se retrouver dans le groupe de clients profitables. Selon les résultats, le client #1 a la plus grande probabilité d'être parmi le groupe de clients profitables.

Règles sociodémographiques	
#1	Homme marié moins de 43 ans.
#2	Homme célibataire ou conjoint de fait âgé entre 36 et 50 ans.
#3	Homme célibataire ou conjoint de fait âgé de plus de 56 ans.

Tableau 2-4 : Règles sociodémographiques du groupe de clients profitables

Numéro de client	Classification de la carte auto-organisatrice	Score
#1	Bruit du moteur Économie d'essence	0,9288 0,9113
#27	Bruit du moteur	0,9075
#333	Bruit du moteur Économie d'essence	0,9178 0,9102

Tableau 2-5 : Résultat de la classification des clients qui ne font pas partie du GCP

Lee et Park (2005) concluent que leur système de sondage sur la profitabilité client est une approche efficace et simple. Il est basé sur le sondage de satisfaction, les données sociodémographiques et le data mining. Ils prouvent dans leur cas qu'il est préférable d'utiliser leur modèle que de recourir à un modèle complexe de profitabilité clientèle.

2.4.3 Modèle de Verhoef et Donkers (2001)

Verhoef et Donkers (2001) proposent d'inclure les valeurs sociodémographiques et le comportement d'achat du client dans les modèles : probit multivarié, probit univarié et régression. Le modèle naïf (« naive model ») est celui qui sert de base pour la comparaison, il n'inclut pas les valeurs sociodémographiques et le comportement d'achat du client. Tous ces modèles sont appliqués dans le cas d'une entreprise d'assurance dans les Pays-Bas afin de prédire la valeur potentielle du client. Cette compagnie détient huit types d'assurances allant de l'assurance vol à l'assurance-vie. 2 300 clients ont été sondés, 1 612 clients sont retenus pour l'analyse. L'échantillon a été séparé en deux : 1 000 clients pour l'entrainement et 612, pour la validation. Le but de l'étude est d'estimer le profit potentiel des clients. Avec les données de profit sur le type de l'assurance et des prédictions de taux de probabilité de détention de l'assurance, il est possible d'évaluer le profit potentiel.

Pour chaque type d'assurance, le tableau 2-6 représente le taux de bonne classification par rapport à l'achat de l'assurance dans l'échantillon de validation pour les deux modèles probit. Pour le modèle naïf, les résultats sont obtenus par l'échantillon d'entraînement. Les résultats démontrent que tous les modèles prédisent correctement à plus de 50%. Pour certains types d'assurance (#7,8,11), le modèle naïf performe mieux que les modèles probit. De façon globale, les modèles probit performent mieux que le modèle naïf. Les variables sociodémographiques telles que l'âge, le revenu, le statut matrimonial et la propriété de la résidence sont des prédicteurs pertinents pour la prédiction de l'achat de l'assurance.

Type d'assurance / Modèle	Probit univarié (%)	Probit multivarié (%)	Naïf (%)
5	0,894	0,899	0,892
6	0,758	0,755	0,733
7	0,651	0,657	0,658
8	0,621	0,621	0,635
9	0,655	0,650	0,547
10	0,503	0,503	0,464
11	0,556	0,542	0,577
12	0,634	0,636	0,570

Tableau 2-6 : Taux de bonne classification

Puisque le but de la recherche est d'estimer la valeur potentielle du client, le critère d'erreur absolue moyenne de prédiction (EAMP) (« Mean Absolute Prediction Error ») est utilisé à cette fin. Les modèles probit et naïf donnent une erreur similaire tel que décrit le tableau 2-7. Les modèles probit performent légèrement mieux que le modèle naïf. D'une perspective managériale, la segmentation des clients peut être basée sur le profil sociodémographique des clients à l'aide des modèles probit. En appliquant les modèles pour la segmentation basée sur la valeur potentielle, nous retrouvons deux segments : valeur potentielle élevée et valeur potentielle faible qui sont représentés dans le tableau 2-8. Pour des raisons de confidentialité, le profit a été indexé. La moyenne du profit est de 100. Le segment de valeur potentielle faible a une profitabilité plus basse de 4%-5% que la

moyenne alors que le segment de valeur potentielle élevée a une moyenne de 4% de plus.

	Probit univarié	Probit multivarié	Régression	Naïf
Erreur absolue moyenne de prédiction	19,5%	19,4%	19,4%	20,5%

Tableau 2-7 : Erreur absolue moyenne de prédiction

Segment		Probit choix	Probit multivarié	Régression	Probit segment
Valeur potentielle élevée	Moyenne Écart-type n	104,0 20,1 311	104,0 19,7 308	104,0 19,9 318	101,6 20,3 326
Valeur potentielle faible	Moyenne Écart-type n	97,4 21,4 301	96,0 21,2 304	95,6 20,9 294	98,1 21,3 286
Taux de bonne classification		53,1%	54,6%	55,9%	51,6%

Tableau 2-8 : Profit indexé et taux de bonne classification selon les modèles étudiés

Dans l'étude, les modèles probit donnent des résultats similaires et ne performent que très légèrement mieux par rapport au modèle naïf pour la modélisation du taux de détention de l'assurance. Quant à la prédiction de la valeur potentielle, il est plus approprié d'utiliser la régression car elle donne de meilleurs résultats. Son taux de bonne classification est supérieur (55,9%) par rapport aux autres. En conclusion, les auteurs affirment qu'il n'y a pas de raison théorique qu'un modèle performe mieux qu'un autre.

2.4.4 Modèle de Glady, Baesens et Croux (2009)

Glady, Baesens et Croux (2009) appliquent leur modèle Pareto/Dépendant sur des données de ING Belgique dont la période est de janvier 2000 à décembre

2005. Pour des fins de comparaison, ils appliquent les mêmes données sur le modèle Pareto/Indépendant et la régression linéaire. La base de données comprend un total de 11 068 877 transactions effectuées par 460 566 clients. Les clients sont classés en huit cohortes basées sur les trimestres de 2001 et 2002. Chaque cohorte débute au début des huit trimestres de la période étudiée.

Dans un premier temps, ils étudient le cas pour l'ensemble des huit cohortes. L'erreur quadratique moyenne (EQM) (« root mean square error ») et l'erreur absolue moyenne (EAM) (« mean absolute error ») sont utilisées comme mesure de prédiction. La corrélation de Spearman entre la valeur prédite et la valeur réelle est employée comme mesure complémentaire, car elle est plus robuste que celle de Pearson face aux données extrêmes. Le tableau 2-9 présente les résultats des mesures de prédiction. Le modèle Pareto/Dépendant est certainement le meilleur modèle des trois pour chaque indicateur.

Modèle	Erreur quadratique moyenne	Erreur absolue moyenne	Corrélation (%)
Pareto/Indépendant	946,2	411,9	40,5
Régression linéaire	892,7	340,5	47,9
Pareto/Dépendant	843,4	324,0	51,8

Tableau 2-9 : Résultat des mesures de prédiction

Dans un deuxième temps, Glady, Baesens et Croux (2009) étudient les huit cohortes séparément afin de voir comment les modèles se comportent dans le temps. Le premier constat est que les trois mesures de prédictions performent mieux avec le modèle Pareto/Dépendant et la régression linéaire que le modèle Pareto/Indépendant pour presque chacune des cohortes. Le deuxième constat est que la corrélation de Spearman est beaucoup plus élevée pour le modèle Pareto/Dépendant que la régression linéaire. Pour ce qui est de l'analyse individuelle des huit cohortes, le modèle Pareto/Dépendant demeure le meilleur modèle parmi les trois modèles étudiés. Les résultats des huit cohortes avec les mesures de prédictions (où RMSE est l'erreur quadratique moyenne et MAE est l'erreur absolue moyenne) sont résumés dans les figures 2-3, 2-4 et 2-5.

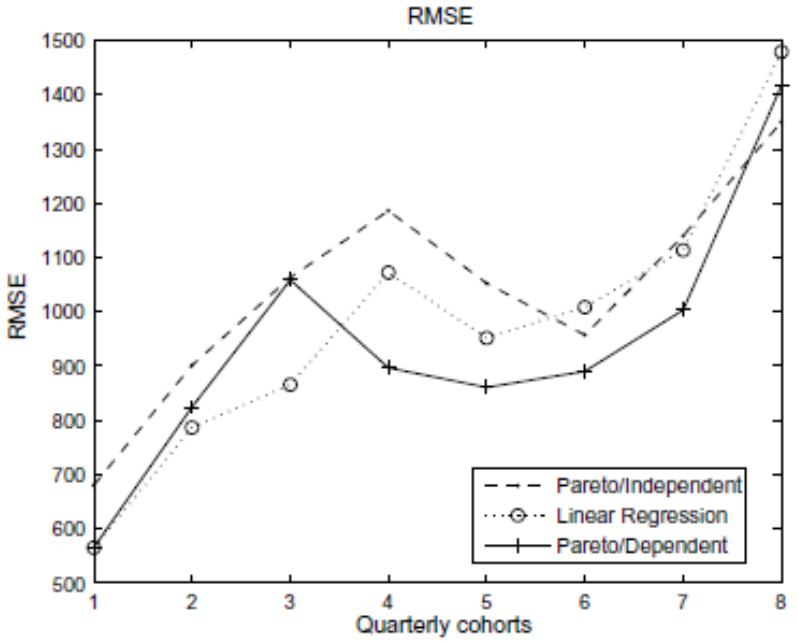

Figure 2-3 : Erreur quadratique moyenne selon les trois modèles

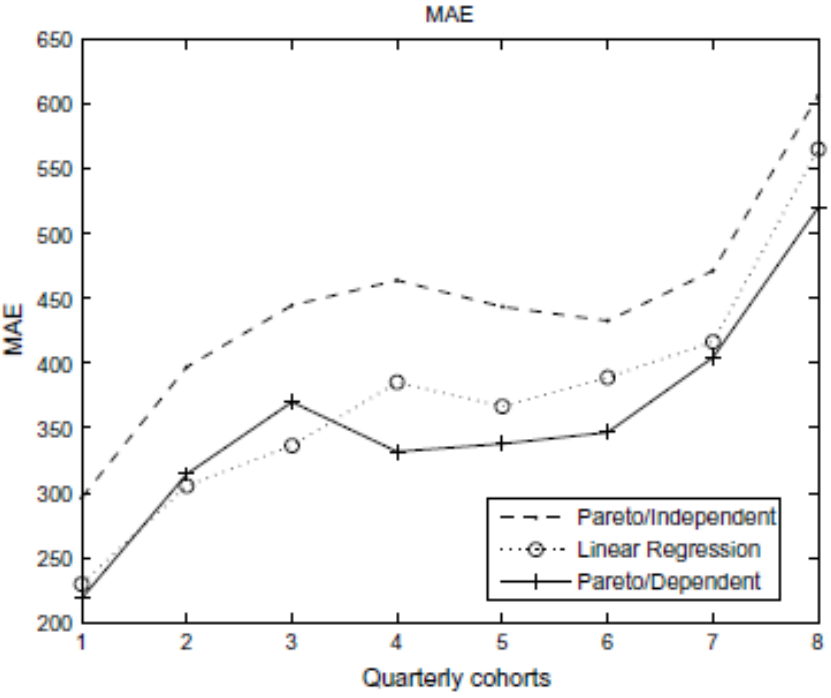

Figure 2-4 : Erreur absolue moyenne selon les trois modèles

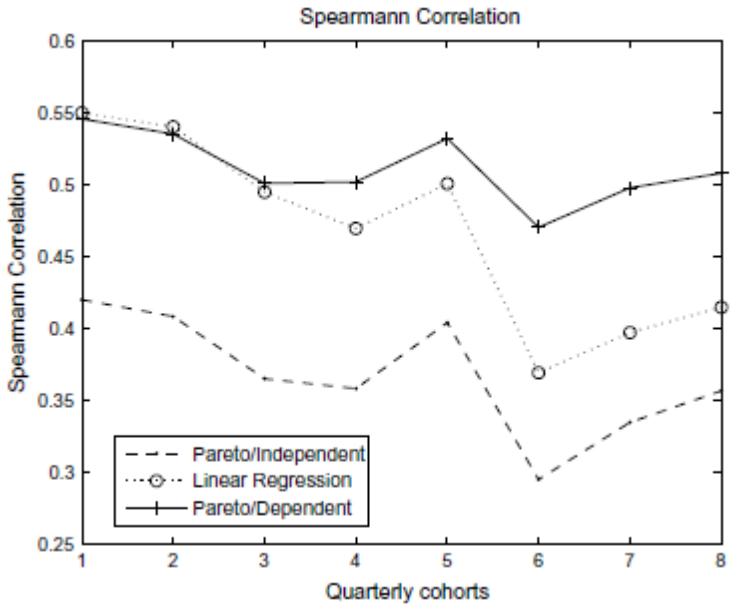

Figure 2-5 : Corrélation de Spearman selon les trois modèles

En résumé, Glady, Baesens et Croux (2009) ne peuvent affirmer que le modèle Pareto/Dépendant est mieux que le modèle Pareto/Indépendant en se basant sur les mesures de prédiction. Toutefois, dans l'étude qu'ils ont effectuée, leur modèle est plus performant. Même si la corrélation demeure faible entre le nombre de transactions et le profit moyen par transaction, cette dépendance n'améliore pas considérablement le modèle, mais au moins il ne le détériore pas.

2.4.5 Modèle de Benoît et den Poel (2009)

Finalement, le modèle de régression quantile développé par Benoît et den Poel (2009) pour évaluer la valeur à vie du client est appliqué sur les données d'une compagnie européenne de services financiers. Des données sur les transactions et le profil sociodémographiques ont été recueillies sur 22 665 ménages ayant au moins un produit actif. Les données sont divisées en deux : 60% pour

l'entraînement et 40% pour la validation. L'étude se fait en deux étapes. La première consiste à évaluer la valeur à vie du client. La deuxième étape est de prédire l'ordre des clients en se basant sur la valeur à vie du client.

En ce qui concerne l'évaluation de la valeur à vie du client, Benoît et den Poel (2009) comparent leur modèle de régression quantile à la régression linéaire et au modèle naïf. Le modèle naïf ne contient pas les variables explicatives. Le tableau 2-10 présente les résultats du taux de succès de ces trois modèles. La moyenne des erreurs au carré (« mean square error ») et l'écart absolu moyen (« mean absolute deviation ») ne sont pas des critères appropriés pour comparer les modèles de régression. Les modèles de régression tendent à optimiser l'un de ces deux critères. Dans l'étude, lorsque le critère de la moyenne des erreurs au carré est employé, la régression linéaire obtient de meilleurs résultats. Si c'est le critère de l'écart absolu moyen, la régression quantile le surpasse. Afin de bien juger les modèles, Benoît et den Poel (2009) adoptent le taux de succès (« hit rate ») comme critère proposé par Donkers, Verhoef et de Jong (2007). Pour obtenir le taux de succès, tous les clients sont classés selon leur valeur à vie du client en quatre groupes. Le taux de succès est le ratio entre le nombre de clients prédits correctement et le nombre total de prédiction.

	Régression linéaire	Régression quantile (r=0,5)	Naïf
Taux de succès	36,39%	37,86%	26,84%

Tableau 2-10 : Taux de succès pour les différents modèles étudiés

Quant à la prédiction de l'ordre des clients basé sur la valeur à vie du client, le résultat de la régression quantile et linéaire est présenté dans la figure 2-6. L'axe horizontal représente le pourcentage des meilleurs clients basés sur la valeur à vie. L'axe vertical est le taux de succès. Tel qu'illustré dans la figure, l'avantage d'utiliser la régression quantile est pertinente lorsque nous sommes intéressés par l'ordonnancement des meilleurs clients. Dans l'étude, il y a un

plus gros écart entre la régression linéaire et quantile lorsque nous sommes intéressés par les 5% meilleurs clients que les 20% meilleurs clients.

Leur constat est que même si divers études démontrent que la régression moyenne est la meilleure technique pour la prédiction; dans le cas de leur recherche, la régression quantile performe mieux tant au niveau de la prédiction que de l'ordonnancement du client basé sur la valeur à vie du client.

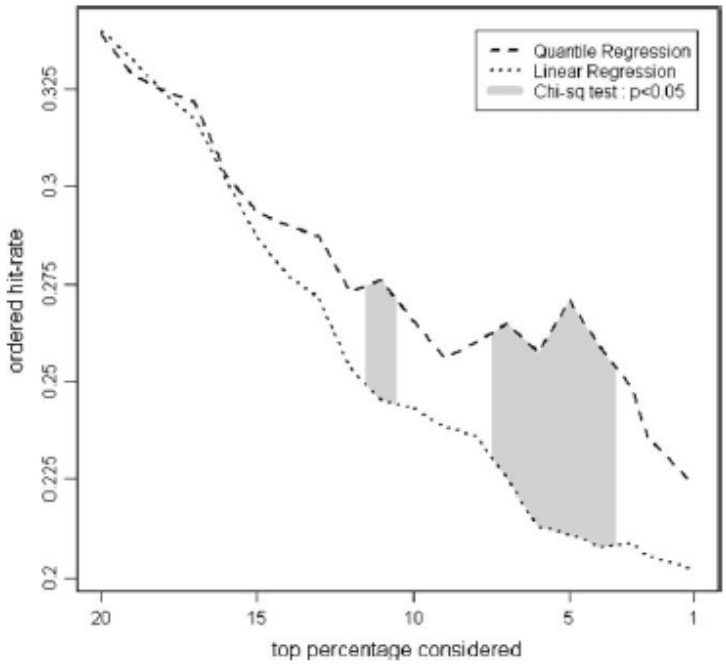

Figure 2-6 : Prédiction de l'ordonnancement du client basé sur la valeur à vie du client

Chapitre 3

Préparation et exploration des données

Nous présentons dans ce chapitre la préparation et l'exploration des données. D'abord, nous traitons du contenu de la base de données. À la section 2, nous abordons la préparation des données, de la façon dont nous les avons épurées et des règles de standardisation que nous avons établies. Ensuite, à la section 3, nous présentons le portrait de la clientèle à l'aide des statistiques descriptives. Enfin, nous quantifions la profitabilité en termes monétaire à l'aide des tests de profitabilité moyenne.

3.1 Description des données

3.1.1 Source de données

Les données de l'étude proviennent de ABC. L'historique disponible remonte en avril 2004, la date d'implantation du système client ABCSoft.

L'information fournie a été extraite en juillet 2009. Tous les clients résidentiels ayant un produit (programme de protection et/ou location) actif à cette date se retrouvent dans la base de données.

La base de données ne contient pas les données sur les anciens clients, c'est-à-dire ceux qui ont quitté l'entreprise avant juillet 2009. Par conséquent, nous ne sommes pas en mesure de déterminer les facteurs de la profitabilité de la clientèle dans son ensemble : les clients actuels et les anciens clients. Les résultats risquent d'être biaisés car ce sont deux profils de clientèle dont le comportement peut différer et expliquer chacun à sa façon, les variables de la profitabilité. Dans notre cas, nous basons notre analyse uniquement sur les clients actuels.

3.1.2 Variables utilisées

Il y a près d'une cinquantaine de variables dans la base de données. Puisqu'il y a deux produits associés à un client, nous n'en étudierons qu'un seul. Toutes les variables qui se rattachent au produit non étudié ne sont pas considérées. Nous expliquons davantage notre décision à la section 3.2.2. De plus, l'ensemble des 19 interventions effectuées chez le client ont été agrégées sous une seule variable « nombre d'interventions ».

La base de données comporte des informations confidentielles sur le client de ABC. Les informations personnelles telles que le nom, l'adresse, le profil sociodémographique n'ont pas été extraites. Seules les données en lien avec le produit sont disponibles.

Les données suivantes englobent toutes les informations pertinentes au produit:

- Nombre d'années-client depuis avril 2004,
- Revenus (du produit et de la facturation supplémentaire pour une intervention non comprise dans le produit),
- Coûts directs (fixés par le directeur de la Finance) où :
    - Main-d'œuvre (salaire, avantages sociaux…) : 55$ / heure / intervention,
    - Accessoires reliés aux interventions (mobile, essence, camion…) : 40$ / heure / intervention,
    - Communication : 0.25$ / produit actif / année,
    - Uniforme : 0.88$ / client / produit actif / année,
    - Transport : 75% du coût de l'intervention,
    - Note : Chacun de ces coûts directs sont présentés de façon agrégée dans la base de données. Il s'agit du coût total pour un produit actif.
- Profitabilité,
- Profil de l'équipement : usage chauffe-eau ou usage chauffage,
- Type de l'équipement,
- Marque, modèle et numéro de série de l'équipement,
- Date de l'installation,
- Type d'interventions sur l'équipement,
- Nombre d'interventions sur l'équipement.

3.2 Préparation des données

3.2.1 Épuration des données

Le nettoyage de la base de données est un facteur clé pour obtenir un modèle de qualité et donc, une interprétation adéquate des résultats. Un effort

considérable a été entrepris afin de rendre les données uniformes et cohérentes entre elles afin de pouvoir les exploiter. Les différentes manipulations ont été effectuées à l'aide du logiciel SPSS et SAS.

Dans un premier temps, les requêtes de fréquences ont permis de visualiser chaque modalité de chacune des variables. Ce faisant, des modalités ont été modifiées; certaines ont été recodées et d'autres, regroupées. La recodification des variables a été une étape importante du nettoyage, car plusieurs appellations existaient pour signifier le même terme. Les libellés ont été uniformisés. Par exemple, le type de l'équipement « air chaud » peut être écrit sous différentes formes : « airchaud », « promo air chaud », « AirChaud », etc. Pour réduire le nombre de modalités, certaines variables ont nécessité un regroupement. Les équipements de chauffage ont été regroupés sous deux catégories : « air chaud » et « eau chaude ». Finalement, certaines modalités ont été retirées de la base de données. Celles qui représentent les équipements relatifs à la clientèle commerciale ont été rejetées car l'étude se base uniquement sur les clients résidentiels (exemple : les aérothermes). Toutefois, certains équipements commerciaux sont conservés tels que le chauffe-eau commercial, car le chauffe-eau résidentiel peut s'avérer insuffisant pour répondre à la demande dans une résidence dont la superficie est très grande. Le profil de l'équipement hybride a aussi été retiré, car nous ne possédons que 23 observations.

À titre d'exemple pour le regroupement des variables, chaque type d'interventions (entretien, mise en marche, panne d'eau chaude, panne de chauffage, etc.) avait sa propre donnée. La liste des interventions individuelles peut être longue et difficile à interpréter. Nous avons regroupé toutes ces différentes interventions (arrondies au centième près) en une seule variable « nombre d'interventions », tel que présenté dans le tableau 3-1.

Nombre d'interventions	Fréquence	%	% cumulatif
0	3 118	13,91%	13,91%
1	3 143	14,02%	27,94%
2	3 275	14,61%	42,55%
3	2 889	12,89%	55,44%
4	2 509	11,20%	66,64%
5	2 339	10,44%	77,08%
6	1 930	8,61%	85,69%
7	1 208	5,39%	91,08%
8	751	3,35%	94,43%
9	460	2,05%	96,48%
10	276	1,23%	97,72%
11	183	0,82%	98,53%
12	119	0,53%	99,06%
13	65	0,29%	99,35%
14	38	0,17%	99,52%
15	31	0,14%	99,66%
16	26	0,12%	99,78%
17	10	0,04%	99,82%
18	10	0,04%	99,87%
19	6	0,03%	99,89%
20	9	0,04%	99,93%
21	1	0,00%	99,94%
22	2	0,01%	99,95%
23	4	0,02%	99,96%
24	1	0,00%	99,97%
25	1	0,00%	99,97%
26	1	0,00%	99,98%
28	3	0,01%	99,99%
48	2	0,01%	100,00%
Total	22 410	100,00%	

Tableau 3-1 : Nombre total des interventions au cours de la période d'étude

Dans un deuxième temps, les valeurs manquantes ou aberrantes sont traitées afin qu'elles deviennent des données cohérentes. Les variables d'interventions tel que nous venons de voir (entretien, mise en marche, panne d'eau chaude,

panne de chauffage, etc.) qui n'affichent aucune donnée ne sont pas des données manquantes. Elles ont été codées avec la valeur 0, car aucune intervention n'a été effectuée sur l'équipement. Quant aux valeurs aberrantes, elles ont été étudiées attentivement cas par cas afin de comprendre la nature extrême de certaines données. Prenons l'exemple des données dont le revenu était de 23 000$ ou de 326 000$, alors que le revenu moyen du produit est de 1 040$. Après validation dans les autres données (nom et adresse), il s'agit de clients commerciaux et institutionnels. Ces deux valeurs ont été rejetées, car l'étude porte sur les clients résidentiels.

Dans un troisième temps, des transformations ont été nécessaires pour les variables nominales ayant k valeurs possibles. Nous avons créé k-1 variables indicatrices afin de les inclure dans le modèle de régression. À titre d'exemple, la variable « gen » pour la génération de l'équipement (1 : première génération, 2 : deuxième génération, 3 : troisième génération) a été recréée en deux variables indicatrices telles qu'indiquées dans le tableau 3-2. La modalité « troisième génération » devient la catégorie de référence.

gen1	gen2	gen
1	0	1
0	1	2
0	0	3

Tableau 3-2 : Transformation de la variable génération de l'équipement en variables indicatrices

La base de données originale contenait 23 603 clients. Suite à l'épuration, elle comprend 22 387 clients.

3.2.2 Règles de standardisation

Plusieurs règles de base sont établies afin de pouvoir travailler les données de façon standardisée pour les objectifs de l'étude.

(1) Absence de données temporelles pour les variables de coûts, revenus, profitabilité et interventions.

- Nous supposons donc une constance dans le temps de ces variables, ce qui est faux en réalité.
- Il est impossible de déterminer la profitabilité pour une période spécifique et de l'attribuer à un équipement ou à toute autre variable. Ainsi, un client peut être très profitable les premières années et ne plus l'être pour les dernières, mais qu'au global, il demeure profitable.

⇨ ***Règle no 1 : Travailler avec une moyenne (en fonction de la date de création du client dans le système) pour les variables d'interventions, de coût, de revenu et de profit.***

(2) Absence de données complémentaires sur les produits.

- ABC offre divers produits à ses clients en termes de couverture (pièces uniquement, pièces et main-d'œuvre, avec ou sans entretien), de durée du contrat (1 an, 2 ans ou 3 ans).

⇨ ***Règle no 2 : Quelles que soient les caractéristiques des différents produits, ceux-ci sont traités comme s'ils étaient tous identiques. Dans la présente étude, nous les nommons « produit ».***

(3) Aucune distinction pour différencier les clients avec un produit ou deux produits.

- Catégorie 1 - Client avec un produit : « programme de protection » ou « programme de location ».
- Catégorie 2 - Client avec deux produits : « programme de protection » et « programme de location ».

⇨ ***Règle no 3 : Le client de la catégorie 1 et le client de la catégorie 2 sont évalués de la même façon. Pour notre étude, le client détient un seul produit de l'entreprise ABC.***

(4) Dans la base de données, il existe deux variables intitulées « équipement 1 » et « équipement 2 ». Elles représentent l'appareil principal couvert par le produit.

- Si le client protège un autre appareil (appelé appareil secondaire) ou des accessoires reliés à l'appareil principal ou secondaire (exemple : une pompe pour le système de chauffage à eau chaude), la base de données n'en fait pas mention. Dans ce cas, il est impossible de déterminer si le total des revenus (du produit ou du revenu supplémentaire relié aux interventions non couvertes par le présent produit) et des dépenses sont attribués uniquement à l'équipement 1 ou 2. Toutes les variables associées à la profitabilité peuvent être erronées car elles peuvent être en lien direct avec un autre appareil (ou accessoire) qui n'est pas présent dans la base de données. Aussi, la différence entre l'équipement 1 et l'équipement 2 peut être une correction au niveau de l'inventaire ou représenter deux des équipements protégés par le présent produit.

⇨ ***Règle no 4 : S'il y a une différence entre « équipement 1 » et « équipement 2 », l'étude présente suppose qu'il y a eu un***

changement d'équipement au cours des 5 dernières années. Les analyses ci-jointes sont basées sur l'équipement 2.

3.2.3 Traitement de la variable dépendante

Dans la présente étude, la variable cible est la profitabilité de chacun des clients de ABC. Nous avons créé cette variable, car elle n'est pas une donnée disponible dans la base de données. Nous avons divisé le revenu net par le nombre d'années-client. La variable dépendante avec laquelle nous travaillons est une valeur moyenne annuelle.

3.3 Statistiques descriptives

3.3.1 Profil des clients actuels

La durée de vie moyenne d'un client chez ABC est de 4,89 ans. Il est à noter que nous disposons d'un historique de cinq ans. La majorité des clients de ABC (71%) sont de profil chauffage et 29% sont de profil chauffe-eau. Parmi les équipements de chauffage, 52% sont de type « eau chaude », 44% sont à « air chaud » et à peine 4% pour les autres types (foyer, radiateur…). Près de la moitié de ces équipement sont à 70% d'efficacité énergétique, car ils sont de la première génération (43%) et un peu plus du tiers (35%) sont de la deuxième, soit à 80% d'efficacité. Presque la totalité (99%) des appareils de chauffe-eau sont de type « résidentiel » et pour lesquels 94% d'entre eux sont de la première génération. Les appareils de la clientèle sont âgés. En se basant sur la génération de l'équipement, 69% des appareils ont plus de 20 ans (première génération), 13% sont de la deuxième (âgés entre 10 et 20 ans) et 8% sont de la troisième génération (âgés de 10 ans et moins), ce qui signifie une efficacité énergétique à 90%. Il y a 10% des équipements pour lesquels la génération est

inconnue. La figure 3-1 dresse un portrait global de l'ensemble de la clientèle de ABC.

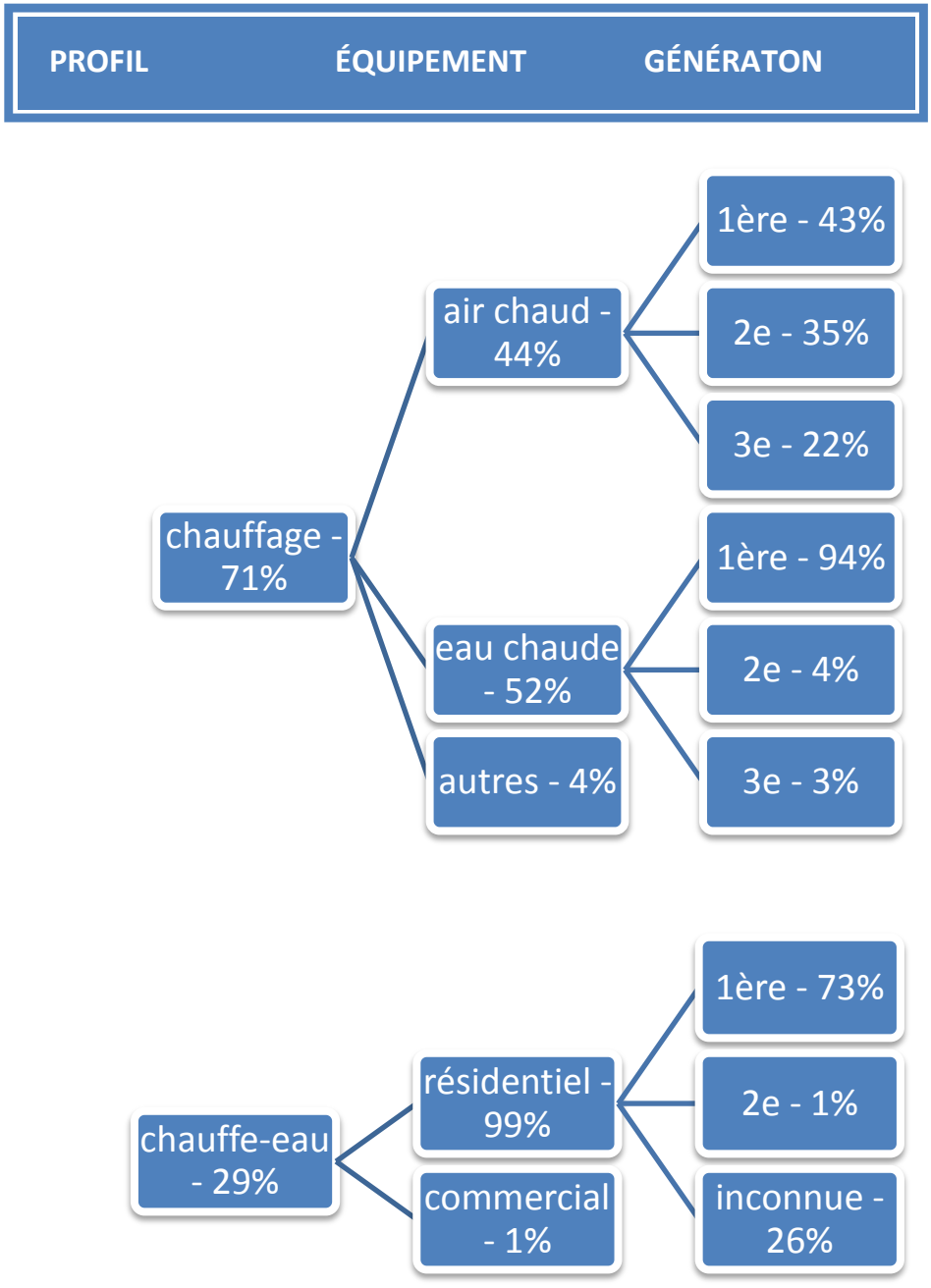

Figure 3-1 : Portrait de la clientèle de ABC

Il y a presqu'autant de clients qui ont changé d'équipement (51%) que ceux qui ont gardé le même (49%) au cours de ces cinq dernières années. Au-delà des revenus du produit, 39% des clients ont été facturés pour des interventions du technicien qui ne sont pas couvertes par le présent produit (facturation supplémentaire). La figure 3-2 illustre le nombre moyen d'interventions (arrondi à l'entier supérieur) par année selon l'équipement (interventions couvertes et non couvertes par le produit).

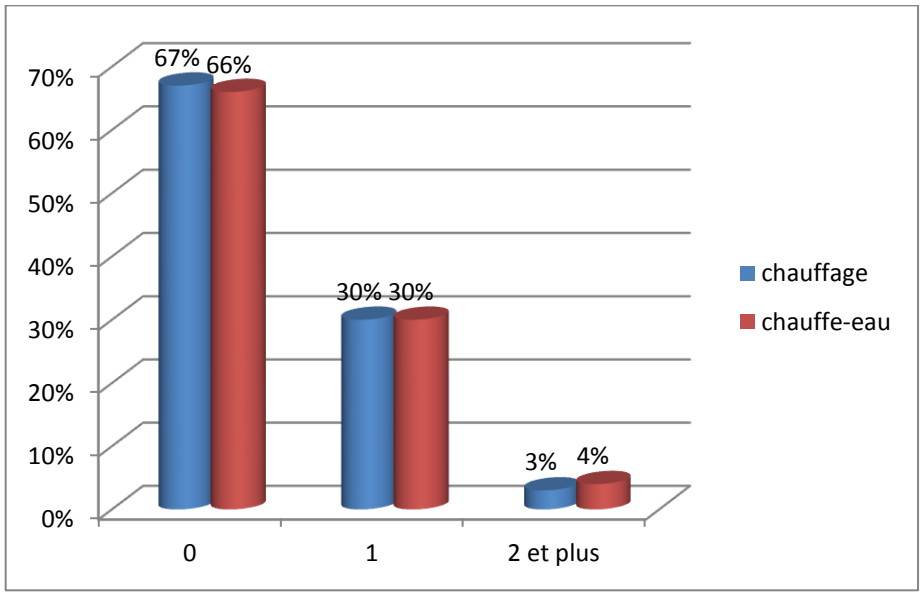

Figure 3-2 : Nombre moyen d'interventions (couvert et non couvert par le produit) par année selon le profil chauffage et chauffe-eau

3.3.2 Profil de la clientèle profitable en termes de proportion

Dans une perspective exploratoire, nous avons analysé les variables individuellement afin de déterminer si elles sont reliées directement ou non à la profitabilité du client. L'analyse est basée sur la durée totale du produit. Il ne

s'agit pas d'une valeur annuelle moyenne. Le concept de profitabilité dans le cas présent se définit comme suit :

- client profitable : si revenu net ≥ 0,
- client non profitable : si revenu net <0,

où

revenu net = revenu total – coût total,

revenu total = revenu du produit + revenu des interventions non couvertes par le produit (facturation supplémentaire),

coût total = somme de tous les coûts de la base de données.

À priori, 88% de la clientèle de ABC sont profitables. Le fait d'avoir changé d'équipement au cours des 5 dernières années n'a pas d'effet sur la profitabilité. La proportion de profitabilité est très élevée et similaire que l'équipement ait été changé ou non (87% et 89% respectivement) tel que nous pouvons le remarquer à la figure 3-3. Pourtant, cette différence de profitabilité est significative car la « p-value » est de 0,00 lors de l'exécution du test de Khi-Carré. La proportion de profitabilité est significativement plus élevée (89%) chez la clientèle ayant gardé son équipement par rapport à celle qui a changé (87%).

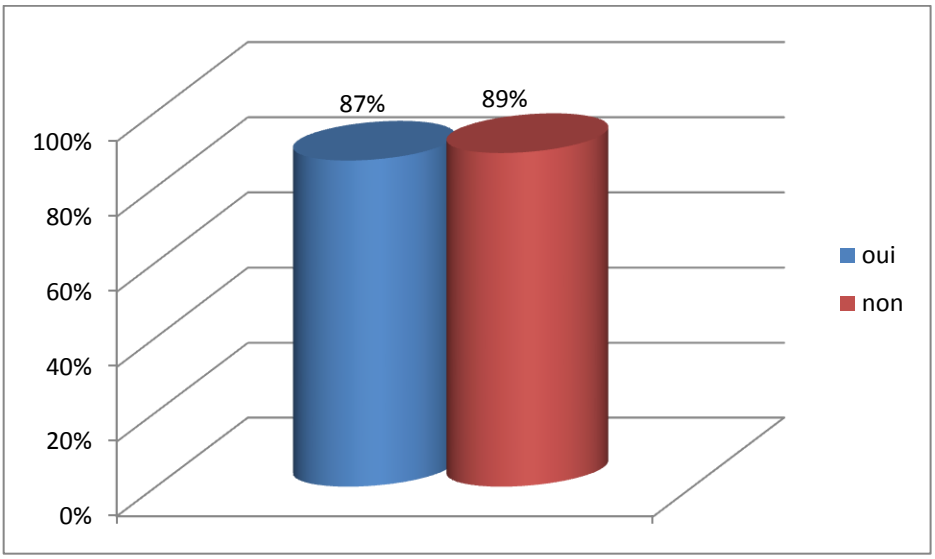

Figure 3-3 : Proportion de profitabilité en fonction du changement d'équipement

Parmi la clientèle profitable, 69% ont un profil chauffage et 31% ont un profil chauffe-eau. En termes de proportion, 95% de ceux qui possèdent un profil chauffe-eau sont profitables, 88% le sont avec un profil chauffage. La figure 3-4 illustre la situation. Cette différence de profitabilité est significative car la « p-value » est de 0,00 lors de l'exécution du test de Khi-Carré.

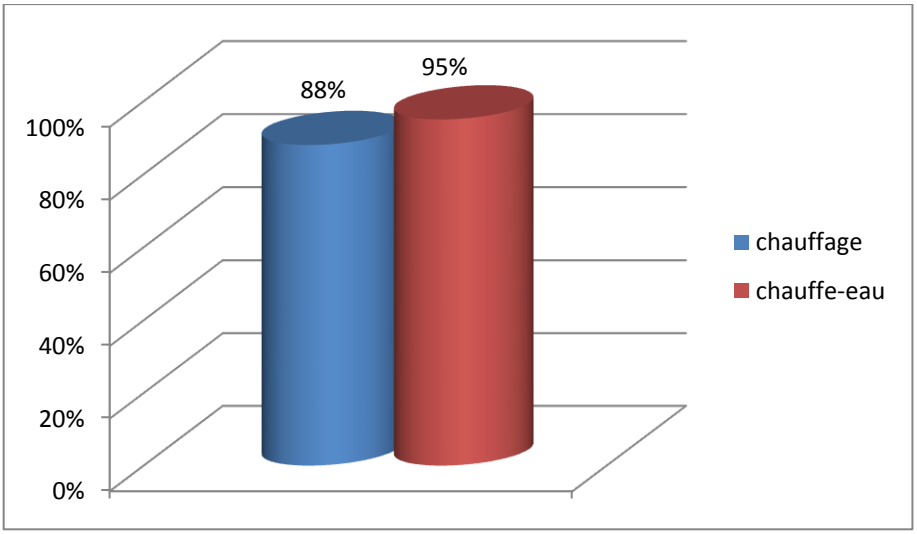

Figure 3-4 : Proportion de profitabilité en fonction du profil de l'équipement

Quant aux équipements, le chauffe-eau résidentiel est celui qui a le plus haut taux de profitabilité à 95%. En appliquant le test de l'analyse de la variance à un 1 facteur contrôlé, la « p-value » du test de l'homogénéité des variances est de 0,00, ce qui rejette l'égalité des variances. Les tests de Welch et de Brown-Forsythe ont donc été exécutés et leur « p-value » est de 0,00, il y a donc au moins une différence significative parmi les équipements. En effectuant le test t pour échantillons indépendants, nous constatons qu'il y a une différence significative entre la plupart des équipements car les « p-values » sont inférieures à 0,05. Par contre, certains équipements ne le pas sont tels que le système de chauffage à l'air chaud avec le chauffe-eau commercial, le système de chauffage à l'eau chaude et le radiateur ainsi que le système de chauffage à l'eau chaude et le radiateur. La proportion de profitabilité d'un chauffe-eau résidentiel 95% > radiateur 89% > système de chauffage à l'air chaud et à l'eau chaude 88% > chauffe-eau commercial 79% > foyer 76%. La figure 3-5 ci-dessous le démontre.

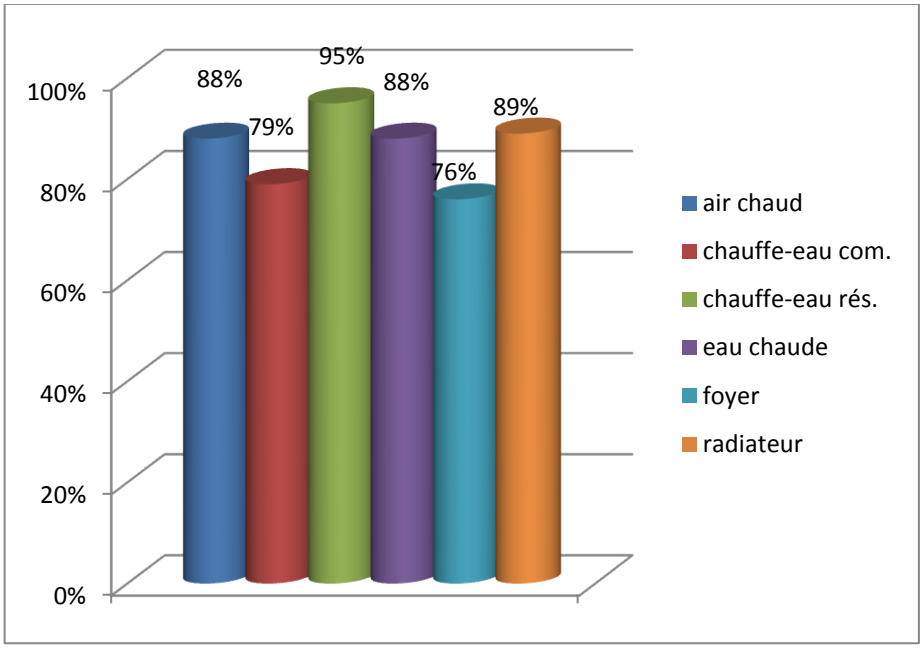

Figure 3-5 : Proportion de profitabilité en fonction de l'équipement

En ce qui concerne la génération connue de l'équipement, la génération qui a la plus forte proportion de profitabilité est la première avec 91%, suivi de la deuxième et de la troisième avec 87% et 83% respectivement. Ces constats, présentés à la figure 3-6, démontrent que plus l'équipement est d'une génération ancienne (dont l'efficacité est plus faible), plus il est profitable.

Afin de déterminer si la différence de proportion est significative entre les trois générations, nous avons procédé avec le test de l'analyse de la variance à un facteur contrôlé. La « p-value » du test de l'homogénéité des variances est de 0,00, ce qui rejette l'égalité des variances. Dans ce cas, les tests de Welch et de Brown-Forsythe ont été effectués et leur « p-value » est de 0,00. Il y a donc au moins une différence significative parmi les 3 générations. En effectuant le test t pour échantillons indépendants, nous constatons qu'il y a une différence significative entre chaque paire de générations : la première et la deuxième, la première et la troisième, et entre la deuxième et la troisième (les « p-values » sont de 0,00).

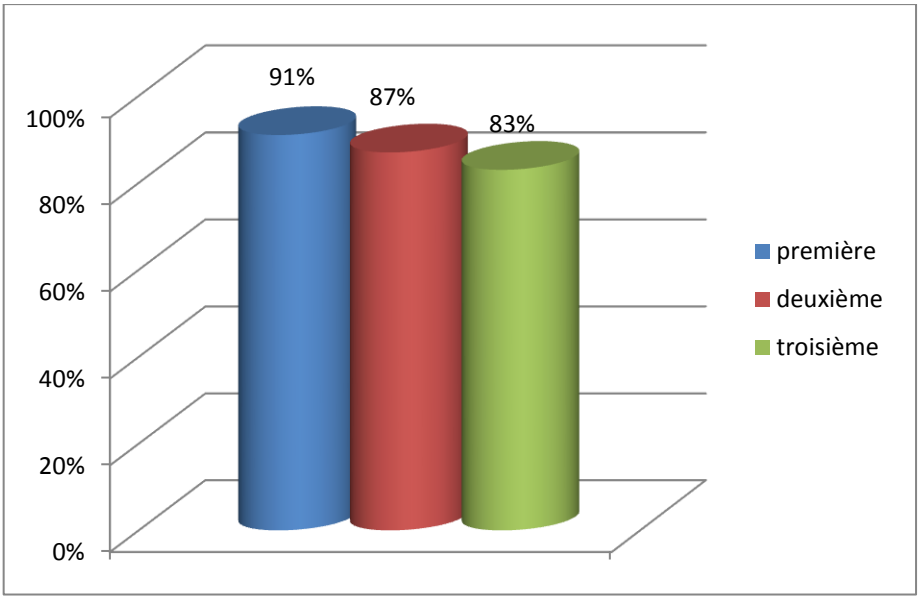

Figure 3-6 : Proportion de profitabilité en fonction de la génération (connue) de l'équipement

Pour ce qui est de la facturation supplémentaire présentée dans la figure 3-7 (interventions non couvertes par le produit), qu'il y en ait ou non, la proportion est relativement semblable (89% - facturation supplémentaire vs 91% - pas de facturation). Cette légère différence de proportion de profitabilité demeure significative. La « p-value » est 0,00 lors de l'exécution du test de Khi-Carré.

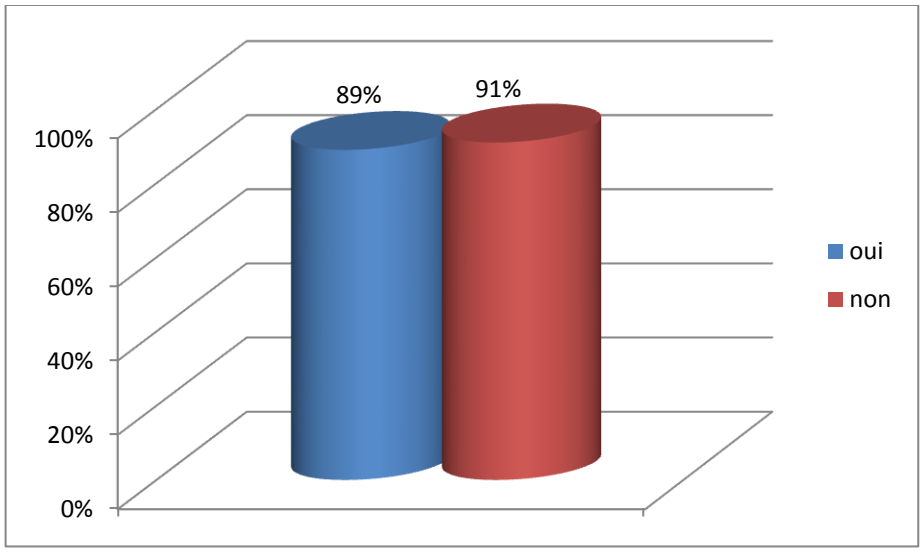

Figure 3-7 : Proportion de profitabilité en fonction de la facturation supplémentaire

En examinant le nombre moyen d'interventions par année, nous remarquons que moins il y a d'interventions sur l'équipement, plus la profitabilité augmente. Ce constat est illustré dans la figure 3-8 ci-dessous.

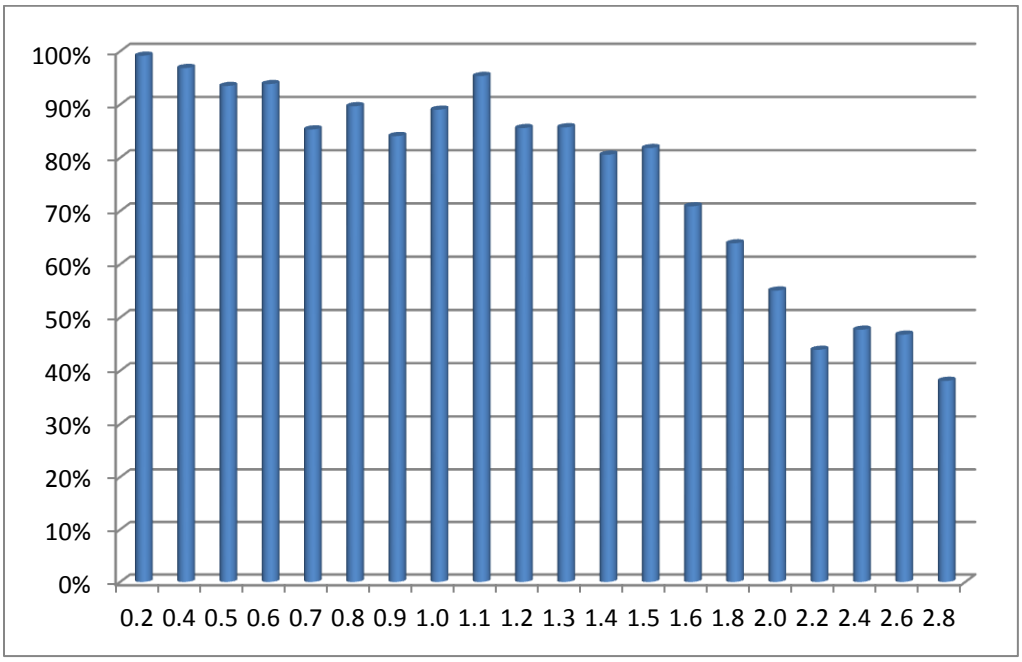

Figure 3-8 : Nombre moyen d'interventions par année

Note : La figure 3-8 ne tient pas compte des données dont le n<30. Le n total est de 22 196.

3.4 Test de profitabilité moyenne

La section précédente a permis de déterminer sous forme de proportion si un client était significativement profitable ou non profitable selon chacune des variables étudiées. L'analyse est basée sur la durée totale du produit du client. La profitabilité d'un client récent par rapport à un client plus ancien peut être difficilement comparable et interprétable; la base de comparaison du temps étant inexistante.

Cette section permet de quantifier la profitabilité en termes monétaire et est ramenée sous forme de moyenne en fonction du nombre d'années-client. Le

concept de profitabilité moyenne dans le cas présent se définit comme suit et est appliquée pour chacune des variables étudiées:

profitabilité moyenne = revenu net / nombre d'années client.

Notre appellation de la profitabilité client rejoint celle utilisée par Berger et Nars (1998). Pour eux, la profitabilité client se traduit par la différence entre les revenus et les coûts (d'attraction, de vente, de services et d'acquisition) actualisés à la valeur actuelle au cours de la relation.

La profitabilité moyenne d'un client de ABC est de 107.09$. L'écart type est de 123.02$, ce qui est élevé par rapport à la moyenne. Les données sont largement dispersées tel que le présente la figure 3-9. Ce grand écart découle du fait qu'il n'est pas possible d'isoler les données spécifiques à un produit particulier. Peu importe le type de produits, nous les considérons tous identiques dans notre analyse. Il n'est pas étonnant de constater que la valeur minimum est de -1 812.78$ et la valeur maximale est de 7 451.24$.

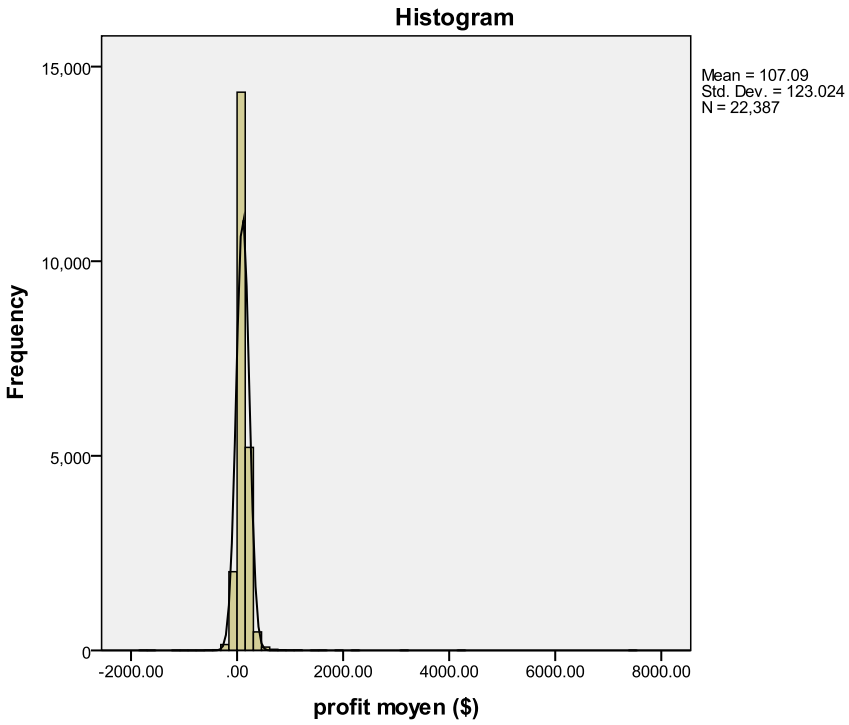

Figure 3-9 : Distribution de la variable dépendante « profit moyen »

Tel que désigne la figure 3-10, le premier quartile indique que 25% des clients ont une profitabilité inférieure à 51,42$. Le deuxième quartile ou la médiane démontre que 50% des clients ont une profitabilité supérieure à 99,03$, donc l'autre 50% ont une profitabilité inférieure à 99,03$. Et 75% ont une profitabilité supérieure à 157,16$.

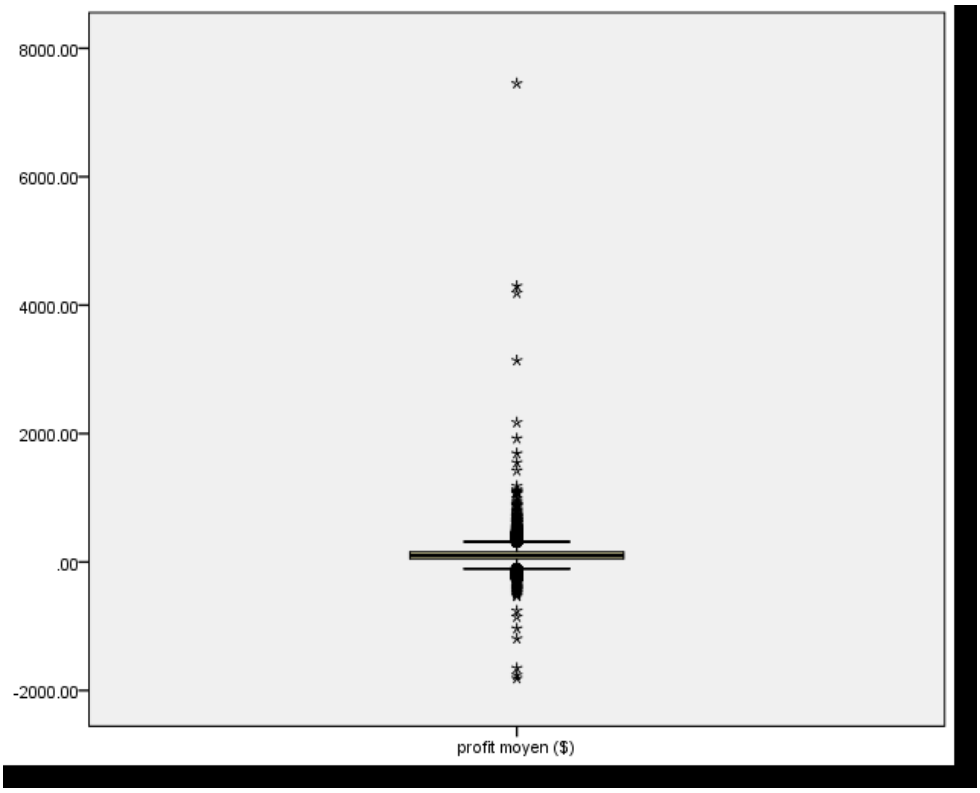

Figure 3-10: Présentation des quartiles de la variable dépendante « profit moyen »

3.4.1 Équipement : profil, type, génération et marque

Dans la section qui suit, nous allons étudier la variable « équipement » dans son intégralité, soit le profil, le type, la génération et la marque de l'équipement. Nous débutons avec la variable profil de l'équipement qui permet de déterminer si le profil chauffage est significativement profitable par rapport au profil chauffe-eau. Le test t pour échantillons indépendants démontre qu'il y a une différence significative entre le profil chauffage et chauffe-eau, car la « p-value » est de 0,00. La profitabilité moyenne du chauffe-eau 147,49$ > chauffage 90,33$. Ces moyennes sont présentées dans la figure 3-11. Le tableau 3-3 présente les statistiques descriptives.

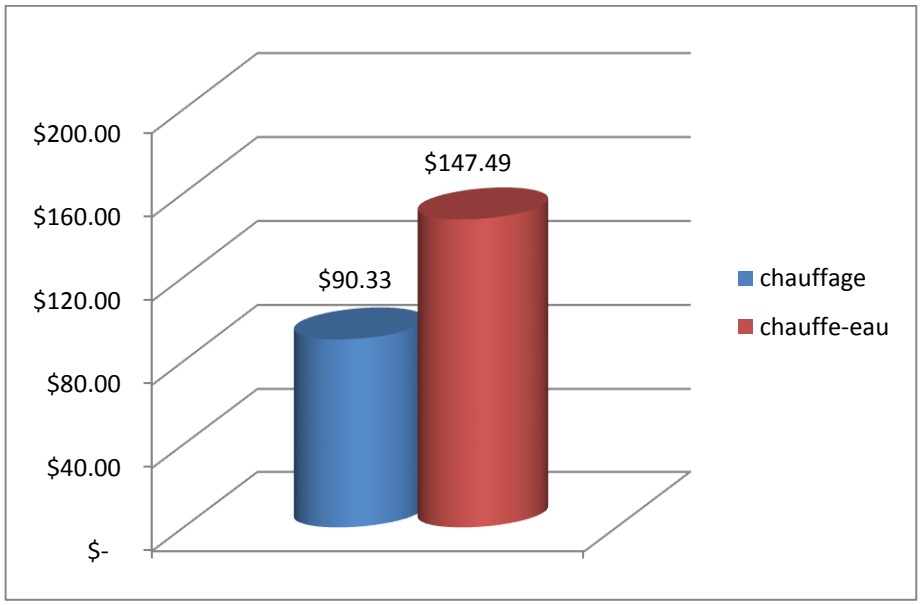

Figure 3-11: Profitabilité moyenne en fonction du profil de l'équipement

Profil	Moy	Écart-type	Min	Max	Quart 1	Méd	Quart3
Chauffe-eau	147,49	140,91	-1760,56	7451,24	195,15	146,97	44,82
Chauffage	90,33	110,54	-1812,78	4297,68	102,30	85,40	65,75
Total	107,09	123,02	-1812,78	7451,24	102,30	99,03	65,75

Tableau 3-3 : Statistiques descriptives en fonction du profil de l'équipement

Ensuite, nous déterminons lesquels des équipements ont le plus d'influence sur la profitabilité et si l'équipement en question est significativement profitable par rapport à un autre. En exécutant le test de l'analyse de la variance à un facteur contrôlé, la « p-value » du test de l'homogénéité des variances est de 0,00. Cette dernière rejette l'égalité des variances. Dans ce cas, les tests de Welch et de Brown-Forsythe sont appliqués et leur « p-value » est de 0,00. Il y a au moins une différence significative parmi les équipements. Le test t pour échantillons indépendants démontre qu'il y a une différence significative pour tous les équipements. La profitabilité moyenne d'un chauffe-eau commercial 235,19$ > chauffe-eau résidentiel 146,66$ > système de chauffage à l'eau

chaude 94,46$ > système de chauffage à l'air chaud 88,13$ > radiateur 62,96$ > foyer 38,95$. La figure 3-12 illustre la profitabilité moyenne des différents équipements. Le tableau 3-4 présente les statistiques descriptives.

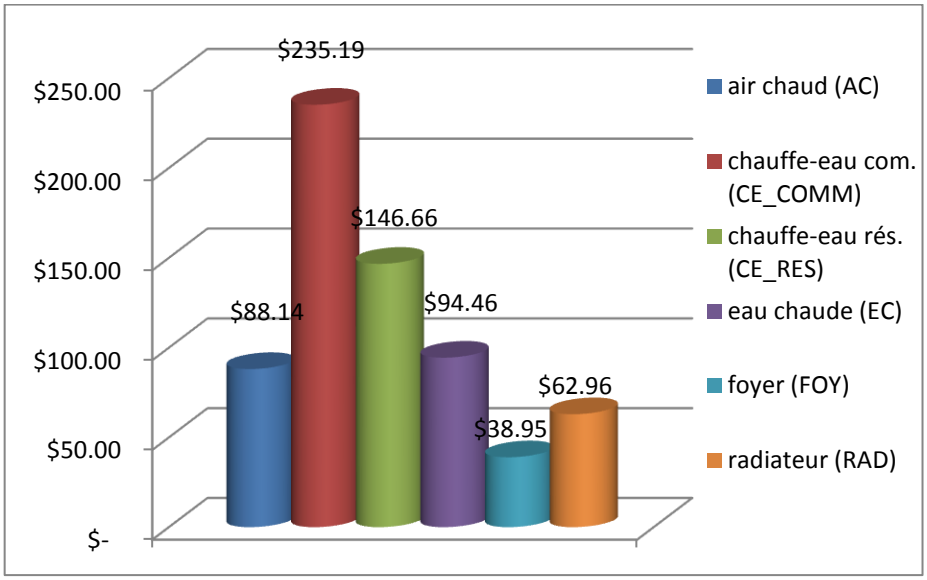

Figure 3-12: Profitabilité moyenne en fonction de l'équipement

Equip	Moy	Écart-type	Min	Max	Quart 1	Méd	Quart3
AC	88,14	90,09	-1037,45	1096,13	161,81	87,55	72,08
CE_COMM	235,19	309,03	-242,01	1181,92	351,82	128,69	-242,01
CE_RES	146,66	138,11	-1760,56	7451,24	195,15	147,04	44,82
EC	94,46	127,36	-1812,78	4297,68	102,30	87,25	65,75
FOY	38,95	72,36	-158,10	228,15	141,87	37,16	6,54
RAD	62,96	64,24	-148,01	437,70	56,13	62,62	-80,55
Total	107,09	123,02	-1812,78	7451,24	102,30	99,03	65,75

Tableau 3-4 : Statistiques descriptives en fonction de l'équipement

Quant à la génération de l'équipement, nous avons procédé avec l'analyse de la variance à un facteur contrôlé pour déterminer s'il existe une différence significative. La « p-value » du test de l'homogénéité des variances est de 0,00, cette dernière rejette l'égalité des variances. Dans ce cas, les tests de Welch et de Brown-Forsythe sont appliqués et leur « p-value » est de 0,00. Il y a au moins une différence significative parmi les trois générations. Le test t pour échantillons indépendants dénote une différence entre chaque génération sauf entre la deuxième et troisième. La profitabilité moyenne de la première génération 109,77$ > deuxième génération 90,39$ > troisième génération 85,18$. La moyenne des trois générations est de 104,88$, ce qui diffère de la moyenne de 107,09$ dû à des valeurs manquantes (n=20 083). Selon les données de l'étude, la tendance démontre que plus la génération est récente, moins la profitabilité est grande telle que démontrée dans la figure 3-13. Le tableau 3-5 présente les statistiques descriptives.

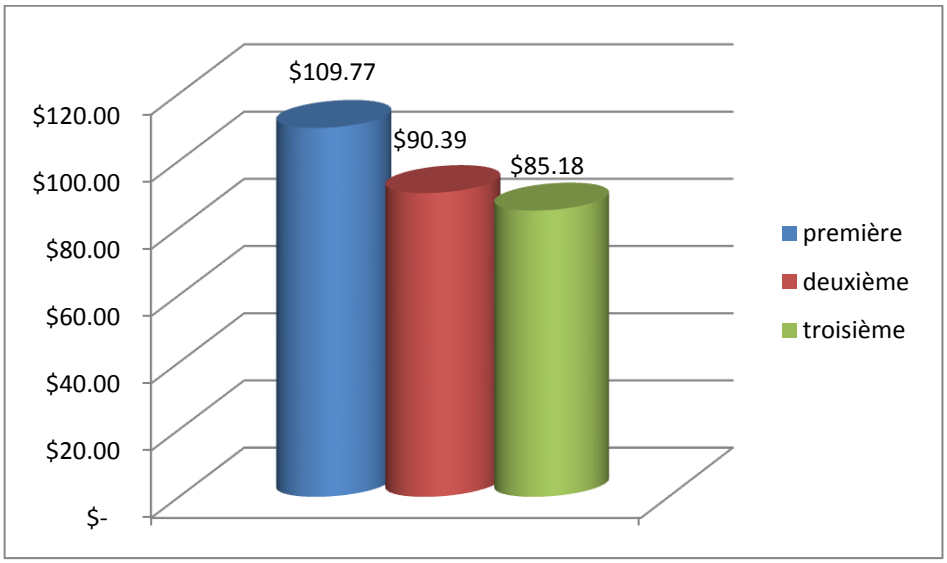

Figure 3-13: Profitabilité moyenne en fonction de la génération (connue) de l'équipement

Gen	Moy	Écart-type	Min	Max	Quart 1	Méd	Quart3
1	109,77	116,55	-1812,78	4297,68	102,30	99,61	65,75
2	90,39	92,36	-437,27	731,35	348,72	88,74	-437,27
3	85,18	124,48	-1194,01	1060,88	87,52	86,37	-477,44
Total	104,88	114,53	-1812,78	4297,68	102,30	98,76	65,75

Tableau 3-5 : Statistiques descriptives en fonction de la génération (connue) de l'équipement

Finalement, la marque permet de conclure ce volet sur l'équipement. La donnée sur la marque de l'équipement est parfois disponible dans la base de données (n=14 025). Étant donné la liste exhaustive de marques sur le marché, le gestionnaire de ABC a sélectionné les marques les plus populaires pour les fins d'analyse que nous présentons à la figure 3-14. Les tests t pour échantillons indépendants démontrent qu'il peut y avoir des différentes significatives entre les marques.

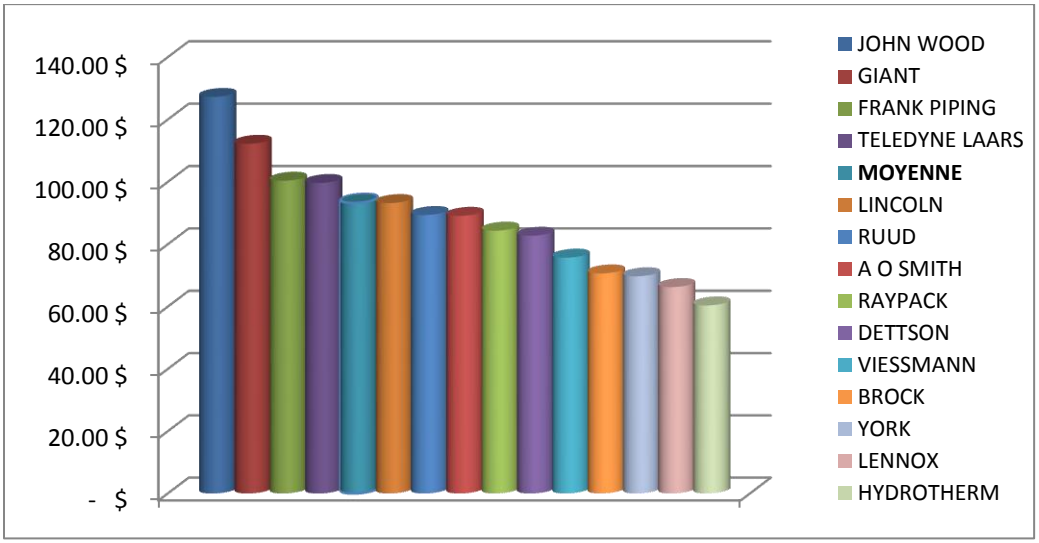

Figure 3-14: Profitabilité moyenne en fonction de la marque (connue) de l'équipement

3.4.2 Changement d'équipement

La profitabilité lors d'un changement d'équipement est supérieure par rapport à un non changement (125,17$ > 88,56$) comme le présente la figure 3-15. Cette différence est significative lorsque le test t pour échantillons indépendants est appliqué, car la « p-value » est de 0,00. Le tableau 3-6 présente les statistiques descriptives.

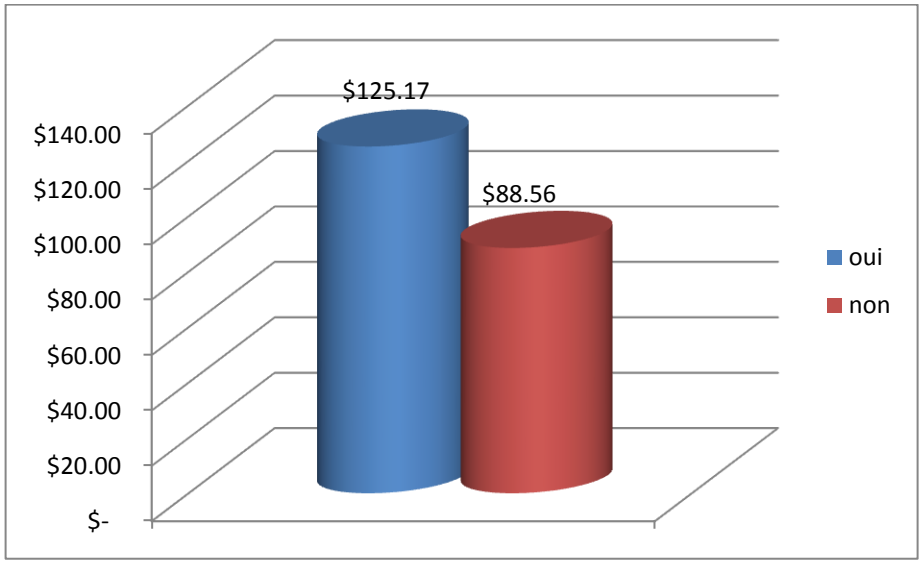

Figure 3-15: Profitabilité moyenne en fonction du changement d'équipement

Chgt	Moy	Écart-type	Min	Max	Quart 1	Méd	Quart3
Oui	125,17	133,95	-1812,78	7451,24	102,30	122,42	44,82
Non	88,56	107,61	-1194,01	4297,68	-163,32	83,94	65,75
Total	107,09	123,02	-1812,78	7451,24	102,30	99,03	65,75

Tableau 3-6 : Statistiques descriptives en fonction du changement d'équipement

3.4.3 Année d'installation et nombre d'interventions

La durée de vie moyenne d'un client de ABC est de 4,89 ans, les données sur les années antérieures ne sont donc pas assez représentatives pour analyser la profitabilité moyenne par année. Néanmoins, nous pouvons analyser les cinq dernières années d'installation des équipements. La profitabilité moyenne de ces années est de 119,38$. Nous avons procédé au test de l'analyse de la variance à un facteur contrôlé. La « p-value » du test de l'homogénéité des variances est de 0,32, cette dernière ne rejette pas l'égalité des variances. Dans ce cas, le test Fisher est appliqué et la « p-value » est de 0,00, car ce test suppose l'égalité des variances. Il y a au moins une différence significative parmi les années d'installation étudiées. Par la suite, le test t pour échantillons indépendants prouve qu'il y a une différence entre 2005-2007, 2005-2008, 2005-2009, 2006-2008,2006-2009 et 2007-2008. La profitabilité moyenne de l'année d'installation en 2005 : 132,56$ > 2006 : 128,59$ > 2007 : 120,85$ > 2008 : 101,88$ > 2009 : 100,23$. La tendance démontre que plus l'année de l'installation de l'équipement est récente, moins la profitabilité est grande. Ce constat est représenté dans la figure 3-16. Le tableau 3-7 présente les statistiques descriptives.

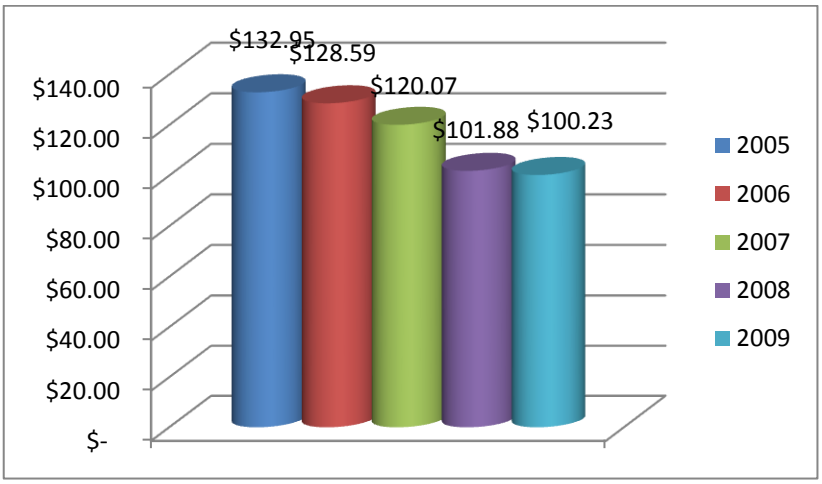

Figure 3-16: Profitabilité moyenne en fonction de l'année d'installation (connue) de l'équipement

Année	Moy	Écart-type	Min	Max	Quart 1	Méd	Quart3
2005	132,95	101,45	-532,89	482,91	195,15	136,21	268,52
2006	128,59	101,61	-230,39	883,97	92,08	128,05	279,30
2007	120,07	98,96	-220,39	649,02	234,70	122,51	317,42
2008	101,88	122,74	-1760,56	808,98	31,15	103,44	-440,21
2009	100,23	128,51	-417,58	704,89	704,89	104,50	51,24
Total	119,24	108,57	-1760,56	883,97	195,15	121,03	268,52

Tableau 3-7 : Statistiques descriptives en fonction de l'année d'installation (connue) de l'équipement

3.4.4 Facturation

Le fait de facturer pour les services d'un technicien (ce qui n'est pas couvert par le produit) contribue à la profitabilité, au-delà du prix déboursé pour le produit tel qu'indiqué dans la figure 3-17. Selon le test t pour échantillons indépendants, la différence est significative. Un client qui est facturé pour les services du technicien voit sa profitabilité supérieure par rapport à celui qui ne l'a pas été (115.57$ > 101.67$). La « p-value » obtenue pour ce test est de 0,00. Rappelons-nous que ces revenus supplémentaires ne sont pas nécessairement associés à l'équipement de la présente base de données. Ces revenus peuvent être appliqués sur des équipements secondaires ou des options non présentes dans nos données. Le tableau 3-8 présente les statistiques descriptives.

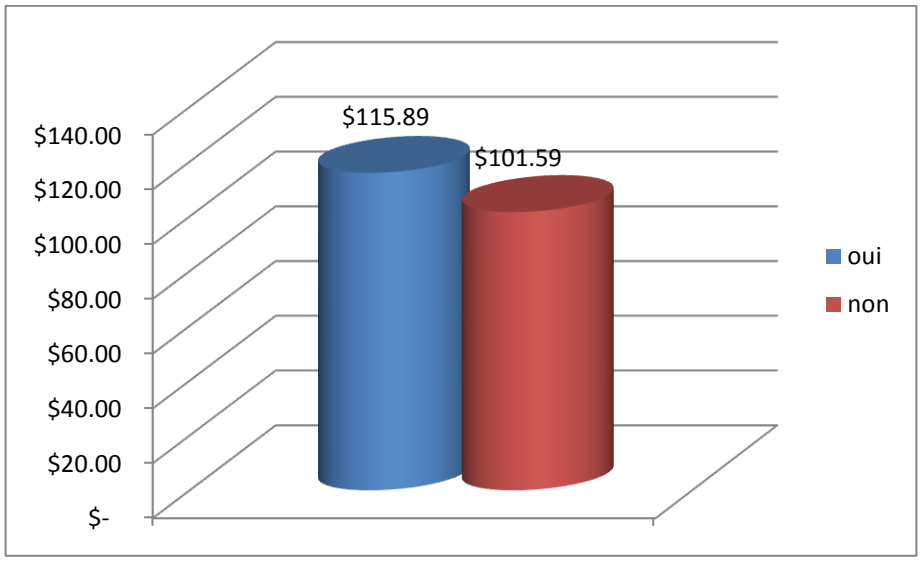

Figure 3-17: Profitabilité moyenne en fonction de la facturation supplémentaire

Fact suppl.	Moy	Écart-type	Min	Max	Quart 1	Méd	Quart3
Oui	115,89	135,56	-1812,78	4297,68	195,15	109,17	65,749
Non	101,59	119,44	-1652,76	7451,24	102,30	96,21	72,077
Total	107,17	126,16	-1812,78	7451,24	102,30	98,99	65,749

Tableau 3-8 : Statistiques descriptives en fonction de l'année d'installation (connue) de l'équipement

Chapitre 4

Modèles et résultats

Dans le cadre de l'étude présente, nous sommes intéressés par la profitabilité client. Tel que nous l'avons constaté dans le chapitre 2, ce concept a plus d'une définition et certaines d'entre elles se contredisent. Pour l'étude, nous nous référons à la profitabilité client définie par Berger et Nars (1998) sous l'appellation de capital client (« customer equity »). Rappelons que le capital client de Berger et Nars (1998) est la différence entre les revenus et les coûts (d'attraction, de vente, de services et d'acquisition) actualisés à la valeur actuelle au cours de la relation.

Nous cherchons à expliquer le comportement de la variable dépendante quantitative « profitabilité moyenne » à l'aide des variables indépendantes suivantes: le nombre d'années-client, le changement d'équipement au cours des cinq dernières années, le profil d'usage de l'équipement, le type de l'équipement et la génération de l'équipement. L'année d'installation de l'équipement n'est pas considérée, car plus de 70% des données pour cette variable sont manquantes. En revanche, la variable génération de l'équipement s'avère être un bon choix de substitut : la première

génération est attribuée pour les équipements âgés de plus de 20 ans, la deuxième génération pour les équipements âgés entre 10 et 20 ans et finalement, la troisième génération pour des équipements plus récents, soit ceux âgés de 10 ans et moins.

Notre échantillon de taille n=22 387 est divisé aléatoirement en deux parties. La première partie comprend 70% des données, ce sont les données de l'apprentissage. La deuxième partie, soit les 30% des données restantes est celle de la validation. Cette façon de procéder permet de construire le modèle et d'ajuster les données en mode apprentissage. Les données de validation permettent d'évaluer la performance du modèle construit.

Les différentes sous-sections sont inspirées du recueil de Larocque (2006). Dans un premier temps, nous recourons à la régression linéaire multiple et à la régression gamma pour comprendre lesquels des prédicteurs expliquent la profitabilité clientèle de l'entreprise ABC. Dans un deuxième temps, nous employons la régression logistique. Cette dernière permet d'identifier si le client est profitable ou ne l'est pas. Dans un troisième temps, nous suggérons un modèle qui combine à la fois la régression linéaire et la régression logistique. Finalement, nous émettons des recommandations à l'entreprise ABC suite à nos différentes analyses et nous discutons des limites de l'étude.

4.1 Régression linéaire multiple : Modèles M1 et M2

4.1.1 Modélisation

La régression linéaire multiple établit la relation linéaire entre la variable dépendante et les variables indépendantes. Le modèle de régression linéaire multiple est décrit sous la forme suivante :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + + \beta_k X_k + \varepsilon ,$$

où

Y : la profitabilité moyenne,
k : le nombre de variables indépendantes,
$X_1, \ldots X_k$: les variables indépendantes,
$\beta_0, \ldots, \beta_k$: les paramètres du modèles à estimer,
ε: l'erreur.

4.1.2 Coefficient de corrélation multiple: R2

D'une valeur comprise entre 0 et 1, le coefficient de corrélation multiple (R^2) mesure la qualité d'adéquation entre le modèle et les données observées. Il permet de quantifier la force de la relation linéaire entre la variable dépendante et les variables indépendantes. Plus sa valeur est élevée, plus il existe une relation linéaire forte entre la variable dépendante et les variables indépendantes.

$$R^2 = 1 - \frac{SCE}{STCE},$$

où

STCE : la somme totale des carrés,
SCE : la somme des carrés des erreurs.

Cependant, lorsqu'une variable quelconque est ajoutée au modèle, le R^2 augmente systématiquement même si cette dernière ne contribue pas à prédire la variable dépendante. Une variante du R^2, appelée le coefficient de

détermination ajusté (R^2 ajusté) diminue, si la variable ajoutée ne contribue pas suffisamment à améliorer le modèle. Plus la valeur du R^2 ajusté est élevée, meilleur est le modèle.

$$R^2 \text{ajusté} = 1 - \frac{\text{EQM}}{\text{STCE}/(\text{n}-1)} = 1 - \frac{SCE/(n-k-1)}{STCE/(n-1)},$$

où

EQM : l'erreur quadratique moyenne,
STCE : la somme totale des carrés,
SCE : la somme des carrés des erreurs.

4.1.3 Multicolinéarité

Dans le cas des modèles de régression, la notion de multicolinéarité prend son importance. Elle réfère à la situation où les variables explicatives sont fortement corrélées entre elles ou qu'une ou plusieurs variables explicatives sont fortement corrélées à une combinaison des autres variables. Ces cas font en sorte qu'un paramètre significatif peut s'avérer non significatif dans le modèle (Larocque, 2006).

Pour détecter la multicolinéarité, il faut d'abord étudier la matrice de corrélation de Pearson que nous retrouvons dans le tableau 4-1. D'une valeur comprise entre -1 et 1, le coefficient de corrélation de Pearson (r) mesure la force de la relation linéaire entre deux variables. Un coefficient de 1 signifie une corrélation positive parfaite entre les deux variables. À l'opposé, un coefficient de -1 indique une corrélation négative parfaite. Donc, plus la valeur du coefficient est proche de 1 ou de -1, plus les deux variables sont fortement corrélées entre elles linéairement.

Lorsque nous examinons la matrice de corrélation de Pearson, nous constatons que la paire de variables « profil » et « changement » sont fortement corrélées entre elles (r>0,6). En d'autres termes, les deux variables expliquent partiellement de la même façon leur influence sur la profitabilité clientèle.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations						
	nb_an_cl	interv	chgt	profil	equip	gen
nb_an_cl nb années-client	1.00000 22387	0.07878 <.0001 22387	0.02779 <.0001 22387	-0.01533 0.0218 22387	-0.03492 <.0001 22387	-0.01105 0.1173 20083
interv intervention	0.07878 <.0001 22387	1.00000 22387	0.03774 <.0001 22387	-0.00610 0.3611 22387	0.09261 <.0001 22387	0.02249 0.0014 20083
chgt changement équip	0.02779 <.0001 22387	0.03774 <.0001 22387	1.00000 22387	-0.62537 <.0001 22387	-0.20572 <.0001 22387	0.16483 <.0001 20083
profil profil équip	-0.01533 0.0218 22387	-0.00610 0.3611 22387	-0.62537 <.0001 22387	1.00000 22387	0.08139 <.0001 22387	0.26657 <.0001 20083
equip equip	-0.03492 <.0001 22387	0.09261 <.0001 22387	-0.20572 <.0001 22387	0.08139 <.0001 22387	1.00000 22387	-0.46939 <.0001 20083
gen gen	-0.01105 0.1173 20083	0.02249 0.0014 20083	0.16483 <.0001 20083	0.26657 <.0001 20083	-0.46939 <.0001 20083	1.00000 20083

Tableau 4-1 : Matrice de corrélation de Pearson

Toutefois, cette matrice n'est pas assez satisfaisante pour détecter le phénomène de la multicolinéarité. Les coefficients individuels peuvent ne pas être élevés même si la combinaison linéaire des autres variables peut être fortement corrélée avec une variable explicative. Il existe d'autres outils pour la détection, le facteur de l'index de la variance (« VIF : variance inflation factor ») et la tolérance qui sont présentés dans le tableau 4-2.

Le VIF se définit tel que :

$$VIF_k = \frac{1}{1-R^2_{(k)}},$$

où

$R^2_{(k)}$: le coefficient de corrélation multiple du modèle de régression de $X_{(k)}$ sur toutes les autres variables explicatives.

Plus la valeur du VIF est petite, moins il y a de colinéarité entre les variables. Il n'y a pas de consensus au sein des chercheurs sur la coupure du VIF qui indique un problème de multicolinéarité. La coupure étant la valeur à partir de laquelle il y a détection de la multicolinéarité. Dans la littérature, certains soulèvent un problème de multicolinéarité lorsque le VIF>10 (Hair et al., 1995; Kennedy, 1992). Pour Rogerson (2011), une coupure à 5 est suffisante et même, pour Pan et Jackson (2008), une coupure à 4 est acceptable. Pour différents chercheurs, le VIF est plus un outil qu'une règle.

Une coupure de VIF>10 signifie que R^2 >0,9, donc 90% de la variabilité de la variable explicative est expliquée par les autres variables explicatives.

La tolérance est un autre indicateur de la colinéarité. D'une valeur comprise entre 0 et 1, elle se définit par 1-R^2. Plus le niveau de tolérance est faible, plus le problème de multicolinéarité est présent. La plupart des auteurs dont Tabachnick et Fidell (2001) soutiennent que le niveau acceptable minimal de tolérance est de 0,10. Tout comme dans le cas du VIF, d'autres recommandent 0,20 (Menard, 1995) et même 0,25 (Huber et Stephens, 1993).

Ainsi, les résultats présentés dans le tableau 4-2 démontrent qu'il n'y pas de problème de multicolinéarité en fonction des valeurs déterminées par les chercheurs pour le VIF et la tolérance.

Parameter Estimates							
Variable	Label	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	345.01375	14.61938	23.60	<.0001	.	0
nb_an_cl	nb années-client	-38.21993	2.87732	-13.28	<.0001	0.99115	1.00893
interv	intervention	-46.31155	1.53170	-30.24	<.0001	0.96057	1.04104
chgt	changement équip	10.45297	2.41360	4.33	<.0001	0.54810	1.82448
profil	profil équip	-46.96775	2.91540	-16.11	<.0001	0.51188	1.95360
equip	equip	3.74875	0.60423	6.20	<.0001	0.72121	1.38656
gen	gen	-1.00873	1.83802	-0.55	0.5831	0.59116	1.69158

Tableau 4-2 : Tableau du facteur d'inflation de la variance

4.1.4 Sélection de modèles

Le modèle sélectionné doit être performant pour la prédiction des futures valeurs de la variable dépendante. Plusieurs méthodes existent pour nous guider sur les performances prévisionnelles de notre modèle. Nous en présentons quelques-unes brièvement.

Estimation directe de l'erreur quadratique moyenne de généralisation (EQMG)

Une première méthode consiste à estimer directement l'erreur quadratique moyenne de généralisation tel que décrit ci-dessous:

$$EQMG = E\left[\left(Y - f\left(X_{1\ldots}\ X_{K)}\right)\right)^{2}\right]$$

Cette quantité mesure la moyenne du carré de l'erreur obtenue de notre modèle. Plus cette quantité est minimisée, meilleur est le modèle. Elle peut être appliquée à l'aide des méthodes de rééchantillonage telles que le bootstrap, la division de l'échantillon et la validation croisée. Le principal désavantage de cette approche est que toutes les données ne sont pas utilisées dans l'ajustement du modèle. Par conséquent, il y a une perte au niveau de la précision.

Modification de l'erreur quadratique moyenne totale

Une deuxième méthode est celle de pénaliser l'erreur quadratique moyenne totale. Si nous estimons uniquement l'erreur quadratique moyenne du modèle, la quantité obtenue n'est pas un bon estimateur de l'erreur quadratique moyenne de généralisation. L'erreur quadratique moyenne totale a tendance à diminuer lorsque la complexité du modèle augmente. Ainsi, l'erreur quadratique moyenne totale tend à surestimer la qualité du modèle en sous-estimant l'erreur quadratique de généralisation (Larocque, 2006).

$$\text{EQMT} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - f\ (X_{1i}, \dots, X_{ki})\right)^2,$$

où

N : le nombre d'observations totales (le nombre de clients).

Pour tenir compte de la complexité du modèle, l'erreur quadratique moyenne totale doit être pénalisée. Parmi les critères de sélection, il y a le R^2 ajusté, l'Akaike Information Criterion (AIC) et le Bayesian Information Criterion (BIC). Il s'agit de la méthode de sélection du meilleur sous-ensemble. Nous en discuterons davantage au point 4.1.5.

Méthode pas-à-pas (« stepwise »)

Une autre approche est celle des méthodes de sélection de modèles séquentielles : ascendante (« forward »), descendante (« backward ») et pas-à-pas (« stepwise »). Tel que l'indiquent leurs noms, elles font une recherche séquentielle parmi un nombre limité de modèles. Ce sont des méthodes fréquemment utilisées car elles sont rapides et facile d'usage. Parmi les trois méthodes, la méthode pas-à-pas est généralement préférable aux deux autres puisqu'elle permet à une variable de rentrer et de sortir tout au long de l'ajustement du modèle. La sélection se termine lorsqu'il n'est plus possible d'ajouter ou de retirer une variable indépendante dans le modèle.

4.1.5 Choix de la méthode pour la sélection de modèles : erreur quadratique moyenne de généralisation

Pour la sélection de notre modèle, nous combinons la méthode de l'estimation de l'erreur quadratique moyenne de généralisation en divisant l'échantillon en deux et celle qui pénalise l'erreur quadratique moyenne totale. La méthode pas-à-pas n'est pas retenue, même si elle donne souvent un très bon modèle. Le problème est qu'elle ne considère pas tous les modèles possibles. Le modèle proposé n'est pas forcément le meilleur.

La méthode de l'estimation de l'erreur quadratique moyenne de généralisation a pour avantage de permettre l'ajustement du modèle en mode apprentissage et de valider sa performance à l'aide de l'échantillon de validation. Le désavantage est qu'elle peut avoir un certain biais, car elle n'utilise pas toutes les données disponibles dans le modèle. Il y a une perte au niveau de la précision du modèle obtenu (Larocque, 2006). En même temps, cette façon de procéder permet de comparer un modèle à un autre type de modèle (par exemple :

modèle de régression linéaire multiple vs modèle de régression gamma). Puisque la méthode de l'estimation de l'erreur quadratique moyenne de généralisation peut avoir un certain biais pour les modèles d'un même type de régression, c'est la raison pour laquelle nous la jumelons avec la méthode de pénalisation de l'erreur quadratique moyenne totale. Cette dernière a des critères de sélection précis tels que le R^2 ajusté, l'AIC et le BIC qui permettent de choisir le meilleur modèle. Enfin, même si la méthode pas-à-pas n'est pas retenue pour le choix de notre modèle, nous pouvons quand même l'utiliser pour vérifier quel modèle elle proposerait par rapport aux autres critères de sélection.

Critère 1 - Coefficient de détermination ajusté (R^2 ajusté)

Tel que mentionné précédemment, le R^2 n'est pas une mesure adéquate pour sélectionner un modèle. Lorsqu'une variable quelconque est ajoutée au modèle, le R^2 augmente. Le R^2 ajusté est quant à lui opté pour comparer des modèles. Cette mesure diminue si la variable ajoutée ne contribue pas suffisamment à améliorer le modèle. Plus la valeur du R^2 ajusté est élevé, meilleur est le modèle. Toutefois, plus la taille de l'échantillon est grande, plus la différence entre le R^2 ajusté et le R^2 est réduite (Baillargeon et Rainville, 1979). Le critère basé sur le R^2 ajusté mène à sélectionner des modèles sur-spécifiés.

À l'aide de ce critère, le modèle à cinq prédicteurs est retenu. Il comprend les prédicteurs suivants : le nombre d'années-client, le nombre moyen d'interventions, le changement d'équipement, le profil et le type de l'équipement. Le R^2 est de 0,1224 et le R^2 ajusté est de 0,1219. Ce dernier modélise la variable type de l'équipement au niveau global. Puisque nous sommes intéressés par les différents équipements, nous modélisons le modèle à

cinq prédicteurs avec l'ensemble des équipements. Ce modèle devient le modèle M1 sélectionné selon le critère R^2 ajusté est représenté dans l'annexe 1.

Modèle M1 :

profitabilité moyenne =
459,52 - 59,47 an_cl - 47,26 interv + 12,10 chgt - 55,88 profil + 123,14 CEcomm + 16,36 Echaude - 84,85 foyer - 35,07 radiateur

où

an_cl : le nombre années-client,
interv : le nombre moyen d'interventions,
chgt : le changement d'équipement (0 : non, 1 : oui),
profil : le profil (0 : chauffe-eau, 1 : chauffage),
CEcomm : le type de l'équipement est le chauffe-eau commercial,
Echaude : le type de l'équipement est le système de chauffage à l'eau chaude,
foyer : le type de l'équipement est le foyer,
radiateur : le type de l'équipement est le radiateur,
air chaud : la catégorie de référence pour les équipements.

Selon le modèle M1, 12,24% de la variabilité de la variable « profitabilité moyenne » est expliquée par les cinq variables du modèle:

- Plus le nombre d'années-client augmente, plus la profitabilité diminue. Pour chaque augmentation de 1 unité du nombre d'années-client, la profitabilité diminue en moyenne de 59,47$.

- Plus il y a d'interventions, plus la profitabilité diminue. Pour chaque intervention additionnelle, la profitabilité diminue en moyenne de 47,26$.
- Un changement d'équipement fait augmenter la profitabilité de 12,10$ en moyenne.
- Les clients dont le profil de l'équipement est le chauffage ont une profitabilité de 55,88$ de moins que ceux qui possèdent un équipement dont le profil est le chauffe-eau.
- Les clients qui possèdent un chauffe-eau commercial ont une profitabilité de 123,14$ de plus que ceux avec un équipement à l'air chaud.
- Les clients dont le type de l'équipement est le système de chauffage à l'eau chaude ont une profitabilité de 16,36$ de plus que ceux qui possèdent le système de chauffage à l'air chaud.
- Les clients dont le type de l'équipement est le foyer ont une profitabilité de 84,85$ de moins que ceux qui possèdent le système de chauffage à l'air chaud.
- Les clients dont le type de l'équipement est le radiateur ont une profitabilité de 35,07$ de moins que ceux qui possèdent le système de chauffage à l'air chaud.

Critère 2 – Akaike Information Criterion (AIC)

Le critère AIC, introduit par Akaike (1973), découle de la méthode d'estimation basée sur le maximum de vraisemblance. Il est qualifié d' « asymptotiquement efficace » (Shibata, 1981). Si le vrai modèle ne se retrouve pas dans l'ensemble des modèles générés, le critère va choisir celui qui minimise l'erreur moyenne quadratique de prédiction. Le critère AIC tend à choisir des modèles complexes lorsque la taille de l'échantillon tend vers l'infini. Le modèle à privilégier est celui qui donne le AIC le plus faible (Hastie et al.,2009). Par contre, plus la taille de l'échantillon est grande, plusieurs modèles peuvent être

près de la valeur minimale du critère AIC. Le modèle optimal peut ne pas être identifié clairement.

$$AIC = N\left(ln\left(\frac{SCE}{N}\right)\right) + 2(k+1)$$

En tenant compte de ce critère, le modèle retenu comporte cinq prédicteurs, ce sont les mêmes que ceux du R^2 ajusté, soit le modèle M1 : le nombre d'années-client, le nombre moyen d'interventions, le changement d'équipement, le profil et le type de l'équipement.

Critère 3 – Bayesian Information Criterion (BIC)

Le critère BIC, introduit par Schwarz (1978), découle du principe du maximum de vraisemblance comme le critère AIC. Il est un critère dit « asymptotiquement constant » (Shibata, 1981). Si le vrai modèle se retrouve dans l'ensemble des modèles générés, la probabilité que le critère BIC sélectionne le vrai modèle tend vers un lorsque la taille de l'échantillon tend vers l'infini. Si l'échantillon est de taille finie, le BIC a tendance à choisir des modèles trop simples. Ce critère pénalise les modèles complexes, ceux qui ont trop de variables et évite le sur-apprentissage. Le modèle à retenir est celui qui présente le BIC le plus faible (Hastie et al.,2009).

$$BIC = N\left(ln\left(\frac{SCE}{N}\right)\right) + (k+1)\ln(N)$$

En tenant compte de ce critère, le modèle retenu comporte également cinq prédicteurs. Ce sont les mêmes que ceux du R^2 ajusté et de AIC, le modèle M1 : le nombre d'années-client, le nombre moyen d'interventions, le

changement d'équipement, le profil et le type de l'équipement. Les coefficients obtenus sont donc identiques.

4.1.6 Comparaison entre les critères : R^2 ajusté, AIC et BIC

Le critère BIC est celui qui pénalise fortement l'ajout de prédicteurs, suivi du critère AIC (Burnham et Anderson, 2004). Pour que le BIC diminue lorsqu'un prédicteur est ajouté, la somme des carrées des erreurs doit diminuer davantage que pour le AIC (lorsque log(n)>2). Il en est également ainsi pour le critère AIC par rapport au R^2 ajusté (Larocque, 2006). Pour que le AIC diminue par rapport au R^2 ajusté lorsqu'un prédicteur est ajouté, la somme des carrées des erreurs doit diminuer davantage pour le AIC.

Le BIC a tendance à choisir des modèles trop simples (modèles sous-spécifiés), contrairement au R^2 ajusté et le AIC avec leurs modèles plus complexes (modèles sur-spécifiés). Il est préférable d'obtenir un modèle sur-spécifié que sous-spécifié. Les variables sont biaisées avec un modèle sous-spécifié, car la précision des estimés diminue en augmentant les écarts-types. Il est moins pénalisant d'avoir plus de variables dans le modèle que d'en oublier.

Des comparaisons entre le AIC et le BIC ont été effectuées par Yang (2005). Il démontre que le AIC est asymptotiquement optimal dans la sélection de modèles avec l'erreur quadratique moyenne sous l'hypothèse que le vrai modèle ne se retrouve pas dans la liste des modèles candidats. Ce qui n'est pas le cas pour le BIC sous les mêmes conditions. De plus, Yang (2005) démontre que le taux auquel converge le AIC vers le vrai modèle est meilleur que le BIC.

Dans le cas de notre régression linéaire multiple, nous n'avons pas à choisir entre les différents critères étudiés, car ils convergent tous vers le même modèle, celui à cinq prédicteurs, le modèle M1.

4.1.7 Résumé des différents résultats

Dans l'étude, le résultat du critère R^2 ajusté, AIC et BIC est de cinq prédicteurs. Nous avons procédé avec la méthode pas-à-pas et nous obtenons le même résultat. L'ensemble des différents critères utilisés pour la sélection de modèles est résumé dans le tableau 4-3. Parmi tous les modèles à k variables, il s'agit du modèle ayant le meilleur R^2 ajusté, AIC et BIC.

k	R^2 ajusté	AIC	BIC	Pas-à-pas	Variables dans le modèle
1	0.0599	131951	131967		• Nombre moyen d'interventions
2	0.1019	131110	131132		• Nombre moyen d'interventions • Profil d'usage
3	0.1132	131133	131163		• Nombre moyen d'interventions • Profil d'usage • Nombre années-client
4	0.1159	131091	131129		• Nombre moyen d'interventions • Profil d'usage • Nombre années-client • Type d'équipement
5	0.1171	131073	131119	x	• Nombre moyen d'interventions • Profil d'usage • Nombre années-client • Type d'équipement • Changement d'équipement

6	0.1170	131075	131128		• Nombre moyen d'interventions • Profil d'usage • Nombre années-client • Type d'équipement • Changement d'équipement • Génération de l'équipement

Tableau 4-3 : Résumé des modèles selon les différents critères de sélection

4.1.8 Ajout d'interactions pour améliorer le modèle

Dans le but d'améliorer la qualité de notre modèle, il est recommandé de valider si l'ajout de certains termes d'interaction est approprié.

Le nombre d'interventions et le changement d'équipement

Le choix des termes d'interaction à tester nous proviennent du gestionnaire de ABC selon ses connaissances dans le domaine. Pour débuter, examinons si la variable changement d'équipement a une interaction avec le nombre d'interventions. Le gestionnaire de ABC présume que plus l'équipement est neuf, moins il y a d'interventions. Toutefois, un équipement neuf a plus de composantes électroniques qu'un équipement plus ancien. La réparation devient plus complexe et le coût des composantes électroniques est plus dispendieux. Nous voulons valider ce point émis par le gestionnaire de l'entreprise.

Le terme d'interaction « iIntChgt » est significatif. Il y a interaction. L'effet du nombre d'interventions n'est pas le même selon la valeur de la variable changement d'équipement. Pour chaque augmentation du nombre

d'interventions, le profit diminue de 38,71$ s'il n'y a pas eu un changement d'équipement au cours des cinq dernières années. À l'inverse, s'il y a eu un changement d'équipement, le profit diminue davantage, soit de 54,92$. Ce constat vient rejoindre la deuxième présomption du gestionnaire. Un équipement neuf contient plus de composantes électroniques, la réparation est plus coûteuse.

profitabilité moyenne =
115,30 - 38,71 interv + 49,80 chgt - 16,21 iIntChgt

si le changement d'équipement=0 (pas de changement):
profitabilité moyenne = 115,73 - 38,71 interv

si le changement d'équipement=1 (changement):
profitabilité moyenne = 165,10 - 54,92 interv

où

iIntChgt : l'interaction entre le changement d'équipement et le nombre d'interventions.

Le nombre d'interventions et le profil de l'équipement

Le gestionnaire tient aussi valider si l'effet du nombre d'interventions est le même en fonction des deux profils d'usage de l'équipement, chauffe-eau et chauffage. Il y a interaction, car le terme d'interaction « iIntProf » est significatif. L'effet du nombre d'interventions n'est pas le même selon la valeur du profil de l'équipement. Le client dont le profil de l'équipement est le chauffage voit son profit diminuer de 41,10$ pour chaque augmentation du

nombre d'interventions. Si le profil est le chauffe-eau, le profit diminue encore plus, soit de 58,00$ pour chaque augmentation du nombre d'interventions.

profitabilité moyenne =
189,58 - 58,00 interv - 70,28 profil + 16,90 iIntProf

si le profil=0 (chauffe-eau) :
profitabilité moyenne = 189,58 - 58,00 interv
si le profil=1 (chauffage) :
profitabilité moyenne = 119,30 - 41,10 interv

où

iIntProf : l'interaction entre le nombre d'interventions et le profil.

Le nombre d'interventions et la génération de l'équipement

Dans le même ordre d'idées, le gestionnaire veut vérifier si le nombre d'interventions est le même en fonction de la génération de l'équipement. Comme l'année d'installation de l'équipement est manquante dans 70% des cas, la génération nous donne une idée approximative de l'âge de l'équipement. Les nouvelles interactions « iGen1_int » et « iGen2_int » créées sont significatives. L'effet du nombre d'interventions n'est pas le même selon la génération de l'équipement. Le client dont la génération de l'équipement est la première voit son profit diminuer de 20,70$ pour chaque augmentation du nombre d'interventions par rapport à la troisième génération. Si c'est la

deuxième génération, le profit diminue de 13,32$ pour chaque augmentation du nombre d'interventions par rapport à la troisième génération.

profitabilité moyenne =
143,41 - 60,47 interv - 5,42 gen1 - 20,70 iIntGen1

profitabilité moyenne =
141,54 - 44,84 interv - 12,29 gen2 - 13,32 iIntGen2

où

gen1 : la première génération de l'équipement,
gen2 : la deuxième génération de l'équipement,
iIntGen1 : l'interaction entre le nombre d'interventions et la première génération,
iIntGen2 : l'interaction entre le nombre d'interventions et la deuxième génération.

Le nombre d'interventions et le type de l'équipement

L'effet du nombre d'interventions n'est pas le même selon le type de l'équipement. Seuls les termes d'interaction entre le nombre d'interventions avec le radiateur et le foyer ne sont pas significatifs. Lorsque le nombre d'interventions augmente en présence d'un chauffe-eau commercial et d'un chauffe-eau résidentiel, la profitabilité diminue de 58,54$ et de 17,97$ respectivement par rapport à un système de chauffage à l'air chaud. Si c'est en lien avec le système de chauffage à l'eau chaude, le profit augmente alors de 8,74$ par rapport au système de chauffage à l'air chaud.

profitabilité moyenne =
139,23 - 46,00 interv + 225,59 CEcomm - 58,54 iInt_CEcomm

profitabilité moyenne =
119,70 - 40,72 interv + 69,19 CEres - 17,97 iInt_CEres

profitabilité moyenne =
145,73 - 48,07 interv - 19,67 Echaude + 8,74 iInt_Echaude

où

CEres : le chauffe-eau résidentiel,
iInt_CEcomm : l'interaction entre le nombre d'interventions et le chauffe-eau commercial,
iInt_CEres : l'interaction entre le nombre d'interventions et le chauffe-eau résidentiel,
iInt_Echaude : l'interaction entre le nombre d'interventions et le système de chauffage à l'eau chaude.

Le changement d'équipement et la génération de l'équipement

L'effet de la profitabilité n'est pas la même selon le changement d'équipement et la génération de l'équipement. Pour un équipement de première génération, s'il y a eu changement, le profit augmente de 22,43$ par rapport à la troisième génération. S'il n'y a pas eu de changement, le profit est moindre, soit à 11,59$. Dans le cas d'un équipement de deuxième génération, le profit augmente de 38,62$ pour un changement par rapport à la troisième génération.

S'il n'y a pas eu de changement pour un équipement de la deuxième génération, le profit demeure pratiquement le même.

profitabilité moyenne =
70,01 + 32,39 chgt + 11,59 gen1 +10,84 iChgtGen1

si le changement d'équipement=0 (pas de changement):
profitabilité moyenne = 70,01 + 11,59 gen1

si le changement d'équipement=1 (changement):
profitabilité moyenne = 102,24 + 22,43 gen1

profitabilité moyenne =
83,31 + 42,39 chgt + 0,04 gen2 - 38,66 iChgtGen2

si le changement d'équipement=0 (pas de changement):
profitabilité moyenne = 83,31 + 0,04 gen2

si le changement d'équipement=1 (changement):
profitabilité moyenne = 125,70 + 38,62 gen2

où

iChgtGen1 : l'interaction entre le changement et la première génération,
iChgtGen2 : l'interaction entre le changement et la deuxième génération.

Le changement d'équipement et le profil de l'équipement

Quant aux variables changement et profil, le terme d'interaction «iChgtProf » est significatif. Il y a interaction. L'effet de la profitabilité n'est pas la même selon le changement d'équipement et le profil de l'équipement.

Lorsqu'il n'y a pas eu de changement d'équipement, le profit est de 43,83$ et de 88,51$ pour le profil chauffe-eau et chauffage respectivement. S'il y a eu un changement d'équipement le profit est de 148,50 $ et de 92,73$ pour le profil chauffe-eau et chauffage respectivement. Donc, lors d'un changement d'équipement, peu importe le profil de l'équipement, le profit est plus élevé que s'il n'y avait pas eu de changement.

- si le changement d'équipement=0 (pas de changement) et le profil=0 (chauffe-eau) :
  profitabilité moyenne = 43,83

- si le changement d'équipement=0 (pas de changement) et le profil=1 (chauffage):
  profitabilité moyenne = 88,51

- si le changement d'équipement=1 (changement) et le profil=0 (chauffe-eau) :
  profitabilité moyenne = 148,50

- si le changement d'équipement=1 (changement) et le profil=1 (chauffage) :
  profitabilité moyenne = 92,73

où

iChgtProf: l'interaction entre le changement d'équipement et le profil.

La génération de l'équipement et le type de l'équipement

L'effet de la profitabilité n'est pas la même selon la génération de l'équipement et le type de l'équipement. La profitabilité d'un équipement de chauffe-eau commercial de première génération est de 233,30$. Alors que celui de la deuxième génération engendre une perte de 222,86$. Quant à un chauffe-eau résidentiel de première génération, la perte est de 15,72$. Le profit d'un système de chauffage à l'eau chaude de première et de deuxième génération est de 19,25$ et de 19,67$ respectivement.

profitabilité moyenne =
107,30 + 7,11 gen1 - 37,42 CEcomm + 233,30 iGen1CEcomm

profitabilité moyenne =
108,30 -17,54 gen2 + 195,84 CEcomm - 222,86 iGen2CEcomm

profitabilité moyenne =
84,28 + 9,06 gen1 + 67,32 CEres - 15,72 iGen1CEres

profitabilité moyenne =
104,69 + 17,19 gen1 - 45,13 Echaude + 19,25 iGen1Echaude

profitabilité moyenne =
119,30 -27,76 gen2 - 25,18 Echaude + 19,67 iGen2Echaude

où

iGen1CEcomm : l'interaction entre la première génération et le chauffe-eau commercial,
iGen2CEcomm : l'interaction entre la deuxième génération et le chauffe-eau commercial,
iGen1CEres : l'interaction entre la première génération et le chauffe-eau résidentiel,
iGen1Echaude : l'interaction entre la première génération et le système de chauffage à l'eau chaude,
iGen2Echaude : l'interaction entre la deuxième génération et le système de chauffage à l'eau chaude.

Intégration des termes d'interaction dans la régression linéaire multiple

Tous les termes d'interaction décrits précédemment qui sont significatifs sont intégrés dans le modèle M1. Par contre, si ces termes d'interaction une fois dans le modèle s'avèrent non significatifs, ils sont retirés. Les paramètres du modèle avec les interactions sont décrits dans l'annexe 2. Pour les références futures, nous le nommons modèle M2.

Modèle M2 :

profitabilité moyenne =
380,77 - 62,04 an_cl - 67,32 interv + 127,40 chgt + 56,20 profil + 90,12 CE_comm - 60,10 Echaude - 96,00 foyer - 49,95 radiateur - 26,14 gen1 - 11,04 gen2 + 20,16 iIntGen1 -64,26 iInt_CEcomm + 14,20 iInt_Echaude - 110,01 iChgtProfil + 152,24 iChgt_CEcomm - 12,79 iChgt_Echaude + 74,30 iGen1Echaude + 60,76 iGen2Echaude

Le R^2 est de 0,1345 pour le modèle M2. Ce qui signifie que 13,45% de la variabilité de la variable « profitabilité moyenne » est expliquée par les différentes variables du modèle.

En termes d'interprétation, nous pouvons dire que :

- Plus le nombre d'années-client augmente, plus la profitabilité diminue. Pour chaque augmentation de 1 unité du nombre d'années-client, la profitabilité diminue en moyenne de 62,04$.
- Plus il y a d'interventions, plus la profitabilité diminue. Pour chaque intervention additionnelle, la profitabilité diminue en moyenne de 67,32$.
- Un changement d'équipement fait augmenter la profitabilité de 127,40$ en moyenne par rapport à un non changement.
- Les clients dont le profil de l'équipement est le chauffage ont une profitabilité de 56,20$ en moyenne de plus que ceux qui ont le profil chauffe-eau.
- Les clients qui possèdent un chauffe-eau commercial ont une profitabilité de 90,12$ de plus que ceux avec le système de chauffage à l'air chaud, mais cette variable n'est pas significative dans le modèle. Elle est présente, car elle est impliquée dans un terme d'interaction qui est significatif.

- Les clients dont le type de l'équipement est le système de chauffage à l'eau chaude, le foyer et le radiateur ont une profitabilité de 60,10$, 96,00$ et 49,95$ de moins respectivement que ceux qui possèdent le système de chauffage à l'air chaud.
- Les clients qui possèdent un équipement dont la génération est la première et la deuxième ont une profitabilité de 26,14$ et de 11,04$ de moins respectivement qu'un équipement de la troisième génération.
- L'effet du nombre d'interventions n'est pas le même selon la génération de l'équipement. Le client dont la génération de l'équipement est la première voit son profit diminuer de 20,15$ pour chaque intervention additionnelle par rapport à la troisième génération.
- L'effet du nombre d'interventions n'est pas le même selon le type de l'équipement. Le client dont l'équipement est le chauffe-eau commercial voit son profit diminuer de 64,26$ pour chaque intervention additionnelle par rapport au système de chauffage à l'air chaud. Si l'équipement est le système de chauffage à l'eau chaude, le profit augmente de 14,20$ pour chaque intervention additionnelle par rapport au système de chauffage à l'air chaud.
- Lorsqu'il n'y a pas eu de changement d'équipement et que le profil de l'équipement est le chauffage, le profit est de 436,97$. S'il y a eu changement d'équipement et que le profil est le chauffage, le profit est de 454,36$.
- S'il y a eu un changement d'équipement du chauffe-eau commercial, le profit est de 623,12$. Si le changement est en lien avec le système de chauffage à l'eau chaude, le profit est de 307,88$.
- Le profit d'un système de chauffage à l'eau chaude de première et de deuxième génération est de 74,30$ et de 60,76$ respectivement.

D'une perspective affaires, prenons le cas d'un client de cinq ans d'ancienneté avec l'entreprise et qui n'a pas changé son équipement. Si le profil de son

équipement est le chauffage et qu'il a une seule intervention, le profit est de 10,09$. Dès la deuxième intervention, l'entreprise encaisse déjà une perte de 57,23$ et s'il y a une troisième intervention, la perte est estimée à 124,55$. Pour les mêmes variables, changeons le profil chauffage pour le chauffe-eau. Dans le cas d'une première intervention, le profit est de 66,29$. À la deuxième intervention, nous sommes en présence du seuil critique, la perte est de 1,03$. Et à la troisième, la perte est de 68,35$. Quant à la comparaison d'un équipement avec le système de chauffage à l'air chaud, le chauffe-eau commercial représente celui qui a la plus grande profitabilité alors que le foyer est celui qui génère la plus grande perte.

4.1.9 Comparaison entre les modèles M1 et M2

Dans les sections précédentes, nous avons discuté des critères de sélection de modèles en termes de R^2 ajusté, de AIC et de BIC pour la régression linéaire multiple. Nous avons retenu deux modèles, celui de base M1 et celui avec interactions, M2. Le tableau 4-4 compare les deux modèles en fonction des critères de sélection ainsi que l'erreur quadratique moyenne et l'erreur de prévision en mode d'entraînement. L'erreur de prévision est estimée à partir de la racine carrée de l'erreur quadratique moyenne. Plus l'erreur est petite, meilleure est la performance du modèle. Selon les résultats, le modèle M2 est plus performant que le modèle M1 sur tous les critères de sélection et au niveau des erreurs.

Modèle	R^2	R^2 ajusté	AIC	BIC	EQM	Erreur de prévision
M1	0,1224	0,1219	149188	149247	13797,37	117,46
M2	0,1345	0,1335	148890	149136	13606,86	116,65

Tableau 4-4 : Comparaison des critères de sélection et des erreurs entre M1 et M2

4.2 Régression gamma : Modèles M3 et M4

4.2.1 Modélisation

Pour obtenir un autre point de vue sur notre modèle de profitabilité, il est intéressant de construire un modèle selon la régression gamma. La régression gamma établit la relation entre le logarithme de la variable dépendante et les variables indépendantes. Pour appliquer cette régression, les deux conditions suivantes doivent idéalement être respectées:

1 – la variable dépendante doit avoir des valeurs strictement positives,

2 – la distribution de la variable dépendante est asymétrique.

Dans notre étude, la variable dépendante « profitabilité moyenne » contient des valeurs positives et des valeurs négatives. Afin de remédier à la situation, la variable dépendante a été translatée pour que les valeurs négatives deviennent positives. La valeur de la plus petite profitabilité moyenne est (1 812,78$). Pour garder la même distribution de profitabilité, nous avons additionné 1 812,78$ à toutes les valeurs de la profitabilité. Ainsi, les données ne font que translater de 1 812,78$. La régression gamma qui se base sur une fonction logarithmique peut alors être appliquée. L'interprétation des coefficients de la régression gamma se fait en termes de pourcentages car les valeurs d’un logarithme sont comprises entre 0 et 1.

Le modèle de régression gamma est décrit sous la forme suivante :

$$Log(E[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + + \beta_k X_k + \varepsilon ,$$

où

Log(E[Y]) : le logarithme de l'espérance de la profitabilité moyenne,
k : le nombre de variables indépendantes,
$X_1, \ldots X_k$: les variables indépendantes,
$\beta_0, \ldots, \beta_k$: les paramètres du modèles à estimer,
ε: l'erreur des résidus.

4.2.2 Sélection de modèles

Tout comme la régression linéaire multiple, les critères AIC et BIC aident à sélectionner le meilleur modèle. Comme il n'est pas possible de sélectionner directement le modèle à l'aide de ces critères dans SAS, nous devons procéder autrement. Nous optons pour une méthode séquentielle descendante. Pour débuter, toutes les variables sont incluses dans le modèle. Ensuite, la variable dont la « p-value » est la plus grande est retirée. Le processus est répété jusqu'à ce que toutes les variables soient retirées. À chaque étape, nous déterminons le AIC et BIC du modèle obtenu. Notons qu'en général, cette procédure ne garantit pas que nous allons ainsi trouver les meilleurs modèles selon AIC et BIC parmi tous les modèles possibles. En procédant ainsi, c'est le modèle à six prédicteurs qui donnent le plus petit résultat pour les deux critères tel que présenté dans le tableau 4-5. Le modèle de la régression gamma quant à lui est décrit dans l'annexe 3. Pour les références futures, nous nommons ce modèle M3.

k	AIC	BIC	Variables dans le modèle
6	176219	176279	• Nombre moyen d'interventions • Profil d'usage • Nombre années-client • Type d'équipement • Changement d'équipement • Génération de l'équipement
5	176230	176283	• Nombre moyen d'interventions • Profil d'usage • Nombre années-client • Type d'équipement • Génération de l'équipement
4	197073	196119	• Nombre moyen d'interventions • Profil d'usage • Nombre années-client • Type d'équipement
3	196077	196115	• Nombre moyen d'interventions • Profil d'usage • Nombre années-client
2	196343	196374	• Nombre moyen d'interventions • Profil d'usage
1	196979	197001	• Nombre moyen d'interventions

Tableau 4-5 : Résumé des modèles selon les critères de sélection AIC et BIC

Modèle M3 :

Log(E[profitabilité moyenne] =

7,73 - 0,0279 an_cl - 0,0249 interv + 0,0063 chgt - 0,0285 profil + 0,0622 CEcomm + 0,0088 Echaude - 0,0448 foyer - 0,0203 radiateur - 0,0021 gen1 - 0,0030 gen2

Le fait d'avoir changé d'équipement au cours des cinq dernières années n'a pas une incidence forte sur la profitabilité, car le profit du client augmente

seulement de 1% par rapport à un non changement. Pour chaque interventions additionnelle, la profitabilité diminue de 2%. Pour chaque augmentation du nombre d'années-client, la profitabilité diminue de 3%. Posséder un profil chauffage fait diminuer la profitabilité de 3% par rapport au profil chauffe-eau. Avoir un chauffe-eau commercial augmente la profitabilité de 6% par rapport à un système de chauffage à l'air chaud. Avoir un système de chauffage à l'eau chaude augmente la profitabilité de 1% par rapport à un équipement à l'air chaud. Posséder un foyer fait diminuer la profitabilité de 4%. Posséder un radiateur fait diminuer la profitabilité de 2%. Un équipement de la première ou de la deuxième génération n'a aucune incidence sur la profitabilité, ces deux variables ne sont pas significatives.

4.2.3 Ajout d'interactions pour améliorer le modèle

Encore dans l'optique d'améliorer le modèle, l'ajout des termes d'interactions s'avère une option intéressante. Les interactions dans une régression gamma sont les mêmes qu'en régression linéaire multiple. Il s'agit de faire le produit entre deux variables. Nous reprenons les mêmes interactions effectuées pour la régression linéaire multiple que nous appliquons dans la régression gamma.

Le nombre d'interventions et le changement d'équipement

Débutons avec l'interaction entre le changement et le nombre d'interventions. Le paramètre d'interaction «ıIntChgt » est significatif. La profitabilité n'est pas la même en fonction du nombre d'interventions et le changement d'équipement ou non au cours des cinq dernières années. Dans le cas qu'il n'y a pas eu de

changement, pour chaque intervention additionnelle, le profit diminue de 1% par rapport à un changement.

Log(E[profitabilité moyenne] =
7,56 - 0,0207 interv + 0,0257 chgt - 0,0082 iIntChgt

Le nombre d'interventions et le profil de l'équipement

L'effet du nombre d'interventions n'est pas le même selon le profil de l'équipement. Dans le cas du profil chauffage, pour chaque intervention additionnelle, le profit augmente de 1% par rapport au profil de chauffe-eau.

Log(E[profitabilité moyenne] =
7,60 - 0,0299 interv - 0,0358 profil + 0,0081 iIntProf

Le nombre d'interventions et la génération de l'équipement

L'effet du nombre d'interventions n'est pas le même selon la génération de l'équipement. Dans le cas d'un équipement de la première génération, pour chaque intervention additionnelle, le profit augmente de 1% par rapport à un équipement de la troisième génération. Quant à la deuxième génération, le profit diminue de 1%.

Log(E[profitabilité moyenne] =
7,58 - 0,0323 interv - 0,0032 gen1 + 0,0114 iIntGen1

Log(E[profitabilité moyenne] =
7,58 - 0,0236 interv - 0,0062 gen2 - 0,0073 iIntGen2

Le nombre d'interventions et le type de l'équipement

L'effet du nombre d'interventions n'est pas le même selon l'équipement. Dans le cas d'un équipement de chauffe-eau commercial, pour chaque intervention additionnelle, le profit diminue de 3% par rapport à un équipement à un système de chauffage à l'air chaud. Pour le chauffe-eau résidentiel, le profit diminue de 1%. Pour le système de chauffage à l'eau chaude, cela n'affecte pas le profit, car le coefficient de 0,0044 n'est pas statistiquement significatif.

Log(E[profitabilité moyenne] =
7,57 - 0,0242 interv + 0,1137 CEcomm - 0,0299 iIntCEcomm

Log(E[profitabilité moyenne] =
7,57 - 0,0216 interv - 0,0353 CEres - 0,0086 iIntCEres

Log(E[profitabilité moyenne] =
7,58 - 0,0252 interv - 0,0101 Echaude + 0,0044 iIntEchaude

Le changement d'équipement et le profil de l'équipement

La profitabilité varie en fonction du changement d'équipement ou non et le profil de l'équipement. S'il y a eu changement d'équipement et que le profil est le chauffage, le profit est de 3% plus élevé par rapport à un non changement d'équipement et un profil chauffe-eau. S'il y a eu changement d'équipement mais que le profil est le chauffe-eau, le profit est de 6% plus élevé par rapport à un non changement et un profil chauffage.

Log(E[profitabilité moyenne] =
7,53 + 0,0238 profil + 0,0548 chgt - 0,0527 iChgtProf

Le changement d'équipement et la génération de l'équipement

La profitabilité varie selon qu'il y ait eu un changement d'équipement ou non et la génération de l'équipement. S'il y a eu changement d'équipement et que c'est un équipement de première génération, le profit augmente de 1 % par rapport à un non changement d'équipement et un équipement de troisième génération. Si c'est un équipement de la deuxième génération, le profit diminue de 2%.

Log(E[profitabilité moyenne] =
7,54 + 0,017 chgt + 0,0061 gen1 + 0,0055 iChgtGen1

Log(E[profitabilité moyenne] =
7,55 + 0,022 chgt + 0,0 gen2 - 0,0201 iChgtGen2

Le changement d'équipement et le type de l'équipement

L'effet de la profitabilité n'est pas le même en fonction du changement d'équipement ou non au cours des cinq dernières années et le type de l'équipement en question. Dans le cas qu'il y a eu un changement, le profit augmente de 13% et de 5% respectivement pour le chauffe-eau commercial et le chauffe-eau résidentiel. Quant au système de chauffage à l'eau chaude, le profit diminue de 3% par rapport à un non changement et à un système de

chauffage à l'air chaud. Le terme d'interaction entre le nombre d'interventions avec foyer et le radiateur ne sont pas significatifs.

Log(E[profitabilité moyenne] =
7,55 + 0,0184 chgt - 0,038 CEcomm + 0,1219 iChgtCEcomm

Log(E[profitabilité moyenne] =
7,55 + 0,0033 chgt - 0,0201 CEres + 0,0473 iChgtCEres

Log(E[profitabilité moyenne] =
7,55 + 0,0269 chgt + 0,008 Echaude - 0,0295 iChgtEchaude

Le profil de l'équipement et la génération de l'équipement

L'effet de la profitabilité n'est pas le même en fonction du profil de l'équipement et la génération de l'équipement. Dans le cas du profil chauffage avec un équipement de la deuxième génération, le profit augmente de 2% par rapport à un profil chauffe-eau et un équipement de la troisième génération. L'interaction entre le profil et la première génération n'est pas significative.

Log(E[profitabilité moyenne] =
7,58 - 0,0298 profil - 0,0172 gen2 + 0,0167 iProfGen2

Le type de l'équipement et la génération de l'équipement

L'effet de la profitabilité n'est pas le même en fonction du type de l'équipement et la génération de l'équipement. Notons que le foyer et le radiateur n'ont pas de génération.

Par rapport à un système de chauffage à l'air chaud de la troisième génération, le profit augmente de 12% pour un chauffe-eau commercial de la première génération. S'il s'agit d'un chauffe-eau commercial de la deuxième génération, le profit diminue de 11%. Pour un chauffe-eau résidentiel de la première génération, le profit diminue de 1%. L'interaction entre le chauffe-eau résidentiel et la deuxième génération n'est pas significative. Et que ce soit un système de chauffage à l'eau chaude de la première ou de la deuxième génération, le profit augmente de 1%.

Log(E[profitabilité moyenne] =
7,56 + 0,0037 gen1 - 0,0197 CEcomm + 0,1168 iGen1CEcomm

Log(E[profitabilité moyenne] =
7,57 - 0,0092 gen2 + 0,0971 CEcomm - 0,1114 iGen2CEcomm

Log(E[profitabilité moyenne] =
7,55 + 0,0048 gen1 + 0,0349 CEres - 0,0082 iGen1CEres

Log(E[profitabilité moyenne] =
7,56 + 0,0089 gen1 - 0,0238 Echaude + 0,0103 iGen1Echaude

Log(E[profitabilité moyenne] =
7,57 - 0,0145 gen2 - 0,0131 Echaude + 0,0084 iGen2Echaude

Intégration des termes d'interaction significatifs dans la régression gamma

Tous les termes d'interaction décrits précédemment qui sont significatifs sont intégrés dans le modèle M3. Si ces termes d'interaction une fois dans le modèle s'avèrent non significatifs, ils sont retirés. Les paramètres du modèle de la régression gamma avec les termes d'interaction sont décrits dans l'annexe 4. Pour les références futures, nous le nommons modèle M4.

Modèle M4 :

Log(E[profitabilité moyenne] =
7,68 - 0,0291 an_cl - 0,0354 interv + 0,0730 chgt + 0,0348 profil + 0,0360 CEcomm - 0,0298 Echaude - 0,0470 foyer - 0,0247 radiateur - 0,0126 gen1 + 0,0008 gen2 + 0,0107 iIntGen1 - 0,0602 iChgtProf - 0,0289 iIntCEcomm + 0,0070 iIntEchaude - 0,0103 iChgtGen2 - 0,0097 iChgtEchaude + 0,0892 iGen1CEcomm + 0,0385 iGen1Echaude + 0,0330 iGen2Echaude

Le fait que le client ait changé son équipement au cours des cinq dernières années, le profit augmente de 8%. Plus il y a d'interventions sur l'équipement, plus la profitabilité diminue. Pour chaque augmentation du nombre années-client et chaque intervention additionnelle, le profit diminue de 3%. Un profil chauffage fait augmenter le profit de 4% par rapport au profil chauffe-eau. Le chauffe-eau commercial fait augmenter le profit de 4%, mais cette variable n'est pas significative. Elle est présente dans le modèle, car elle est impliquée dans des termes d'interaction qui sont significatifs. Le système de chauffage à l'eau chaude, le foyer et le radiateur font diminuer le profit de 3%, 5%, 4 % et

de 2% respectivement par rapport à un système de chauffage à l'air chaud. La génération de l'équipement semble avoir peu d'incidence sur la profitabilité, car la première génération fait diminuer le profit de 1% alors que la deuxième, la fait augmenter de 1%. Il est à noter que la deuxième génération n'est pas une variable significative. Elle est présente dans le modèle, car elle est impliquée dans des termes d'interaction qui sont significatifs.

La profitabilité varie en fonction d'un changement d'équipement et du profil de l'équipement. S'il y a eu un changement d'équipement et que le profil est le chauffage, le profit augmente de 5%. Si le profil est le chauffe-eau, le profit augmente de 8%. Il en est également ainsi pour le changement et le type de l'équipement. S'il y a eu un changement et que l'équipement est le système de chauffage à l'eau chaude, le profit diminue de 1%. Un changement d'équipement avec un équipement de la deuxième génération fait diminuer le profit de 1%.

L'effet de la profitabilité diffère selon le nombre d'interventions avec le type de l'équipement et la génération de l'équipement. Pour chaque intervention additionnelle en lien avec le chauffe-eau commercial ou le système de chauffage à l'eau chaude, le profit diminue de 3% et augmente de 1% respectivement. Quant à l'interaction entre le nombre d'interventions et la première génération de l'équipement, le profit augmente de 1%.

Finalement, la profitabilité n'est pas la même selon le type de l'équipement et la génération de l'équipement. Pour un chauffe-eau commercial de la première génération, le profit augmente de 9%. Quant au système de chauffage à l'eau chaude de la première génération, le profit augmente de 4%. Le profit augmente de 3% si c'est le même système de chauffage, mais de deuxième génération.

4.2.4 Comparaison entre les modèles M3 et M4

Dans les sections précédentes, nous avons discuté des critères de sélection de modèles en termes de AIC et de BIC pour la régression gamma. Nous avons retenu deux modèles, celui de base M3 et celui avec interactions, M4. Le tableau 4-6 compare les deux modèles en fonction des critères de sélection en mode d'entraînement. Selon les résultats, le modèle M4 est plus performant que le modèle M3 sur tous les critères de sélection. Notons que c'est la même conclusion que pour la régression linéaire multiple. Le modèle avec les interactions est plus performant que le modèle de base.

Modèle	AIC	BIC
M3	195928	196019
M4	195751	195911

Tableau 4-6 : Comparaison des critères de sélection entre M3 et M4

4.2.5 Comparaison avec la régression linéaire multiple entre les modèles M2 et M4

En règle générale, les analyses de régression sur des valeurs positives se font à l'aide de la régression linéaire et de la régression gamma (Firth, 1988; Das et Lee, 2009). Le point à soulever est qu'habituellement les valeurs positives ont une distribution non normale et que la variance peut être constante ou non. Das (2011) compare trois cas réels pour les deux types de régression. Deux cas démontrent que la régression linéaire et la régression gamma arrivent pratiquement aux mêmes résultats. Pour l'autre cas, la régression gamma performe légèrement mieux que la régression linéaire.

Selon nos résultats, les modèles avec les interactions sont plus performants que les modèles de base. Il est intéressant de savoir lequel des deux est celui qui

prédit le mieux la profitabilité de la clientèle, entre le modèle de la régression linéaire multiple et de la régression gamma. Pour ce faire, nous comparons l'erreur de prévision en mode d'entraînement. L'erreur de prévision est estimée à partir de la racine carrée de l'erreur quadratique moyenne. Les deux erreurs sont présentées dans le tableau 4-7. Plus l'erreur est petite, meilleure est la performance du modèle. La comparaison entre le modèle de régression gamma à celui de la régression linéaire multiple se fait sans aucun biais, car notre échantillon est divisé en deux parties. L'erreur du modèle M4 est légèrement plus petite que le modèle M2 (116,65 vs 116,49). Nous constatons d'un point de vue pratique que la différence est négligeable. Nous pouvons donc nous baser sur le modèle qui est le plus facile à interpréter : la régression linéaire multiple.

Modèle	EQM	Erreur de prévision
M2	13606,86	116,65
M4	13569,80	116,49

Tableau 4-7 : Comparaison des erreurs entre M2 et M4

4.3 Régression logistique : Modèles M5 et M6

Une autre méthode d'apprentissage supervisé que nous pouvons considérer est la régression logistique. Elle s'apparente à la régression linéaire car ses objectifs sont les mêmes : comprendre l'influence des variables indépendantes sur la variable dépendante et prévoir les valeurs futures de la variable dépendante à partir des variables indépendantes. Contrairement à la régression linéaire qui peut avoir plus d'une valeur, la variable réponse ne peut prendre que deux valeurs. Dans l'étude, nous avons transformé la variable « prof_moy » en deux : 0 pour les clients non profitables et 1 pour les clients profitables. Nous présentons dans cette section les concepts reliés à cette méthode.

4.3.1 Modélisation

Tel que nous l'avons introduit, le concept de la régression logistique est de tester un modèle de régression pour la laquelle la valeur de la variable dépendante est dichotomique. Elle prend la valeur 0 ou 1. Le poids de chaque variable indépendante est représenté par un coefficient de régression. C'est ce qui permet au modèle de prédire la probabilité qu'un événement survienne (valeur 1) ou non (valeur 0). Puisque la variable dépendante est dichotomique, l'espérance est la probabilité que la variable dépendante soit égale à 1. En ajustant le modèle de la régression linéaire avec une variable dichotomique, le modèle obtenu serait le suivant :

$$P(Y = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

où

P(Y=1) : la probabilité que Y =1,
$X_1, \ldots X_k$: les variables indépendantes,
$\beta_0, \ldots, \beta_k$: les paramètres du modèle à estimer.

Toutefois, le modèle ajusté ci-dessus est inadéquat. La problématique est que la variable dépendante ne peut prendre que deux valeurs. Pour remédier à la situation, il s'agit de transformer le côté gauche du modèle afin qu'il puisse prendre les mêmes valeurs que le côté droit. Ainsi, le modèle de la régression logistique est exprimé en termes de probabilité. L'interprétation est entermes de odds, car nous comparons la probabilité que Y=1 avec la probabilité que Y=0.

Le modèle de régression logistique est décrit sous la forme suivante :

$$P(Y = 1) = \frac{1}{1 + exp(\beta_0 - \beta_1 X_1 - X_2 - \cdots - \beta_k X_k)}$$

où

P(Y=1) : la probabilité que Y arrive,
X_1, …X_k : les variables indépendantes,
β_0,…, β_k : les paramètres du modèle à estimer.

La méthode la plus couramment utilisée pour estimer les paramètres de la régression logistique est le maximum de vraisemblance (« maximum likelihood »). Le choix des coefficients de la régression logistique se basent sur l'obtention des valeurs prédites de la variable indépendante qui sont le plus près des valeurs observées. Le modèle considéré comme étant de base est celui dont la catégorie (0 ou 1) possède la fréquence la plus élevée. Tout comme la régression linéaire, les critères AIC et BIC servent aussi à sélectionner le modèle (Larocque, 2006).

4.3.2 Comparaison avec la régression linéaire

Pohar et al. (2004) soutiennent que la régression logistique est une méthode robuste, flexible et facile à utiliser. Le but est de trouver le meilleur modèle et le plus parcimonieux pour décrire la relation entre la variable dépendante et les variables indépendantes. Elle n'émet aucune hypothèse concernant la distribution des variables explicatives contrairement à la régression linéaire. Les auteurs affirment que la régression linéaire donne de meilleurs résultats lorsque la distribution des variables dépendantes est normale. Une distribution est dite normale lorsque la distribution des données est symétrique par rapport à

la moyenne. Toutefois, lorsque la taille de l'échantillon est grande, les deux types de régression s'équivalent. Dans le cas pour laquelle la distribution des variables indépendantes n'est pas normale, il est plus approprié d'utiliser la régression logistique. En fait, la régression logistique donne de bons résultats peu importe le type de distribution des données. Hastie et al. (2009) arrivent à la même conclusion. Pour eux, la régression logistique est un modèle plus robuste que la régression linéaire peu importe le type de distribution des données.

Puisque nous sommes intéressés par la profitabilité du client, la régression logistique permet de prévoir à l'aide des variables explicatives, si le client est profitable ou ne l'est pas. Il n'y a que deux réponses possibles. La régression linéaire multiple quant à elle permet de quantifier la valeur de la profitabilité. L'avantage pour ABC de recourir à la régression logistique est de permettre dans un premier temps de prévoir si le client est profitable ou ne l'est pas. Si le client prédit est profitable, ABC pourrait quantifier sa valeur de profitabilité en termes monétaires à l'aide de la régression linéaire multiple. Nous allons étudier davantage le modèle combiné de la régression logistique et de la régression linéaire multiple à la section 4.4.

4.3.3 Sélection de modèles

Les critères AIC et BIC servent de critère pour sélectionner le meilleur modèle comme dans le cas de la régression linéaire et de la régression gamma. Pour y parvenir, nous devons trouver les meilleurs modèles possibles de une variable jusqu'aux six variables que nous étudions. L'option « score » dans SAS trouve le meilleur modèle à chacune des étapes jusqu'au modèle à six variables. Le « score » est le gain amené par le modèle par rapport à un modèle sans variable explicative. Ensuite, chacun des six modèles est ajusté pour déterminer les

meilleurs modèles en fonction des critères AIC et BIC tel que présenté dans l'annexe 5. Le modèle à cinq prédicteurs répond à ces critères, nous le nommons M5. Le modèle de la régression logistique M5 comprend le nombre moyen d'interventions, le profil, le type de l'équipement, le nombre d'années-client et la génération de l'équipement qui est détaillé dans l'annexe 6. Le meilleur point de coupure est à 0,52 pour lesquelles 89,8% des observations sont prédites correctement. Dans l'échantillon, 10,2% des données réelles sont des valeurs égales à 0 qui représentent les clients non profitables et 89,8% sont égales à 1, les clients profitables. Au-delà du taux de bonne classification, d'autres mesures sont intéressantes à analyser : la sensitivité et la spécificité. Ce sont deux mesures qui permettent d'évaluer la performance du modèle à détecter les valeurs à Y=0 et à Y=1. La sensitivité est la probabilité qu'une observation appartenant à la catégorie Y=1 soit bien classifiée. Quant à la spécificité, elle est la probabilité qu'une observation appartenant à la catégorie Y=0 soit bien classifiée. La sensitivité de M5 est de 99% et la spécificité est de 11%. Il est à noter que plus le point de coupure est élevé, plus la sensitivité diminue et à l'inverse, la spécificité augmente.

Modèle M5 :

$$log\left(\frac{p}{1-p}\right) = 1{,}321 - 0{,}175\text{inter} + 1{,}603\text{an\_cl} - 3{,}543\text{profil} - 0{,}703\text{AChaud} - 0{,}167\text{CEcomm} + 0{,}6324\text{CEres} - 1{,}394\text{gen1} - 1{,}082\text{gen2}$$

Le modèle M5 stipule que lorsque le nombre années-client augmente d'un an, l'odds de s'avérer être un client profitable augmente de 60%. Lorsque le nombre d'interventions augmente de 1, l'odds d'être profitable diminue de 18%. L'odds de profitabilité du profil chauffage augmente de 88% par rapport

au profil chauffe-eau. L'odds de profitabilité du système de chauffage à l'air chaud augmente de 44% alors que celui du chauffe-eau commercial diminue de 34%. L'odds de profitabilité de la première génération est augmentée de 22% contrairement à la première génération qui diminue de 94%.

4.3.4 Ajout d'interactions pour améliorer le modèle

Encore dans l'optique d'améliorer le modèle, l'ajout d'interactions s'avère une option intéressante. En ajoutant les termes d'interaction, la plupart des variables s'avèrent non significatives, tel qu'illustré dans l'annexe 7. Nous nommons ce modèle, le modèle M6. Le meilleur point de coupure est à 0,56 pour lesquelles 89,8% des observations sont prédites correctement. Dans l'échantillon, 10,2% des données réelles sont des valeurs égales à 0 qui représentent les clients non profitables et 89,8% sont égales à 1, les clients profitables. La sensitivité de M6 est de 99% et la spécificité est de 15%.

Modèle M6 :

$$
\begin{aligned}
log\left(\frac{p}{1-p}\right) = {} & 0{,}639 + 1{,}1125\text{interv} + 0{,}6178\text{an}_{\text{cl}} - 0{,}6102\text{iIntAn} \\
& - 0{,}048\text{Achaud} - 0{,}721\text{CEcomm} + 1{,}21\text{CEres} \\
& - 0{,}126\text{iIntAchaud} - 0{,}151\text{iIntCEcomm} - 0{,}038\text{iIntCEres} \\
& + 0{,}371\text{gen1} - 0{,}063\text{gen2} - 0{,}3093\text{iAchaudGen1} \\
& + 0{,}0397\text{iAchaudGen2} + 0{,}7511\text{iCEcommGen2} \\
& - 0{,}583\text{iCEresGen1}
\end{aligned}
$$

4.3.5 Comparaison entre les modèles M5 et M6

En mode d'entraînement, les résultats des modèles M5 et M6 sont relativement similaires lorsque sont comparés le point de coupure, le taux de bonne classification, la spécificité et la sensitivité tel que l'indique le tableau 4-8. La principale différence réside dans le fait que la plupart des variables dans M6 s'avèrent non significatives.

Modèle	Point de coupure	Taux de bonne classification	Spécificité	Sensitivité
M5	0,52	89,8%	11%	99%
M6	0,56	89,8%	15%	99%

Tableau 4-8 : Résumé des résultats pour M5 et M6

4.4 Modèle combiné : Modèle M9

Une autre perspective d'établir la profitabilité clientèle est de faire une régression linéaire multiple distinctement pour les clients profitables M7 et les clients non profitables M8 à l'aide de R^2 sur l'échantillon d'entraînement. Le but consiste à vérifier si les R^2 ajustés de ces deux nouveaux modèles sont supérieurs au R^2 ajusté obtenu lors de la régression linéaire multiple avec les interactions, le modèle M2. Ce dernier a une meilleure erreur de prévision que le modèle M1 tel que constaté dans la section 4.1 (117,46 vs 116,65). Ensuite, nous utilisons l'échantillon de validation pour évaluer la performance des modèles construits. Les résultats prédits par le modèle de la régression logistique M5 sont utilisés pour appliquer les modèles M7 et M8. Même si M5 et M6 donnent le même taux de bonne classification, nous privilégions le modèle M5 car la plupart de ses variables sont significatives contrairement au modèle M6. S'il est prédit par M5 que le client est profitable, alors le modèle M7 est appliqué. S'il est prédit par M5 que le client est non profitable, alors le modèle M8 est appliqué. Le modèle M9 est le résultat des prédictions de M5

sur lequel est appliqué le modèle M7 ou M8 selon la valeur de la prévision. Le modèle M9 est donc un modèle combiné de la régression logistique et de la régression linéaire multiple.

Afin de sélectionner distinctement le meilleur modèle de régression linéaire multiple pour les clients profitables et les non profitables, l'annexe 6 dresse le résultat du nombre de variables dans le modèle selon le critère du R^2 ajusté dans l'échantillon d'entraînement. En d'autres termes, parmi tous les modèles à k variables, il s'agit du modèle ayant le meilleur R^2 ajusté pour les clients profitables et les clients non profitables.

4.4.1 Clientèle profitable : modèle M7

Débutons avec la clientèle profitable. Le modèle qui donne le R^2 ajusté le plus élevé est le modèle à six prédicteurs, il est de 0,1260 tel qu'indiqué dans l'annexe 8. Puisque nous sommes intéressés par les différents équipements, nous modélisons le modèle à six prédicteurs avec l'ensemble des équipements. Ce modèle devient le modèle M7. Le R^2 est de 0,1374 et le R^2 ajusté est de 0,1368. Le R^2 du modèle M7 est légèrement plus élevé que celui du modèle M2 (0,1335). Les prédicteurs du modèle M7 sont : le nombre d'années-client, le changement d'équipement au cours des cinq dernières années, le profil et le type de l'équipement et la génération de l'équipement tel que présenté dans l'annexe 9.

Modèle M7 :

profitabilité moyenne =
660,81 - 103,22 an_cl - 10,85 interv + 16,83 chgt - 43,34 profil + 190,36 CEcomm + 13,23 Echaude - 100,68 foyer - 37,04 radiateur - 7,43 gen1 - 5,56 gen2

4.4.2 Clientèle non profitable: modèle M8

De l'autre côté, procédons de la même façon avec les clients non profitables. Le modèle qui donne le R^2 ajusté le plus élevé est le modèle à cinq prédicteurs, il est de 0,3172 tel présenté dans l'annexe 8. En modélisant ce modèle avec l'ensemble des équipements, nous obtenons le modèle M8. Le R^2 est 0,3094 et le R^2 ajusté est de 0,3050. Le R^2 du modèle M8 est 2,3 fois plus élevé que celui du modèle M2. Les prédicteurs du modèle M8 sont : le nombre d'années-client, le changement d'équipement au cours des cinq dernières années, le type de l'équipement et la génération de l'équipement. Toutes les variables de M8 se retrouvent dans l'annexe 10.

Modèle M8 :

profitabilité moyenne =
- 708,57 + 145,36 an_cl - 42,75 interv - 19,20 chgt + 18,09 CEcomm - 2,79 CEres - 7,33 Echaude + 41,02 foyer - 1,50 radiateur + 6,35 gen1 - 0,18 gen2

4.4.3 Performance du modèle M9

Afin de juger la performance du modèle combiné M9, nous comparons son erreur quadratique moyenne et son erreur de prévision par rapport aux modèles M1 et M2 en mode validation. Les erreurs sont illustrées dans le tableau 4-9. Nous constatons que l'erreur de prévision la plus petite est celle du modèle M2. À l'inverse, l'erreur de prévision la plus élevée est celle du modèle combiné M9. Ainsi, le modèle M9 semble faire plus souvent de grandes erreurs que le modèle M2. Il est donc peu fiable pour prédire les valeurs de profitabilité. Cependant, le R^2 ajusté des modèles intermédiaires M7 et M8 sont plus élevés que ceux de M1 et M2. En séparant les clients profitables des clients non profitables, la mesure d'adéquation entre la variable dépendante et les variables indépendantes est plus forte.

Modèle	Erreur quadratique moyenne	Erreur de prévision
M1	11763,04	108,46
M2	11628,31	107,83
M9	13731,55	117,18

Tableau 4-9 : Résumé des erreurs des modèles

Pour comprendre davantage le modèle M9, quelques analyses supplémentaires s'avèrent nécessaires. Rappelons-nous que ce modèle est construit à l'aide des modèles intermédiaires M5, M7 et M8. Dans le tableau 4-10 ci-dessous, nous retrouvons l'erreur quadratique et l'erreur de prévision des modèles M7 et M8 en comparaison avec le M9. Le tableau révèle que le modèle M7 possède la plus petite erreur quadratique et erreur de prévision (même en comparaison avec M1 et M2). À l'opposée, les erreurs du modèle M8 sont les plus élevées (même en comparaison avec M1 et M2). Il est à noter que les erreurs de M7 et de M8 ne sont pas calculées sur la base de tous les clients comme dans le cas de M1 et M2. Le modèle M7 est construit uniquement sur les clients profitables, alors que le modèle M8 est construit sur les clients non profitables.

Modèle	Erreur quadratique moyenne	Erreur de prévision
M7	11541,97	107,43
M8	36213,23	190,30
M9	13731,55	117,18

Tableau 4-10 : Résumé des erreurs des modèles

Cette fois-ci, en abordant le modèle M9 sous l'angle de la régression logistique pour la prédiction des clients profitables et non profitables, plusieurs points méritent d'être étudiés davantage. Selon les valeurs réelles, 90,5% des clients sont profitables et 9,5% ne le sont pas. Le modèle de la régression logistique M5 prédit que 91,1% des clients sont profitables et 8,9% ne le sont pas. Les résultats sont similaires entre la réalité et la prédiction et ne sont pas alarmants. De plus, le taux de la bonne classification est établi à 88,2% pour M9. Ce dernier est illustré dans le tableau 4-11. Le taux de spécificité est de 36,6% et le taux de sensitivité est de 93,2%.

	y=0	y=1
$\hat{y}$=0	219	379
$\hat{y}$=1	419	5721

Tableau 4-11 : Taux de bonne classification de M9

La problématique dans le modèle M9 est son incapacité à détecter les clients qui sont non profitables. Le taux de spécificité est de 34,3%, donc à peine le tiers des clients non profitables sont classifiés correctement. Dans ce cas, le modèle M8 (régression linéaire sur les clients non profitables) est appliqué sur 65,7% des clients qui ne sont pas supposés être des clients non profitables. Cette situation fait inévitablement en sorte que l'erreur de prévision soit encore plus grande que si le client avait été prédit profitable. C'est la raison pour laquelle les erreurs de M8 sont les plus élevées. Donc, même si les modèles intermédiaires M7 et M8 ont de meilleurs R^2 ajustés, en les appliquant sur la mauvaise classification de clientèle selon les prédictions de M5, le résultat

global fait en sorte que la performance du modèle M9 se trouve affectée.

4.5 Modèle recommandé à ABC : Modèle M2

Dans les sections précédentes, nous avons étudié trois modèles de régression : linéaire multiple, gamma et logistique. À la lumière de nos analyses, nous recommandons un seul modèle à ABC pour évaluer la profitabilité de la clientèle. Celle que nous choisissons est la régression linéaire multiple avec les interactions, soit le modèle M2. Dans la littérature, il a été démontré que la régression linéaire multiple et gamma arrivent sensiblement aux mêmes résultats (Das, 2011). Pour ce qui est de notre étude, l'erreur de prévision de la régression gamma est légèrement inférieure à celle de la régression linéaire multiple. Cette différence est non significative du point de vue pratique. Puisque l'interprétation d'un modèle gamma se fait à l'aide de pourcentage, nous privilégions la régression linéaire multiple, car elle est plus facile à interpréter. De plus, la raison pour laquelle nous ne retenons pas la régression logistique est qu'elle permet uniquement de prédire si le client est profitable ou ne l'est pas, ce qui n'est pas une information suffisante pour prendre des décisions éclairées.

4.6 Recommandations managériales

4.6.1 Variables les plus discriminantes

Suite aux analyses du modèle M2, il s'avère que les variables qui influencent le plus fortement la profitabilité sont le nombre d'interventions, le profil et le type de l'équipement.

Dans un premier temps, le nombre d'interventions joue un rôle critique dans la profitabilité. Dans les analyses préliminaires, il a été démontré que plus le nombre d'interventions augmente, plus le profit diminue. Une intervention génère une perte moyenne de 62,04$. Le seuil critique de la profitabilité est de deux interventions. Donc, à partir de deux interventions, le client engendre déjà une perte pour l'entreprise. Actuellement, ABC ne facture aucune surprime pour chaque intervention effectuée chez le client. Nous recommandons qu'à partir de la deuxième intervention, ABC facture au minimum 95$/h qui est le coût de la main-d'œuvre et des accessoires reliés à l'intervention. Elle pourrait également le présenter sous forme de franchise. Cette nouvelle disposition devrait être incluse dans les nouveaux contrats des programmes de protection. Par contre, cette surprime ou cette franchise seraient moins attrayantes pour les clients. ABC pourrait également augmenter le prix de ses produits, ce qui permettrait d'augmenter le nombre d'interventions possibles tout en demeurant profitable. Il faudrait alors évaluer l'impact sur la demande.

Dans un deuxième temps, les équipements dont le profil est le chauffage sont beaucoup plus profitables que ceux du profil chauffe-eau. Reprenons l'exemple de notre client avec une ancienneté de cinq ans et qui n'a pas changé son équipement. En présence du profil chauffe-eau avec une intervention, le profit est de 10,09$. Dès la deuxième intervention, l'entreprise subit une perte de 57,23$ et s'il y a une troisième intervention, la perte est estimée à 124,55$. Lorsque le profil est le chauffage, dans le cas d'une première intervention, le profit est de 66,29$. À la deuxième intervention, nous sommes en présence du seuil critique, la perte est de 1,03$. À la troisième intervention, la perte est de 68,35$. Il y a donc une plus grande marge de manœuvre avec le profil chauffage puisque le prix actuel permet d'inclure deux interventions avant que le produit ne soit plus profitable. Afin de ramener la même profitabilité, il faudrait penser à augmenter le prix pour les équipements de profil chauffe-eau d'au moins 60$.

Finalement, lorsque nous arrivons au niveau de l'équipement, celui qui génère le plus de profit est le chauffe-eau commercial avec une moyenne de 90,12$. Celui qui engendre la plus grosse perte est le foyer. La perte moyenne est estimée à 96,00$. ABC pourrait retirer la couverture du foyer de sa gamme de produits ou d'ajuster le prix de la couverture du foyer afin de refléter un niveau de profitabilité respectable. Au niveau global, il est souhaitable que tous les clients soient profitables. Cependant peut-être qu'une petite perte peut être compensée par le fait que le client est loyal et assure des revenus mensuels stables dans le temps.

En somme, nous recommandons trois actions précises à ABC. La première est d'augmenter le prix de ses produits. La deuxième est de facturer les interventions ou d'introduire un concept de franchise après un certain nombre d'interventions pour assurer la profitabilité. La dernière est d'exclure certains équipements de son programme de protection.

4.6.2 Erreurs de prévision

Le modèle M2 que nous avons retenu permet d'expliquer à 13,45% la « profitabilité moyenne ». Lorsque nous appliquons ce modèle sur l'ensemble de nos données, l'estimation de la profitabilité du client dans 7% des cas est prédite dans un intervalle de 5% de la profitabilité réelle. Dans 13% des cas, l'erreur de prévision se situe entre 5% et 10%. Il est évident que si l'entreprise est prête à tolérer un plus gros pourcentage d'erreur de prévision tel que 40%, dans ce cas, ce sont près de la moitié des clients (47%) qui ont une profitabilité estimée jusqu'à 40% plus ou moins élevée que la profitabilité réelle. Le détail des intervalles d'erreur de prévision est représenté dans le tableau 4-12.

% Erreur de prévision	Fréquence	Pourcentage	Pourcentage cumulé
< 5%	1498	6.69	6.69
5%-10%	1456	6.50	13.20
11%-20%	2733	12.21	25.40
21%-30%	2733	11.73	37.13
31%-40%	2220	9.92	47.05
41%-50%	2007	8.97	56.01
51% +	9848	43.99	100.00

Tableau 4-12 : Pourcentage des intervalles des erreurs de prévision

Au-delà des erreurs de prévision, l'erreur la plus grave est de prédire qu'un client est profitable alors qu'il ne l'est pas. Selon notre modèle, moins de 10% des clients prédits comme étant profitables ne le sont pas en réalité. Le tableau 4-13 présente la profitabilité vs la profitabilité prédite par M2.

Clients	Non profitables	Profitables
Prédits non profitables	163	111
Prédits profitables	2 042	20 071

Tableau 4-13 : Nombre de clients profitables vs prédits profitables

4.7 Limites du modèle retenu

Nous tenons à préciser que le modèle M2 recommandé à l'entreprise ABC pour la prédiction de la profitabilité du client est en fonction des données mises à notre disposition. Ces données ne sont pas précises à 100%. Tel que discuté dans le chapitre 3, plusieurs règles de standardisation sont établies afin que nous puissions travailler les données et répondre aux objectifs de l'étude. Le manque de précision des données a certainement un impact sur les résultats.

Revenons sur quelques unes des règles de standardisations que nous avons établies au chapitre 3. Le problème majeur de la base de données de ABC est que les données sont agrégées pour toutes les variables de revenus, de coûts et de nombre d'interventions. De plus, nous ne disposons pas de liste complète des équipements couverts par le produit. La base de données n'en présente que deux. Tel que nous l'expliquons dans le chapitre 3, le client peut posséder un équipement secondaire dont ne mentionne pas la base de données. Il est alors impossible de déterminer la profitabilité spécifique pour une période et de déterminer avec certitude que l'intervention est attribuée au bon équipement. En d'autres termes, une meilleure précision des données aurait amélioré le résultat de nos différents modèles de régression.

Afin d'améliorer la qualité des données, ABC devrait détailler le niveau de l'information par produit et par intervention. Les informations suivantes devraient être disponibles :

- le type de produit (couverture de 1 an, 2 ans ou 3 ans, le nombre d'interventions inclus dans le produit : 0, 1, 2 ou 3),
- la date de début et de fin de la couverture du produit,
- le revenu,
- le coût,
- les équipements principaux et secondaires ainsi que les options,
- le type d'intervention sur lequel des équipements,
- la date de l'intervention,
- le coût de l'intervention que celle-ci soit incluse ou non dans le produit.

Chapitre 5

Conclusion

L'objectif de la présente étude était d'élaborer un modèle de profitabilité d'un client de ABC à partir de la base de données fournie par l'entreprise. Ce mémoire a d'abord établi une revue de la littérature sur le concept de la profitabilité client. Une confusion règne autour de la terminologie de ce concept. Pour l'étude, nous avons adopté la profitabilité client définie par Berger et Nars (1998) sous l'appellation de capital client.

Dans le but de dresser le profil des clients actuels, nous avons procédé à la préparation et l'exploration des données. Ensuite, pour expliquer la profitabilité client, nous avons fait appel à trois modèles classiques de régression : linéaire multiple, gamma et logistique. Nous les avons également évalués avec et sans les interactions. Pour notre quatrième modèle, nous avons combiné la régression logistique et la régression linéaire multiple. Le constat est que les modèles avec interactions performent mieux que les modèles sans interactions. L'erreur de prévision de la régression gamma est légèrement plus petite que celle de la régression linéaire multiple, mais cette différence est

négligeable du point de vue pratique. La régression logistique performe bien, mais permet uniquement de prédire si le client est profitable ou ne l'est pas, ce qui n'est pas une information suffisante pour prendre une décision éclairée. Quant au modèle combiné, son erreur de prévision est plus grande par rapport à la régression linéaire multiple. Pour toutes ces raisons, nous recommandons le modèle de régression linéaire multiple (M2) à ABC. De plus, ce modèle est plus facile à interpréter pour les gestionnaires.

Finalement, nous recommandons trois actions précises à ABC. La première est d'augmenter le prix de ses produits. La deuxième est de facturer les interventions ou d'introduire un concept de franchise après un certain nombre d'interventions pour assurer la profitabilité. La dernière est d'exclure certains équipements de son programme de protection.

De futures recherches pourraient être entreprises pour comprendre davantage la clientèle de ABC. Ce que nous avons étudié est la profitabilité des clients actuels. En analysant la profitabilité des clients ayant mis fin à leur contrat, nous pourrions déterminer si l'entreprise a perdu des occasions d'affaires intéressantes. Cette analyse pourrait être reliée à une étude portant sur la fidélisation des clients. L'entreprise pourrait tester différents concepts pour récompenser les clients les plus loyaux pour assurer leur maintien à long terme.

Annexes

Annexe 1

Modèle M1 - Modèle de la régression linéaire multiple selon le critère du R^2 ajusté

Root MSE	117.49600	**R-Square**	0.1224
Dependent Mean	106.57752	**Adj R-Sq**	0.1219
Coeff Var	110.24463		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	459.52464	15.23482	30.16	<.0001
nb_an_cl	nb années-client	1	-59.46660	3.06936	-19.37	<.0001
interv	intervention	1	-47.25624	1.61576	-29.25	<.0001
chgt	changement équip	1	12.10304	2.45398	4.93	<.0001
profil	profil équip	1	-55.87755	2.81423	-19.86	<.0001
equip22	chauffe-eau com	1	123.13594	18.01982	6.83	<.0001
equip25	eau chaude	1	16.36131	2.35215	6.96	<.0001
equip26	foyer	1	-84.84902	17.31280	-4.90	<.0001
equip27	radiateur	1	-35.07107	6.16438	-5.69	<.0001

Annexe 2

Modèle M2 - Modèle de la régression linéaire multiple avec l'ajout de termes d'interaction

Root MSE	116.71932	**R-Square**	0.1345
Dependent Mean	106.57752	**Adj R-Sq**	0.1335
Coeff Var	109.51589		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	380.77348	26.74999	14.23	<.0001
nb_an_cl	nb années-client	1	-62.03685	3.05981	-20.27	<.0001
interv	intervention	1	-67.32261	3.02464	-22.26	<.0001
chgt	changement équip	1	127.40097	22.57721	5.64	<.0001
profil	profil équip	1	56.20061	22.60829	2.49	0.0129
equip22	chauffe-eau com	1	90.11833	55.96478	1.61	0.1074
equip25	eau chaude	1	-60.09569	10.96181	-5.48	<.0001
equip26	foyer	1	-96.00183	17.37662	-5.52	<.0001
equip27	radiateur	1	-49.94808	6.68326	-7.47	<.0001
gen21	génération 1	1	-26.14468	4.09549	-6.38	<.0001
gen22	génération 2	1	-11.04104	3.95238	-2.79	0.0052
iIntGen1	int interv et gen1	1	20.15804	3.84817	5.24	<.0001
iInt_CEcomm	int interv et CE comm	1	-64.26419	24.69131	-2.60	0.0093
iInt_Echaude	int interv et eau chaude	1	14.20451	3.52932	4.02	<.0001
iChgtProfil	int chgt et profil	1	-110.01392	22.82448	-4.82	<.0001
iChgt_CEcomm	int chgt et CE comm	1	152.23691	53.28996	2.86	0.0043
iChgt_Echaude	int chgt et eau chaude	1	-12.79265	5.26052	-2.43	0.0150
iGen1_Echaude	int gén1 et eau chaude	1	74.30167	10.27039	7.23	<.0001
iGen2_Echaude	int gén2 et eau chaude	1	60.75987	12.92150	4.70	<.0001

Annexe 3

Modèle M3 - Modèle de régression gamma selon la méthode séquentielle descendante

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	7.7301	0.0086	7.7134	7.7469	817042	<.0001
nb_an_cl	1	-0.0279	0.0017	-0.0313	-0.0246	270.29	<.0001
interv	1	-0.0249	0.0009	-0.0267	-0.0231	749.86	<.0001
chgt	1	0.0063	0.0015	0.0034	0.0092	18.29	<.0001
profil	1	-0.0285	0.0018	-0.0321	-0.0250	247.58	<.0001
equip22	1	0.0622	0.0101	0.0424	0.0821	37.69	<.0001
equip25	1	0.0088	0.0015	0.0059	0.0117	35.09	<.0001
equip26	1	-0.0448	0.0098	-0.0640	-0.0256	20.91	<.0001
equip27	1	-0.0203	0.0037	-0.0275	-0.0131	30.22	<.0001
gen21	1	-0.0021	0.0016	-0.0053	0.0011	1.59	0.2072
gen22	1	-0.0030	0.0021	-0.0071	0.0011	2.04	0.1534
Scale	1	229.4261	2.5918	224.4021	234.5626		

Annexe 4

Modèle M4 - Modèle de la régression gamma avec l'ajout de termes d'interaction

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	7.6795	0.0142	7.6517	7.7072	294462	<.0001
nb_an_cl	1	-0.0291	0.0017	-0.0324	-0.0258	296.16	<.0001
interv	1	-0.0354	0.0017	-0.0387	-0.0320	431.07	<.0001
chgt	1	0.0730	0.0116	0.0502	0.0958	39.41	<.0001
profil	1	0.0348	0.0117	0.0120	0.0576	8.91	0.0028
equip22	1	0.0360	0.0272	-0.0173	0.0893	1.75	0.1858
equip25	1	-0.0298	0.0062	-0.0419	-0.0178	23.43	<.0001
equip26	1	-0.0470	0.0098	-0.0662	-0.0279	23.13	<.0001
equip27	1	-0.0247	0.0038	-0.0321	-0.0172	42.59	<.0001
gen21	1	-0.0126	0.0023	-0.0172	-0.0081	29.32	<.0001
gen22	1	0.0008	0.0031	-0.0053	0.0070	0.07	0.7883
iIntGen1	1	0.0107	0.0022	0.0064	0.0150	24.19	<.0001
iChgtProfil	1	-0.0602	0.0118	-0.0833	-0.0370	25.95	<.0001
iInt_CEcomm	1	-0.0289	0.0143	-0.0569	-0.0008	4.07	0.0437
iInt_Echaude	1	0.0070	0.0020	0.0031	0.0109	12.29	0.0005
iChgtGen2	1	-0.0103	0.0037	-0.0176	-0.0029	7.56	0.0060
iChgt_Echaude	1	-0.0097	0.0032	-0.0159	-0.0034	9.21	0.0024
iGen1_CEcomm	1	0.0892	0.0241	0.0419	0.1364	13.67	0.0002
iGen1_Echaude	1	0.0385	0.0058	0.0272	0.0498	44.65	<.0001
iGen2_Echaude	1	0.0330	0.0073	0.0188	0.0473	20.64	<.0001
Scale	1	232.2973	2.6242	227.2104	237.4981		

Annexe 5

Résumé des modèles selon les différents critères de sélection pour la régression logistique

K	Score	AIC	BIC	Variables dans le modèle
1	8618	8622	8637	• Nombre moyen d'interventions
2	7892	7898	7921	• Nombre moyen d'interventions • Profil d'usage
3	7884	7892	7923	• Nombre moyen d'interventions • Profil d'usage • Type d'équipement
4	7812	7823	7860	• Nombre moyen d'interventions • Profil d'usage • Type d'équipement • Nombre années-client
5	7792	7804	7850	• Nombre moyen d'interventions • Profil d'usage • Type d'équipement • Nombre années-client • Génération de l'équipement
6	7786	7806	7860	• Nombre moyen d'interventions • Profil d'usage • Type d'équipement • Nombre années-client • Génération de l'équipement • Changement d'équipement

Annexe 6

Modèle M5 – Modèle de la régression logistique selon le critère AIC et BIC

Analysis of Maximum Likelihood Estimates							
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**	**Exp(Est)**
Intercept		1	1.3212	0.4477	8.7071	0.0032	3.748
interv		1	-1.7424	0.0502	1206.6624	<.0001	0.175
nb_an_cl		1	0.4716	0.0896	27.6818	<.0001	1.603
equip	**1**	1	-0.2698	0.1131	5.6899	0.0171	0.764
equip	**2**	1	-0.4452	0.2991	2.2147	0.1367	0.641
equip	**3**	1	0.6324	0.1178	28.8256	<.0001	1.882
gen	**1**	1	0.1953	0.0553	12.4567	0.0004	1.216
gen	**2**	1	-0.0582	0.0593	0.9652	0.3259	0.943

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
interv	0.175	0.159	0.193
nb_an_cl	1.603	1.344	1.910
profil 0 vs 1	3.543	2.232	5.621
equip 1 vs 5	0.703	0.599	0.826
equip 2 vs 5	0.167	0.050	0.551
gen 1 vs 3	1.394	1.138	1.707
gen 2 vs 3	1.082	0.873	1.342

Annexe 7

Modèle M6 - Modèle de la régression logistique avec l'ajout des termes d'interaction

Analysis of Maximum Likelihood Estimates							
Parameter			**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept			1	0.6393	0.5503	1.3493	0.2454
interv			1	1.1125	0.7497	2.2018	0.1378
nb_an_cl			1	0.6178	0.0871	50.2696	<.0001
interv*nb_an_cl			1	-0.6102	0.1470	17.2338	<.0001
equip	**1**		1	-0.0484	0.3717	0.0170	0.8964
equip	**2**		1	-0.7206	1.0288	0.4906	0.4836
equip	**3**		1	1.2101	0.5040	5.7642	0.0164
interv*equip	**1**		1	-0.1259	0.2417	0.2712	0.6025
interv*equip	**2**		1	-0.1509	0.6905	0.0478	0.8270
interv*equip	**3**		1	-0.0384	0.2492	0.0238	0.8774
gen	**1**		1	0.3715	0.2496	2.2148	0.1367
gen	**2**		1	-0.0628	0.0826	0.5791	0.4467
equip*gen	**1**	**1**	1	-0.3093	0.2559	1.4619	0.2266
equip*gen	**1**	**2**	1	0.0397	0.0825	0.2318	0.6302
equip*gen	**2**	**1**	1	0.7511	0.6668	1.2689	0.2600
equip*gen	**2**	**2**	0	0	.	.	.
equip*gen	**3**	**1**	1	-0.5826	0.4101	2.0175	0.1555
equip*gen	**3**	**2**	0	0	.	.	.

Annexe 8

Meilleurs modèles à k variables selon le critère R^2 ajusté pour les clients profitables et non profitables

Clientèle profitable				**Clientèle non profitable**		
k	R^2 ajusté	Variables dans le modèle		k	R^2 ajusté	Variables dans le modèle
1	0.686	• Nombre années-client		1	0.2243	• Nombre années-client
2	0.1189	• Nombre années-client • Profil d'usage		2	0.3103	• Nombre années-client • Nombre moyen d'interventions
3	0.1222	• Nombre années-client • Profil d'usage • Changement d'équipement		3	0.3168	• Nombre années-client • Nombre moyen d'interventions • Changement d'équipement
4	0.1235	• Nombre années-client • Profil d'usage • Changement d'équipement • Type d'équipement		4	0.3172	• Nombre années-client • Nombre moyen d'interventions • Changement d'équipement • Génération de l'équipement
5	0.1254	• Nombre années-client • Profil d'usage • Changement d'équipement • Type d'équipement • Nombre moyen d'interventions		5	0.3176	• Nombre années-client • Nombre moyen d'interventions • Changement d'équipement • Génération de l'équipement • Type d'équipement
6	0.1260	• Nombre années-client • Profil d'usage • Changement d'équipement • Type d'équipement • Nombre moyen d'interventions • Génération de l'équipement		6	0.3172	• Nombre années-client • Nombre moyen d'interventions • Changement d'équipement • Génération de l'équipement • Type d'équipement • Profil d'usage

Annexe 9

Modèle M7 – Modèle de la régression linéaire multiple sur la clientèle profitable

Root MSE	104.03640	**R-Square**	0.1374
Dependent Mean	125.70172	**Adj R-Sq**	0.1368
Coeff Var	82.76450		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	660.81310	14.81597	44.60	<.0001
nb_an_cl	nb années-client	1	-103.22162	2.94377	-35.06	<.0001
interv	intervention	1	-10.85062	1.68079	-6.46	<.0001
chgt	changement équip	1	16.83440	2.46885	6.82	<.0001
profil	profil équip	1	-43.34129	3.02050	-14.35	<.0001
equip22	chauffe-eau com	1	190.35699	18.20419	10.46	<.0001
equip25	eau chaude	1	13.22790	2.48698	5.32	<.0001
equip26	foyer	1	-100.68427	17.86963	-5.63	<.0001
equip27	radiateur	1	-37.04330	6.13720	-6.04	<.0001
gen21	génération 1	1	-7.42925	2.70692	-2.74	0.0061
gen22	génération 2	1	-5.55967	3.55172	-1.57	0.1175

Annexe 10

Modèle M8 – Modèle de la régression linéaire multiple sur la clientèle non profitable

Root MSE	89.42173	**R-Square**	0.3094
Dependent Mean	-65.28397	**Adj R-Sq**	0.3050
Coeff Var	-136.97349		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	-708.57002	30.34722	-23.35	<.0001
nb_an_cl	nb années-client	1	145.36291	6.17493	23.54	<.0001
interv	intervention	1	-42.75110	3.27014	-13.07	<.0001
chgt	changement équip	1	-19.20225	5.80821	-3.31	0.0010
equip22	chauffe-eau com	1	18.09391	28.67495	0.63	0.5281
equip23	chauffe-eau res	1	2.79279	8.56794	0.33	0.7445
equip25	eau chaude	1	-7.33369	5.96657	-1.23	0.2192
equip26	foyer	1	41.01576	26.77936	1.53	0.1258
equip27	radiateur	1	-1.49654	15.92418	-0.09	0.9251
gen21	génération 1	1	6.34791	7.25270	0.88	0.3816
gen22	génération 2	1	-0.17585	7.95598	-0.02	0.9824

Bibliographie

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Dans B. N. Petrov, et F. Csaki, (Eds.) *Second International Symposium on Information Theory*. Akademiai Kiado, Budapest, 267-281.

Anderson, E. W., Fornell, C., et Lehmann, D. R. (1994). Customer Satisfaction, Market Share, and Profitability: Findings from Sweden. Journal of Marketing, vol. 58, 53-66.

Baillargeon, G., et Rainville, J. (1979). Statistique appliquée, Tome 3, Régression multiple (2e édition). Les éditions SMG, Trois-Rivières, 1034 p.

Benoit, D. F., et den Poel, D. V. (2009). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. Expert Systems with Applications, vol 36, n^{o} 7, 10475-10484.

Berger, P. D., et Nasr, N. I. (1998). Customer lifetime value: marketing models and applications. Journal of Interactive Marketing, vol 12, n^{o} 1, 17–30.

Blattberg, R. C., et Deighton, J. (1996). Manage Marketing by the Customer Equity Test. Harvard Business Review, vol. 74, n^{o} 4, 136-144.

Burnham, K. P., et Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in Model Selection. Sociological Methods and Research, vol 33, 261–304.

Das, R. N. (2011). Slope Rotatability With Correlated Errors Calcutta Statistical Association Bulletin, vol. 54, 57-71.

Das, R. N., et Lee, Y. (2009). Log-normal versus gamma models for analyzing data from quality-improvement experiments. Quality Engineering, vol 21, n° 1, 79-87.

Donkers, B., Verhoef, P., et de Jong, M. (2007). Modeling CLV: A test of competing models in the insurance industry. Quantitative Marketing and Economics, vol 5, n° 2, 163–190.

Duboff, R. S. (1992). Marketing to maximize profitability. The Journal of Business Strategy, vol 13, n° 6, 10–13.

Firth, D. (1988). Multiplicative errors: log-normal or gamma? Journal of the Royal Statistical Society: Series B, vol 50, n° 2, 266-268.

Glady, N., Baesens, B., et Croux, C. (2009). A Modified Pareto/NBD Approach for Predicting Customer Lifetime Value. Expert Systems With Applications, vol 36, n°2, 2062-2071.

Gleaves, R., Burton, J., Kitshoff, J., Bates, K., et Whittington, M. (2008). Accounting is from Mars, marketing is from Venus: Establishing common ground for the concept of customer profitability. Journal of Marketing Management, vol 24, 825–845.

Gloy, B. A., Akridge, J. T., et Preckel, P. V. (1997). Customer lifetime value: An application in the rural petroleum market. Agribusiness, vol 13, n° 3, 335–347.

Griffin, J. (2003). Customer segmentation: divide and prosper, iQ Magazine, Cisco Systems, n° March/April.

Gupta S., et Lehmann D. R. (2005). Managing Customers as Investments: The Strategic Value of Customers in the Long Run, Upper Saddle River, NJ: Wharton School Publishing, Pearson Education, Inc., 224 p.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L., et Black, W. C. (1995). Multivariate Data Analysis (3rd ed). New York: Macmillan.

Hastie T., Tibshirani R., et Friedman J. (2009). The Elements of Statistical Learning (2nd edition), Springer-Verlag, 763 p.

Huber, E., et Stephens, J. D. (1993). Political Parties and Public Pensions: A Quantitative Analysis, *Acta Sociologica*, vol 36, 309-325.

Hwang, H., Jung, T., et Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. Expert Systems with Applications, vol 26, n^{o} 2, 181–188.

Jain, D., et Singh, S. S. (2002). Customer Lifetime Value Research in Marketing: A Review and Future Directions. Journal of Interactive Marketing, vol. 16, n^{o} 2, 34-46.

Kim, S. Y., Jung, T. S., Suh, E. H., et Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert Systems with Applications, vol 31, n^{o} 1, 101–107.

Kennedy, P. (1992). A Guide to Econometrics (3rd ed.). Oxford: Blackwell. 424 p.

Koenker, R. (2005). Quantile regression. Cambridge: Cambridge University Press.

Koenker, R., et Basset, G. (1978). Regression Quantiles. Econometrica, vol 46, n^{o} 1, 33–50.

Kohonen, T. (1989). Self-organization & associative memory (3rd ed.). Springer-Verlag, 312 p.

Kumar, V., Petersen, J. A., et Leone, R. P. (2010). Driving profitability by encouraging customer referrals: Who, when and how. Journal of Marketing, vol 74, 1–17.

Larocque, Denis (2006). Analyse Multidimensionnelle, Recueil-6-602-07 A07, note de cours, Montréal, École des hautes études commerciales, 408 p.

Lee, J. H., et Park, S. C. (2005). Intelligent profitable customers segmentation system based on business intelligence tools. Expert Systems with Applications, vol 29, n° 1, 145–152.

McManus, L. (2007). The construction of a segmental customer profitability analysis. Journal of Applied Management Accounting Research, vol 5, 59–74.

McManus, L., et Guilding, C. (2008). Exploring the potential of customer accounting: A synthesis of the accounting and marketing literatures. Journal of Marketing Management, vol 24, 771–795.

Menard, S. (1995). Applied Logistic Regression Analysis: Sage University Series on Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage.

Mulhern, F. J. (1999). Customer profitability analysis: Measurement, concentration, and research directions. Journal of Interactive Marketing, vol 13, n° 1, 25–40.

Pan, Y, et Jackson, R. T. (2008). Ethnic difference in the relationship between acute inflammation and and serum ferritin in US adult males. Epidemiology and Infection, 136, 421-431.

Pfeifer, P. E., Haskins, M. E., et Conroy, R. M. (2005). Customer lifetime value, customer profitability, and the treatment of acquisition spending. Journal of Managerial Issues, vol 17, 11– 25.

Pohar, M., Blas, M., et Turk S. (2004). Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. Metodološki zvezki, vol. 1, n° 1, 143-161.

Quinlan, R. J. (1993). C4.5: Programs for machine learning. San Mateo, California: Morgan Kaufmann Publishers.

Reinartz, W. J., et Kumar, V. (2000). On the profitability of long-life customers in a non-contractual setting: An empirical investigation and implications for marketing. Journal of Marketing, vol 64, n° 4, 17–35.

Rogerson, P. A. (2011). Statistical methods for geography (3rd ed.). London: Sage, 368 p.

Schwarz, G. E. (1978). Estimating the dimension of a model. Annals of Statistics, vol 6, n° 2, 461–464.

Schmittlein, D. C., Morrison, D. G., et Colombo, R. (1987). Counting your customers: who are they and what will they do next? Management Science, vol 33, n° 1, 1-24.

Shibata, R. (1981). An optimal selection of regression variables. Biometrika, vol 68, 45-54.

Tabachnick, B. G., et Fidell, L. S. (2001). Using Multivariate Statistics (4th ed.). Boston, MA: Allyn and Bacon, 966 p.

Verhoef, P. C., et Donkers, B. (2001). Predicting customer potential value an application in the insurance industry. Decision Support Systems, vol 32, 189–199.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? Biometrika, vol 92, 37–950.