Projet de fin d'études

Regroupement des succursales d'une institution financière dont les caractéristiques ayant trait à l'épargne sont comparables et prédiction de l'atteinte d'un objectif d'épargne annuel

Laurence Desbois-Bédard
Jonathan Roy
Sarah-Anne Savard
Samuel Tremblay

Université Laval
22 avril 2015

Résumé

Pour une institution financière, le suivi de la performance de ses succursales est un défi considérable. Ce suivi est toutefois nécessaire afin d'intervenir rapidement auprès des succursales en difficulté. Dans ce contexte, la grande quantité de données disponibles est à la fois un avantage et un inconvénient. L'objectif de ce projet est de fournir à l'équipe d'optimisation de la performance des succursales d'une institution financière des outils permettant de faciliter l'identification des succursales en difficulté et des stratégies à utiliser pour les aider, et ce, en utilisant les données disponibles.

Le projet est divisé en trois objectifs. Premièrement, des indicateurs sont créés à partir de certaines variables en utilisant l'analyse factorielle. L'information contenue dans 23 variables est ainsi condensée dans 9 indicateurs interprétables. Le second objectif vise à regrouper les succursales de l'institution selon leur profil d'épargne. La classification non supervisée avec la méthode de Ward effectuée sur certaines variables d'épargne permet de former 8 groupes. Le plus grand groupe contient 42% des succursales et le plus petit contient moins de 2% des succursales. L'analyse discriminante non paramétrique basée sur les *k* plus proches voisins permet de classer une nouvelle succursale dans les groupes formés. Pour sa part, le troisième objectif vise la prédiction de l'atteinte d'un objectif d'épargne. Pour ce faire, des modèles de régression logistique et de régression linéaire sont ajustés. Ces modèles sont aussi utilisés pour déterminer les variables qui expliquent le mieux l'atteinte de l'objectif. Les variables significatives dans les modèles sont notamment l'atteinte de l'objectif en juin, le montant cumulatif des ventes en juin, le sous-secteur auquel appartient la succursale et la valeur des propriétés financées. Ces modèles sont ajustés sur les données de 2014 et validés avec les données de 2013. Les taux de mauvaise prévision varient entre 21 % et 32 %. Il ne semble pas y avoir de lien entre le groupe d'appartenance d'une succursale et l'atteinte de son objectif.

Les résultats présentés dans ce rapport doivent être utilisés avec prudence. En effet, les groupes créés ne sont pas stables dans le temps et quelques modèles ont des coefficients contre-intuitifs pour certaines variables.

Table des matières

1 Description du projet

1.1 L'institution financière et la notion d'épargne

L'entreprise qui chapeaute ce projet de fin d'études est une institution financière canadienne qui, pour des raisons de confidentialité, ne peut être identifiée. Pour cette raison, des termes génériques seront utilisés pour la décrire. Elle sera ci-après appelée l'*institution*.

Du point de vue structurel, l'institution regroupe un assez grand nombre de divisions. L'une d'entre elles, la division *succursales*, s'occupe de la gestion quelques centaines de succursales au Canada. La figure 1 présente grossièrement la structure de cette division : elle gère plusieurs succursales ayant chacune un grand nombre de clients

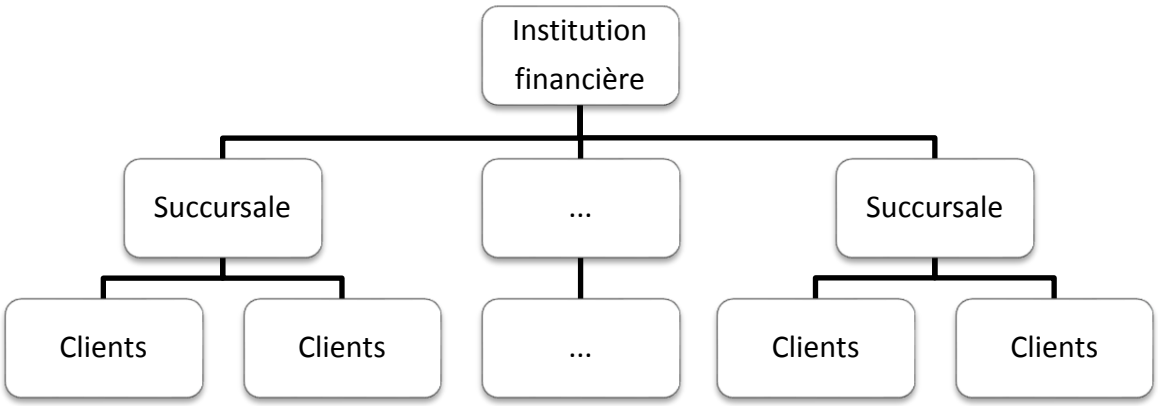

Figure 1 : Structure de la division *succursales* de l'institution financière

Au sein de cette division, plusieurs sphères de l'activité financière sont touchées. En effet, les succursales offrent à leurs clients des prêts hypothécaires, des produits de crédit, d'épargne et de placement, de la gestion de patrimoine, etc. La division *succursales* doit veiller au bon fonctionnement et à la rentabilité de ces transactions. Cela dit, ce projet ne concerne que les produits d'épargne, alors seul cet aspect du domaine financier est ici présenté.

Une bonne compréhension du concept d'épargne est importante afin de saisir les enjeux de ce projet. Il est aisé de donner une définition de l'épargne au niveau de l'individu. Selon Le Petit Robert, l'épargne est la « part du revenu qui n'est pas consacrée à la consommation ». Cette définition est simple et intelligible. Cependant, pour décrire l'épargne d'une succursale, elle doit être adaptée. L'épargne d'une succursale est plutôt définie comme le total de la valeur investie par les clients dans les différents produits d'épargne à un moment précis. Les produits d'épargne sont nombreux, variés et incluent notamment les certificats de placement garantis et les certificats de placement garantis liés aux marchés, lesquels sont des placements sécuritaires à taux d'intérêt fixe ou flexible[1]. Ces types de placements sont

[1] http : //www.rbcbanqueroyale.com

offerts en diverses déclinaisons : les régimes enregistrés d'épargne-retraite (RÉER), les régimes enregistrés d'épargne-étude (RÉEE), les comptes d'épargne libres d'impôts (CÉLI), etc.

Pour traiter de l'épargne d'une succursale, l'expression *ventes-retraits* sera utilisée dans ce rapport. Les *ventes-retraits* sont obtenues en additionnant le montant total de la vente de produits d'épargne et en soustrayant le montant total des retraits entre deux temps d'observation. Une vente a lieu, par exemple, lorsqu'un client achète un REER. Lorsqu'il le retire, en tout ou en partie, il s'agit plutôt d'un retrait. Compte tenu de la nature des *ventes-retraits*, il est possible que leur valeur soit négative à certains temps d'observation.

L'épargne est, bien entendu, un élément clé dans la réussite d'une institution financière. C'est pour cette raison que l'institution y consacre de nombreuses ressources.

1.2 L'équipe d'optimisation de la performance et son travail

Dans la division identifiée à la section précédente, une équipe a pour mandat d'optimiser la performance des succursales. Pour ce faire, elle évalue ces dernières chaque semaine par rapport à plusieurs critères et élabore des stratégies afin de les aider à remplir certains objectifs. Son champ d'action est transversal à tous les départements touchant l'épargne.

Ce projet est réalisé en collaboration avec deux conseillères en intelligence d'affaires de cette équipe. L'intelligence d'affaires, domaine d'expertise relativement récent, est une «expression consacrée pour décrire les systèmes qui combinent les technologies de bases de données, d'entreposage de données, d'exploitation de données et de systèmes d'aide à la décision[2]».

Les conseillers en intelligence d'affaires travaillent principalement pour les gestionnaires de l'équipe d'optimisation. Ceux-ci ont besoin d'information et les conseillers sont là pour leur en fournir. Pour répondre à une question précise, ils doivent forer les entrepôts de données afin d'en extraire les données pertinentes, réaliser des analyses statistiques et résumer l'information dans des tableaux de bord que les gestionnaires consultent pour guider leurs décisions.

Le contexte dans lequel s'insère ce projet est directement relié au mandat de l'équipe d'optimisation de la performance des succursales. Chaque année, l'institution fixe, pour chacune de ses succursales, un objectif de *ventes-retraits* devant être rempli le 31 décembre. L'équipe d'optimisation aide les succursales à atteindre cet objectif. Pour ce faire, l'une des stratégies qu'elle utilise est de fixer un objectif de *ventes-retraits* hebdomadaire pour chaque succursale. En suivant le progrès des succursales d'une semaine à l'autre, il est alors plus aisé de savoir à quel moment intervenir.

[2] http : //www.hec.ca

Évidemment, les *ventes-retraits* de la succursale sont directement liées à l'épargne des clients faisant affaire avec celle-ci. Il est clair que les *ventes-retraits* augmentent lorsque la vente de produits d'épargne augmente et que les retraits diminuent. Qui plus est, certains de ces produits sont plus rentables que d'autres. Par exemple, les produits fabriqués par l'institution elle-même, les *produits maison*, rapportent davantage que les produits génériques.

Lorsque l'atteinte des objectifs de *ventes-retraits* semble compromise pour une ou plusieurs succursales, l'institution déploie des stratégies d'action afin de rectifier la situation. Elle peut, par exemple, aider les succursales en instaurant des promotions ou en proposant des activités de développement des affaires.

À l'usage, l'équipe a noté une difficulté majeure concernant l'élaboration de ces stratégies en raison du grand nombre de succursales et de l'hétérogénéité de leurs profils d'épargne. En effet, les différences entre les succursales compliquent grandement le choix de la stratégie à déployer. En raison de contraintes de temps, il est important pour l'équipe d'optimisation de trouver une solution à ce problème afin d'accélérer le processus de redressement des succursales en difficulté. L'idée proposée par les conseillères est que, pour bien évaluer la performance d'une succursale, l'équipe doit la comparer à un groupe de succursales possédant des caractéristiques semblables. À l'heure actuelle, les groupes utilisés pour ce faire ont été créés pour d'autres besoins. Ils ne regroupent donc pas nécessairement des succursales ayant des caractéristiques semblables au niveau de l'épargne.

1.3 Le projet étudiant

Afin d'améliorer la situation, l'objectif principal de ce projet est de créer un outil de classification permettant de regrouper les succursales dont les caractéristiques en lien avec l'épargne sont similaires. La figure 2 résume le contexte dans lequel s'insère cet outil. Il sera utile au moment de l'identification du groupe des succursales en difficulté.

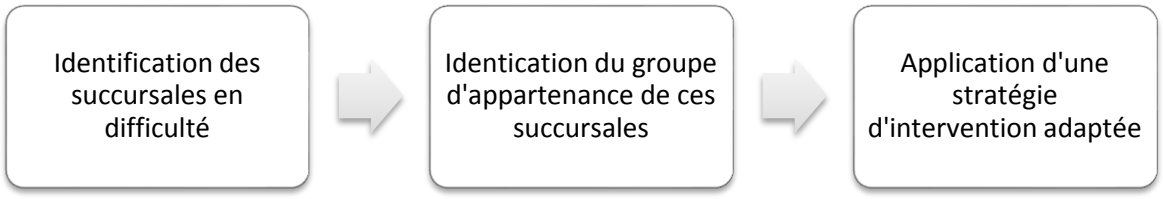

Figure 2 : Schéma résumant le processus d'intervention auprès des succursales en difficulté

Comme mentionné précédemment, l'équipe d'optimisation de la performance utilise déjà des groupes de succursales. Toutefois, ils ne sont pas adaptés pour analyser spécifiquement les données concernant l'épargne. L'objectif principal de ce projet est donc de créer un nouvel outil adapté à l'épargne des succursales. Afin de structurer le travail, le projet est divisé en trois objectifs spécifiques.

1.3.1 Objectif 1 : Création d'indicateurs

D'abord, la dimension du jeu de données est réduite en créant quelques indicateurs à partir de quelques-unes des centaines de variables continues. Ces indicateurs permettent de synthétiser l'information contenue dans les diverses variables et sont plus faciles à manipuler que le jeu de données complet. De plus, ils permettent de réduire la multicolinéarité entre les variables. Certains d'entre eux sont utilisés dans la suite du projet.

Du point de vue pratique, les indicateurs créés peuvent aussi être utilisés par les gestionnaires au moment de la prise de décision. Ils permettent en effet d'avoir rapidement une vue d'ensemble de la situation d'épargne d'une succursale.

1.3.2 Objectif 2 : Regroupement des succursales semblables et prédiction

Par la suite, l'outil de classification est élaboré. Il se divise en deux parties. D'une part, les succursales sont regroupées en tenant compte de leurs caractéristiques d'épargne. D'autre part, une méthode permettant de connaître le groupe d'appartenance d'une nouvelle succursale est envisagée.

Objectif 2 a : Classification des succursales

Les groupes de succursales sont formés sans considérer de groupes de référence. Les indicateurs d'épargne créés à l'objectif 1 ainsi que certaines variables brutes du jeu de données sont utilisés pour la classification.

Les différents regroupements obtenus sont comparés entre eux afin de choisir ceux qui semblent les plus pertinents. Les facteurs qui distinguent les groupes de façon importante sont analysés afin d'interpréter les groupes de manière intelligible. Avec l'aide des conseillères en intelligence d'affaires, un ensemble de groupes est sélectionné.

Objectif 2 b : Assignation d'un groupe à une nouvelle succursale

Une méthode est proposée afin de déterminer le groupe d'appartenance d'une nouvelle succursale. Les groupes de référence sont ceux choisis à l'objectif 2 a. Cette méthode de classification est utile lorsque, pour des raisons administratives ou suite à une fusion, de nouvelles succursales doivent être assignées à un groupe. Elle permet aux gestionnaires de pouvoir classer rapidement ces nouvelles succursales et d'avoir une idée de leur performance éventuelle.

En définitive, cet outil de classification est important car il permet d'associer un profil d'épargne à une succursale. Le comportement des succursales ayant un certain profil peut être analysé afin de mieux comprendre les facteurs affectant leur performance quant à l'épargne. De plus, les différentes succursales d'un même profil peuvent être comparées entre elles.

1.3.3 Objectif 3 : Prédiction de l'atteinte de l'objectif annuel de *ventes-retraits*

Une fois les indicateurs et les groupes créés, des méthodes sont explorées afin de prédire l'atteinte de l'objectif annuel de *ventes-retraits* d'une succursale. L'atteinte de l'objectif est définie par un montant cumulatif de *ventes-retraits* supérieur à l'objectif annuel en date du 31 décembre de l'année en cours. Deux modèles de prévision sont construits afin de prédire l'atteinte de l'objectif d'une succursale à partir des observations du mois de juin.

L'une des questions des gestionnaires est de savoir si le profil d'épargne d'une succursale au mois de juin est associé à l'atteinte de son objectif. L'influence des groupes formés à l'objectif 2 selon l'atteinte de l'objectif annuel est donc évaluée.

La prédiction de l'atteinte de l'objectif annuel permet à l'équipe d'optimisation de la performance des succursales d'avoir une idée des succursales ou des groupes de succursales risquant de ne pas atteindre leur objectif individuel. L'équipe peut alors déployer des stratégies d'intervention semblables au sein d'un même groupe afin de redresser la situation et de permettre au plus grand nombre de succursales d'atteindre leur objectif annuel.

2 Collecte et jeu de données

Les données utilisées pour ce projet ne proviennent pas d'une expérience ou d'un échantillonnage, mais plutôt d'observations faites sur chacune des succursales à intervalle régulier. Le jeu de données est créé par les conseillères responsables du projet à partir des entrepôts de données de l'institution. Étant donné que ces entrepôts sont mis à jour de façon hebdomadaire, les observations sont disponibles pour chacune des semaines de l'année. Pour ce projet, les observations fournies sont celles qui correspondent à la dernière semaine de chaque trimestre entre décembre 2012 et juin 2014[3]. Il s'agit donc d'une image de la situation à sept moments fixes du temps. Les données de *ventes-retraits* de décembre 2014 sont aussi fournies afin d'ajuster le modèle de prévision.

Comme il est mentionné à la section 1.1, l'institution est composée de quelques centaines de succursales au Canada. Comme les observations ont été faites à sept reprises, le jeu de données contient quelques milliers d'observations. Pour des raisons de confidentialité, le nombre exact d'observations ne peut pas être divulgué. Pour ce qui est des variables, il en contient environ 200. La majorité des variables sont des moyennes et des proportions qui tracent le profil de la clientèle et de l'épargne d'une succursale à chaque temps d'observation.

2.1 Méthode de collecte

La collecte des informations servant à créer le jeu de données disponible pour ce projet se divise en trois grandes étapes : la saisie, l'entreposage et l'extraction.

[3]Cela correspond aux mois suivants : décembre 2012 (temps 1), mars 2013 (temps 2), juin 2013 (temps 3), septembre 2013 (temps 4), décembre 2013 (temps 5), mars 2014 (temps 6) et juin 2014 (temps 7)

La première étape, la saisie, est effectuée par les commis et les planificateurs financiers dans chaque succursale. Ce sont eux qui entrent dans le système informatique toutes les informations relatives à la clientèle et à son utilisation des services offerts. D'une part, les informations personnelles du client sont saisies, par exemple son sexe, son état civil et son âge. Ces informations permettent de connaître le profil sociodémographique de chaque succursale. D'autre part, des informations concernant les ventes et les retraits effectués dans la succursale sont aussi enregistrées.

Toutes les semaines, l'ensemble des données de chaque succursale est mis en entrepôt. À cette étape, un système de vérification automatisé s'assure de la validité des données recueillies. Si un signal est lancé par le système, une vérification manuelle de la donnée problématique est effectuée. Les entrepôts sont évidemment beaucoup trop vastes pour ce projet puisque, d'une part, seulement la vente et le retrait de produits ayant un lien avec l'épargne sont étudiés et que, d'autre part, les données utilisées sont agrégées par succursale. En effet, dans ces banques de données, une observation correspond à un client.

La troisième étape est donc celle où les conseillères extraient, au moyen d'une procédure informatique, les informations composant le jeu de données pour ce projet. Les données brutes sont alors manipulées pour en faire des moyennes et des proportions par succursale.

2.1.1 Variables modalités

Une remarque est à faire au sujet de certaines variables. Le jeu de données utilisé est créé en agrégeant les données par succursale à partir de banques de données où l'unité d'observation est le client. Cependant, certaines variables se trouvant dans les banques de données sont catégoriques. Au moment de créer le jeu de données, les modalités de chacune de ces variables deviennent elles-mêmes des variables. L'exemple de la variable donnant le cycle de vie financier d'un client permet de clarifier cette situation. Dans les banques de données, la variable cycle de vie est une variable catégorique à huit modalités. Elle représente l'appartenance du client à l'un des huit cycles de vie financiers définis par l'institution. Lors de la création du jeu de données, ces huit modalités deviennent des variables représentant les proportions de clients dans la succursale se trouvant dans chacun des huit cycles de vie financiers. À des fins pratiques, les variables de ce type sont nommées *variables modalités* dans la suite du rapport.

2.2 Problèmes observés et réduction du jeu de données

Suite à l'étude des variables et de leur description, certaines d'entre elles ont été retirées du jeu de données. Les différentes raisons ayant guidé ces choix sont énumérées dans cette section. La description détaillée des variables utilisées pour les analyses présentées dans ce rapport se trouve à la prochaine section.

Information redondante

Certaines variables présentent les mêmes informations. Par exemple, la variable « anmois » peut prendre sept modalités identifiant l'année et le mois où les données ont été récoltées (par exemple « 201212 », « 201303 », etc.). De son côté, la variable « temps » contient aussi sept modalités (les entiers 1 à 7) correspondant aux différents temps identifiés par la variable « anmois ». Ainsi, la connaissance de la variable « temps » entraîne la connaissance de la variable « anmois » et vice-versa. La variable « anmois » ainsi que toutes les variables de ce type ont donc été retirées.

Variables construites

Plusieurs variables sont des indicateurs construits à partir de variables qui, dans certains cas, sont contenues dans le jeu de données. Comme cela induit des problèmes de multicolinéarité, les variables contenant les plus simples ont été conservées et les autres éliminées. Par exemple, la variable relation d'affaires est un indicateur calculé à partir de l'âge du client, de son épargne marché estimée, du solde de ses comptes, etc. Elle a donc été retirée du jeu de données.

Variables dont la variabilité est trop faible

Pour certaines variables, les données sont concentrées autour d'une certaine valeur avec une variabilité très faible. Par exemple, pour la variable indiquant le pourcentage de clients décédés, l'écart-type est 0.000035. Cette variable ne permet donc pas de discriminer les succursales et a été retirée.

Trop grand pourcentage d'information non disponible

Pour certaines *variables modalités*, la proportion de clients pour lesquels l'information est non disponible est très grande. Par exemple, pour la variable donnant la proportion d'entreprises ayant un certain type de financement, la proportion des entreprises pour lesquelles l'information n'est pas disponible est élevée (33 % en moyenne). Cette variable a donc été enlevée, de même que toutes les *variables modalités* pour lesquelles l'information non disponible est supérieure à 20 % en moyenne.

Problèmes de mise à jour

Pour certaines variables, l'information n'est pas mise à jour régulièrement. Comme il est possible que ces variables ne représentent pas adéquatement la clientèle, elles ont été retirées.

2.3 Description des variables

Les nombreuses variables du jeu de données peuvent être regroupées selon le type d'information qu'elles contiennent. Les catégories retenues sont les suivantes :

1. la variable « temps »
2. les variables permettant de décrire les succursales (10)
3. les variables concernant l'épargne des succursales (23)
4. les variables concernant les prêts (5)
5. les variables concernant les clients qui sont des particuliers (30)
6. les variables concernant les clients qui sont des entreprises (2)
7. les variables donnant les *ventes-retraits* cumulatives et les objectifs à atteindre (2)

Afin d'avoir une idée des données disponibles, ces catégories seront décrites brièvement. La liste complète des variables ainsi qu'une courte description de chacune se trouvent à l'annexe 1.

La variable « temps » est celle qui donne le temps où l'observation a été effectuée. Les sept temps d'observation ont été donnés dans la note de bas de page numéro 3 à la section 2.1.

Les variables de la deuxième catégorie sont celles permettant d'identifier et de décrire les succursales de manière globale. Elles sont, par exemple, le nombre de clients particuliers et le nombre de clients entreprises d'une succursale. Ces variables sont mesurées à la succursale directement.

Les variables des catégories 3 à 6 inclusivement ont été agrégées par succursale à partir de données récoltées sur les clients. La troisième catégorie regroupe les variables décrivant l'épargne des clients d'une succursale. On distingue deux types d'information : des montants moyens et des proportions. D'une part, certaines variables correspondent au montant moyen, en dollars, investi par les clients d'une succursale dans les différents produits d'épargne offerts. D'autre part, certaines variables donnent la proportion de clients d'une succursale ayant investi dans chaque produit d'épargne. Les produits d'épargne en question seront présentés davantage à la section 3.2.

Similairement, les variables de prêts décrivent les montants moyens, en dollars, des prêts accordés aux clients d'une succursale ainsi que la proportion de clients ayant chaque type de prêt.

La cinquième catégorie contient les variables concernant les clients particuliers des succursales. On y trouve, par exemple, la moyenne de l'ancienneté des clients d'une succursale et les proportions de clients de chaque sexe. Plusieurs des variables de cette catégorie sont des *variables modalités* telles que décrites à la section 2.1.1.

Pour ce qui est de la sixième catégorie, concernant les clients entreprises, les variables disponibles sont le chiffre d'affaires estimé moyen par client entreprise d'une succursale et le nombre de clients utilisant des structures particulières de services aux entreprises.

Finalement, la septième catégorie contient deux variables. D'abord, la variable des *ventes-retraits* cumulatives représente la somme de tous les produits d'épargne vendus depuis le début de l'année de laquelle est soustrait l'argent retiré des produits d'épargne pendant cette même période. Comme il a déjà été mentionné, cette variable peut prendre des valeurs négatives. Elle est utile pour mesurer la capacité de ventes et de rétention de la clientèle de chaque succursale. Il est à préciser que lorsqu'un client retire son argent, le montant soustrait de la variable *ventes-retraits* est le montant initialement mis dans le produit d'épargne (sans les intérêts ou les pertes). De plus, lorsqu'un client transfert son argent d'un produit d'épargne à un autre, l'effet sur la variable des *ventes-retraits* cumulatives est nul. Pour les dernières semaines de décembre 2013 et de juin 2014, la variable des *ventes-retraits* cumulatives représente le montant total provenant de la vente et des retraits dans la succursale pendant l'année correspondante. Enfin, la deuxième variable est celle qui donne les objectifs d'épargne à atteindre pour chaque succursale. Les deux variables de cette catégorie ont été mesurées quatre fois : en juin et décembre 2013 et en juin et décembre 2014.

2.4 Manipulations effectuées sur les variables

Suite à l'épuration du jeu de données, les variables restantes ont été renommées avec des noms courts, confidentiels et intelligibles. De plus, certaines manipulations mathématiques ont été effectuées. Pour les *variables modalités* conservées, la démarche suivante a été utilisée :

1. Retirer la modalité correspondant à l'information non disponible
2. Diviser chaque modalité par la somme de toutes les colonnes correspondant à une information disponible
3. Multiplier les résultats par 100 pour obtenir des pourcentages plutôt que des proportions

Les nouvelles *variables modalités* représentent donc le pourcentage de clients dont l'information est disponible et ayant une certaine caractéristique. Dans le cas de variables dichotomiques, seulement la première modalité a été conservée. Enfin, les proportions ont été multipliées par 100 afin d'obtenir des pourcentages.

De plus, la variable donnant le nombre de concurrents de l'institution présents dans l'aire de diffusion[4] de la succursale (nb_concur) a été catégorisée, car elle présentait des valeurs extrêmes. Les catégories ont été créées de telle sorte que chacune contient environ un tiers des observations de juin 2014. Elle a été modifiée de la façon suivante :

1. Si nb_concur = 0, la catégorie « 0 » est attribuée à la nouvelle variable
2. Si nb_concur = 1, 2, 3, 4 ou 5, la catégorie « 1 à 5 » est attribuée à la nouvelle variable
3. Si nb_concur >= 6, la catégorie « 6 et plus » est attribuée à la nouvelle variable

3 Objectif 1 : Création d'indicateurs

Afin de faciliter l'interprétation des groupes formés (section 4) et d'ajuster un bon modèle prédictif (section 5), il est essentiel de réduire davantage la dimension du jeu de données et de chercher une solution au problème de multicolinéarité entre les variables. Pour ce faire, le premier objectif de ce projet est de concentrer l'information contenue dans les variables fortement corrélées en un certain nombre d'indicateurs interprétables. La méthode utilisée afin de créer ces indicateurs ainsi que leur interprétation sont décrites dans cette section.

3.1 Méthode

Afin de répondre à cet objectif, la méthode statistique choisie est l'analyse factorielle. Les analyses présentées dans ce rapport ont été réalisées avec le logiciel SAS (version 9.4). La méthode utilisée pour créer les indicateurs est la suivante :

1. Réaliser une analyse en composantes principales (ACP) sur les corrélations d'un ensemble de variables et sur la totalité des observations disponibles à l'aide de la procédure PRINCOMP
2. Garder le nombre minimal de composantes principales permettant d'expliquer au moins 70% de la variabilité
3. Effectuer une rotation VARIMAX avec la procédure FACTOR
4. Utiliser comme critère final pour le choix des indicateurs la facilité d'interprétation de leurs facteurs

3.2 Indicateurs d'épargne

Pour ce qui est des variables concernant l'épargne des succursales (catégorie 3 de la section 2.3), les analyses ont été effectuées en utilisant trois sous-ensembles de variables représentant des caractéristiques liées à l'épargne des succursales (les montants moyens détenus par la clientèle, les ratios de ces montants par rapport à l'épargne totale de la succursale et les proportions de clients détenant certains produits d'épargne). Le sous-ensemble des variables correspondant aux montants moyens des différents produits d'épargne détenus par les clients a été choisi, car les résultats obtenus

[4] L'aire de diffusion est une unité géographique définie par Statistique Canada

sont plus stables dans le temps[5] et plus faciles à interpréter. Les résultats de l'analyse factorielle sont présentés au tableau 1.

Le facteur 1 correspond à une détention de produits d'épargne répondant aux besoins d'une clientèle aisée[6] et recevant les services de d'autres composantes de l'institution financière. Aussi, il correspond à une offre de produits moins conventionnels (produit d'épargne aisé 1 et produit d'épargne aisé 2). Le facteur 2 correspond plutôt à une détention de produits correspondant à une clientèle fortunée[7]. Cette clientèle a besoin de soutien pour la gestion de son patrimoine et recherche des stratégies de placement fiscalement avantageuses. De son côté, le facteur 3 correspond plutôt au profil des clients utilisant uniquement les services offerts en succursale et recherchant des possibilités de rendement supérieures à celles des produits conventionnels. Par exemple, ces clients détiennent des produits liés au marché. Finalement, le facteur 4 correspond à une détention de produits très sécuritaires dont le rendement est connu à l'avance et qui sont offerts en succursale. Il représente donc une clientèle au profil plus conservateur.

Ainsi, l'interprétation des facteurs indique que les principales composantes de la variabilité sont expliquées par les niveaux de richesse et de tolérance au risque de la clientèle.

Tableau 1 : Résultats de l'analyse factorielle effectuée sur les variables correspondant aux montants moyens, par succursale, des différents produits d'épargne détenus par les clients (le nombre ombragé correspond au score le plus élevé en valeur absolu pour cette variable)

	Facteur 1	**Facteur 2**	**Facteur 3**	**Facteur 4**
Valeurs mobilières	0.91980	0.20527	0.13265	0.00961
VM - Plein exercice	0.90745	0.09278	0.01266	-0.03429
Produit d'épargne aisé 1	0.55513	0.21015	0.50910	0.03910
Produit d'épargne aisé 2	0.48419	0.09876	0.31048	0.07051
Produit d'épargne fortuné 1	0.10559	0.88554	0.05834	-0.00137
Produit d'épargne fortuné 2	0.21252	0.83779	0.12334	0.00241
Produit d'épargne lié au marché 1	0.01937	0.00873	0.89024	0.08934
Produit d'épargne lié au marché 2	0.44773	0.21222	0.61094	-0.09922
Compte d'épargne conventionnel	-0.21879	-0.07835	-0.04569	0.78956
Certificat de placement garanti	0.27613	0.08663	0.11728	0.75009

[5] Au départ, les analyses sur les trois groupes de variables ont été effectuées sur les données du temps 7 seulement. Ensuite, les mêmes analyses ont été reprises sur les données de tous les temps. Seul le groupe de variables retenu présentait les mêmes résultats. Les résultats ont donc été considérés plus stables dans le temps.
[6] Les clients aisés sont ceux qui détiennent entre 100 000 $ et 1 000 000 $ en produits d'épargne.
[7] Les clients fortunés sont ceux qui détiennent plus de 1 000 000 $ en produits d'épargne.

3.3 Indicateurs sociodémographiques

Les variables sociodémographiques sont pour la plupart des *variables modalités* telles que définies à la section 2.1.1. Évidemment, les différentes modalités de ces variables sont corrélées. Afin de réduire la dimension du jeu de données et de créer de nouvelles variables indépendantes les unes des autres, des scores sont calculés sur chaque *variable modalité* avec la méthode décrite plus haut.

La variable cycle de vie

Cette *variable modalité* représente les pourcentages de clients se trouvant dans chacun des huit cycles de vie financiers définis par l'institution. Les huit modalités sont réduites à trois scores en conservant 76.64 % de la variabilité. Le tableau 2 présente les résultats de l'analyse factorielle.

Le facteur 1 décrit une clientèle très active. En effet, il correspond à une succursale pour laquelle beaucoup de clients sont dans les catégories projets divers et accédant à la propriété et où peu de clients sont retraités. Le facteur 2 correspond à une clientèle très jeune, où il y a beaucoup de clients de moins de 18 ans, d'étudiants et de jeunes travailleurs et peu de clients en préparation à la retraite. Le facteur 3 correspond à une succursale où les clients sont de nouveaux propriétaires.

Tableau 2 : Résultats de l'analyse factorielle effectuée sur les variables correspondant aux différents pourcentages de clients à l'intérieur de chaque cycle de vie financiers (le nombre ombragé correspond au score le plus élevé en valeur absolue pour cette variable)

	Facteur 1	**Facteur 2**	**Facteur 3**
Projets divers	0.85764	-0.01674	-0.28140
Accédant propriété	0.84029	0.20748	0.07205
Retraités	-0.69135	-0.37546	-0.44826
Moins de 18 ans	-0.29299	0.78236	0.28582
Étudiants	0.36354	0.68959	0.14989
Jeunes travailleurs	0.51212	0.62464	-0.13894
Préparation retraite	-0.23285	-0.83368	0.03359
Nouveaux propriétaires	-0.05261	0.06237	0.95040

La variable niveaux de richesse

Cette *variable modalité* représente les pourcentages de clients de 18 ans et plus dans chacun des cinq niveaux de richesse définis par l'institution. Le niveau de richesse est une variable construite avec l'estimation de l'épargne du client en considérant aussi la détention de produits financiers auprès d'institutions concurrentes. Les cinq modalités sont réduites à deux scores qui permettent d'expliquer 83.36 % de la variabilité. Le tableau 3 présente les résultats de l'analyse factorielle.

Le premier facteur correspond aux pratiques d'épargne chez les clients n'étant ni aisés, ni fortunés. Il décrit une succursale ayant beaucoup de clients accumulateurs et peu de clients utilisateurs. Les clients accumulateurs sont ceux qui achètent des produits d'épargne, alors que les clients utilisateurs ont peu d'économies (moins de 10 000 $), sont sans emploi ou reçoivent de l'aide au logement. Le deuxième facteur correspond à la richesse de la clientèle de la succursale. En effet, il décrit une succursale ayant beaucoup de clients aisés et de clients fortunés et peu de clients bâtisseurs. Les clients bâtisseurs sont généralement des étudiants ou de jeunes travailleurs accédant à la propriété.

Tableau 3 : Résultats de l'analyse factorielle effectuée sur les variables correspondant aux différents pourcentages de clients à l'intérieur de chaque niveau de richesse (le nombre ombragé correspond au score le plus élevé en valeur absolue pour cette variable)

	Facteur 1	**Facteur 2**
Accumulateurs	0.92573	0.09029
Utilisateurs	-0.95971	-0.02256
Fortunés	0.10108	0.83480
Aisés	0.65913	0.71925
Bâtisseurs	0.02573	-0.84960

3.4 Discussion

En résumé, des indicateurs ont été créés pour les variables correspondant à des montants moyens de produits d'épargne et pour les variables « cycle de vie » et « niveau de richesse ». Ces indicateurs présentent de nombreux avantages : ils ne sont pas corrélés entre eux, ils sont facilement interprétables et ils permettent de réduire le nombre de variables du jeu de données. Ils peuvent être observés pour avoir une idée du profil d'une succursale. Ils peuvent aussi être utilisés dans des modèles statistiques comme ceux présentés à la section 5. Enfin, la méthode proposée pourrait être utilisée pour créer d'autres indicateurs au besoin.

Néanmoins, cette méthode a quelques désavantages. D'abord, le calcul des facteurs doit être fait pour chaque temps auquel on souhaite utiliser les indicateurs. Heureusement, la méthode est facilement reproductible avec le logiciel SAS. Il faut aussi noter que bien que les indicateurs semblent stables dans le temps, il est possible que des analyses ultérieures donnent des résultats très différents de ceux présentés ici. Néanmoins, comme il s'agit d'analyses simples à effectuer, il est tout à fait possible de les répéter et d'en faire une nouvelle interprétation le cas échéant.

4 Objectif 2 : Regroupement des succursales semblables et prédiction

Comme il a été mentionné à la section 1.3, l'objectif principal de ce projet est d'élaborer un outil de classification des succursales basé sur des caractéristiques en lien avec l'épargne. Le regroupement des succursales permet à l'équipe d'optimisation de la performance de cibler les actions à poser lorsqu'une succursale éprouve des difficultés, et ce, en la comparant à des succursales similaires. La création de cet outil se divise en deux étapes. Premièrement, des groupes de succursales sont formés à partir d'un sous-ensemble de variables concernant l'épargne. Deuxièmement, une méthode permettant de classer une nouvelle succursale dans l'un des groupes formés est présentée.

4.1 Objectif 2a : Classification des succursales

Afin de créer des groupes de succursales dont les caractéristiques liées à l'épargne sont similaires, la classification non supervisée (*clustering*) est utilisée. Cette approche statistique permet de former des groupes d'observations de sorte que les observations d'un même groupe se ressemblent et que les observations de groupes différents se distinguent.

Ultimement, le choix des groupes proposé est guidé par les besoins de l'équipe d'optimisation. Les contraintes spécifiées par les conseillères sont les suivantes :

- le nombre de groupe doit être entre 5 et 15
- les groupes doivent idéalement être de taille semblable
- les groupes doivent être interprétables

4.1.1 Méthode

La méthode présentée dans cette section est réalisée avec le logiciel SAS. Elle comporte de nombreux choix subjectifs.

Parmi toutes les variables du jeu de données, les variables retenues pour effectuer la classification sont celles décrivant l'épargne des succursales. Ce sous-ensemble de variables a été jugé le plus approprié pour former des groupes de succursales dont les caractéristiques ayant trait à l'épargne sont comparables. Comme les variables disponibles sont continues, le choix d'un algorithme de classification ascendant est privilégié.

Différentes méthodes sont alors possibles pour regrouper les observations à chaque étape de l'algorithme. Elles se distinguent par la technique utilisée pour mesurer la distance entre les observations. Plusieurs de ces méthodes sont implémentées en SAS. Chacune a ses avantages et ses inconvénients, et il n'existe pas de choix optimal. Les essais réalisés ont été faits avec les méthodes de la moyenne, du centroïde et de Ward. Ces deux premières méthodes ont l'avantage d'être robustes en présence de données aberrantes. Par contre, elles forment généralement un très gros groupe et de très petits groupes. Pour sa part, la méthode de Ward est sensible aux données aberrantes, mais a l'avantage de former des groupes de taille plus équilibrée.

Une fois les observations regroupées de manière hiérarchique, il faut choisir le nombre de groupes à conserver. Les critères utilisés pour guider ce choix sont le *cubic clustering criterion* (CCC), la statistique pseudo-t^2 et la statistique pseudo-F.

L'interprétation des groupes est ensuite effectuée à l'aide de comparaisons multiples sur les variables d'intérêt avec la procédure GLM. Les moyennes à l'intérieur de chaque groupe sont comparées afin d'identifier les groupes qui se démarquent sur certains facteurs.

Deux notes importantes sont à faire. D'abord, la classification est faite à partir des variables standardisées, car celles-ci n'ont pas toutes le même ordre de grandeur. Cela évite que les variables ayant une grande variabilité influencent davantage les regroupements faits par l'algorithme. Deuxièmement, les données de juin 2014 sont utilisées afin de former les groupes à partir des données les plus récentes.

En résumé, la démarche utilisée est la suivante :

1. Standardiser les données avec la procédure STANDARD
2. Effectuer la classification non supervisée avec la procédure CLUSTER en spécifiant une méthode pour regrouper les observations (option METHOD=)
3. Étudier les graphiques des critères de sélection (CCC, pseudo-t^2 et pseudo-F) en fonction du nombre de groupes afin de choisir le nombre de groupes à conserver
4. Associer les groupes créés à chaque observation du jeu de données avec la procédure TREE
5. Interpréter les groupes formés en effectuant des comparaisons multiples sur les variables jugées pertinentes avec la procédure GLM (énoncé LSMEANS)

Plusieurs tentatives de création de groupes ont été effectuées avec différentes combinaisons de variables et les trois méthodes énoncées plus haut. Le choix retenu est celui qui répond le mieux aux demandes des conseillères.

Il est à noter que les pourcentages présentés dans cette section et la suivante ont été arrondis afin de simplifier la présentation et d'assurer la confidentialité des données

4.1.2 Présentation et interprétation des groupes retenus

Pour créer les groupes présentés dans cette section, la méthode de Ward et les variables suivantes ont été utilisées :

- les quatre scores d'épargne créés à l'objectif 1
- les pourcentages de clients faisant affaire avec l'équipe 1 et l'équipe 2
- le montant de ventes virtuelles
- le montant moyen du volume d'affaires de la succursale

Avec ces huit variables, la statistique pseudo-t^2 et la statistique pseudo-F concordent et suggèrent de conserver huit groupes pour classer les succursales. De plus, parmi ces huit groupes, le plus petit contient un peu plus de 1 % des succursales, ce qui n'est pas énorme, mais tout de même satisfaisant. Seulement trois des huit groupes contiennent chacun moins de 2% des succursales.

Une fois les groupes obtenus, des comparaisons multiples ont été effectuées pour les comparer. Seules les différences significatives entre les groupes ont été considérées. Le tableau 4 présente les caractéristiques des différents groupes ainsi que le pourcentage du nombre de succursales qu'ils contiennent.

Tableau 4 : Pourcentage (%) du nombre de succursales contenu dans chacun des groupes formés par classification non supervisée avec la méthode de Ward et interprétation des groupes formés

Groupe	%	Caractéristiques et interprétation
1	42%	Le groupe 1 fait partie de ceux ayant un faible volume d'affaires. En effet, seuls les groupes 3 et 8 ont un volume d'affaires moyen inférieur ou égal à celui du groupe 1. Ce groupe se distingue du groupe 3 par des scores d'épargne plus élevés et des pourcentages plus élevés pour l'équipe 1 et l'équipe 2[8]. En raison de leur faible volume d'affaires, ces succursales sont probablement petites.
2	19%	Le groupe 2 est celui dont les succursales ont les plus grandes valeurs sur le premier score d'épargne. Ce sont donc des succursales ayant beaucoup de clients aisés et recevant les services de d'autres composantes de l'institution financière.
3	15%	Le groupe 3 est celui ayant le plus faible volume d'affaires moyen. Le pourcentage de clients faisant affaire avec l'équipe 3 est plus élevé que dans le groupe 1. Ce groupe est parmi ceux ayant les moyennes les plus basses pour tous les scores d'épargne. Ces deux dernières caractéristiques suggèrent que les clients de ces succursales ne sont pas de grands épargnants. Les succursales de ce groupe sont probablement petites.
4	12%	Le groupe 4 est parmi les plus élevés pour le troisième score d'épargne. Ce groupe a un pourcentage de clients faisant affaire avec l'équipe 2 parmi les plus élevés. Le premier score d'épargne, le pourcentage de clients faisant affaire avec l'équipe 1 et le montant moyen du volume d'affaires de ce groupe sont plus élevés que ceux du groupe 8. Les succursales de ce groupe ont un gros volume d'affaires, une clientèle plutôt aisée et vendent beaucoup de *produits maison*.
5	7%	Le groupe 5 se distingue comme étant celui ayant les plus fortes ventes virtuelles. Selon les conseillères, ces succursales se trouvent peut-être dans de grands centres urbains.

[8] L'équipe 1 est spécialisée en épargne alors que l'équipe 2 est une équipe généraliste qui offre des services de financement et d'épargne. Pour sa part, l'équipe 3 est celle qui s'occupe des petites transactions.

Groupe	%	Caractéristiques et interprétation
6	2%	Le groupe 6 se distingue comme étant celui ayant la moyenne la plus élevée pour le quatrième score d'épargne et le plus bas pourcentage de clients faisant affaire avec l'équipe 2. Ainsi, il semble que sa clientèle soit plus conservatrice et possiblement plus âgée.
7	2%	Le groupe 7 se distingue comme étant celui ayant la moyenne la plus élevée pour le deuxième score d'épargne et celui dont le pourcentage d'équipe 1 est le plus élevé. Les succursales de ce groupe semblent donc se situer dans des milieux plus favorisés. Elles ont aussi une plus grande proportion de clients fortunés.
8	1%	Le groupe 8 a le plus grand pourcentage de clients faisant affaire avec l'équipe 2. Ce groupe est parmi ceux dont la moyenne du troisième score d'épargne est la plus élevée. Les succursales de ce groupe semblent avoir une clientèle plutôt aisée (mais moins que celles du groupe 4) et semblent vendre beaucoup de *produits maison*.

Comme les caractéristiques des groupes 4 et 8 se ressemblent beaucoup, la fusion de ces groupes pourrait être envisagée. Cela permettrait notamment de simplifier l'outil de prédiction du groupe d'appartenance d'une nouvelle succursale.

4.1.3 Instabilité des groupes dans le temps

Il est important d'apporter quelques nuances concernant la composition des groupes formés pour répondre à cet objectif. Les groupes de succursales ont été créés avec les observations de juin 2014. Il est fort possible qu'à un autre temps, le groupe assigné à chaque succursale soit différent. Des vérifications ont été effectuées afin d'établir si, d'un temps à l'autre, les succursales restent dans le même groupe. Voici la démarche utilisée pour réaliser ces vérifications :

1. Effectuer la classification à l'aide de la démarche présentée à la section 4.1.1 en utilisant les données d'un autre temps d'observation, mais en gardant obligatoirement 8 groupes (donc en n'étudiant pas les critères de sélection)
2. Créer un jeu de données contenant l'identifiant de la succursale, le numéro du groupe dans lequel la succursale est classée en juin 2014 et le numéro du groupe dans lequel la succursale est classée à l'autre temps testé
3. Créer un tableau de fréquences croisant le numéro du groupe dans lequel la succursale est classée en juin 2014 et le numéro du groupe dans lequel la succursale est classée à l'autre temps testé
4. Répéter les étapes 1 à 3 pour tous les temps à tester

Il est à noter que le numéro associé à un groupe dépend uniquement de sa taille. Par exemple, les groupes 3 et 4 en juin 2014 sont de taille comparable. Si quelques succursales ont changé de groupe depuis mars 2014, il se peut que les numéros de ces groupes aient été interchangés. Afin de trouver les correspondances entre les groupes à travers le temps, les groupes sont associés selon les fréquences

croisées maximales. Le tableau ci-dessous présente le tableau de fréquences croisées obtenu en comparant les groupes obtenus en juin 2014 et ceux obtenus en mars 2014 pour chaque succursale.

Tableau 5 : Tableau croisant le groupe d'appartenance des observations de juin 2014 et le groupe d'appartenance des observations de mars 2014, en pourcentage (les cases ombragées correspondent au pourcentage de succursales classées dans le même groupe d'un temps à l'autre)

		Groupes en mars 2014								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**
Groupes en juin 2014	**1**	34	8	0	0	<1	0	0	0	**42**
	2	9	0	9	0	<1	0	0	0	**19**
	3	<1	14	0	0	0	<1	0	0	**15**
	4	2	<1	3	6	0	0	0	0	**12**
	5	<1	0	<1	0	6	0	0	0	**7**
	6	0	0	0	0	0	2	0	0	**2**
	7	0	0	0	0	0	0	1	0	**2**
	8	0	0	0	0	0	0	0	1	**1**
	Total	**46**	**22**	**12**	**6**	**8**	**2**	**2**	**1**	**100**

Le groupe 2 de juin 2014 semble correspondre au groupe 3 de mars 2014 (avec beaucoup d'erreurs de classement). De la même façon, le groupe 3 de juin 2014 semble correspondre au groupe 2 de mars 2014. Cela illustre les variations possibles dans l'attribution du numéro des groupes.

Pour obtenir le pourcentage de classification homogène dans le temps, la proportion de succursales qui semblent appartenir au même groupe est calculée. Dans ce cas-ci, environ 74 % des succursales ont été classées dans le même groupe. Une démarche semblable a montré qu'entre décembre 2012 et juin 2014, 59 % des succursales ont été classées dans le même groupe. Il est donc raisonnable d'affirmer que les groupes d'appartenance des succursales changent avec le temps.

4.1.4 Discussion

Évidemment, l'instabilité des groupes dans le temps est un enjeu important. Il serait inapproprié d'attribuer un groupe à une succursale aveuglément et d'utiliser les interprétations décrites ci-haut. Cependant, il est possible que les succursales qui changent de groupe d'un temps à un autre aient en fait changé de profil. Dans ce cas, le groupe associé à une succursale à un temps donné permettrait de guider le choix d'une stratégie d'intervention adaptée. Quoi qu'il en soit, il n'est pas possible de juger de l'efficacité réelle de cette méthode avec les analyses présentées ici. Les groupes proposés doivent donc être utilisés prudemment. Néanmoins, la méthode proposée semble valable si elle est reproduite à intervalle régulier.

4.2 Objectif 2b

Comme les groupes ne sont pas stables dans le temps, il est peut être inutile de prédire à quel groupe une nouvelle succursale appartiendra. Cependant, si on considère plutôt que les groupes sont des profils d'épargne, il est intéressant de classer une nouvelle succursale. En effet, on lui attribue alors un profil d'épargne pouvant aider l'équipe d'optimisation de la performance dans le choix des stratégies à appliquer. Pour ce faire, l'analyse discriminante est utilisée.

4.2.1 Méthode

En SAS, la procédure DISCRIM, qui permet d'effectuer une analyse discriminante, suppose par défaut la normalité multidimensionnelle des données. Ce postulat est rejeté par le test de Mardia[9] basé sur les coefficients d'asymétrie et d'aplatissement. Une analyse discriminante non paramétrique basée sur les *k* plus proches voisins a donc été privilégiée. Elle est effectuée en spécifiant l'option METHOD=NPAR. La constante *k* a été prise près de la racine carrée du nombre d'observations[10].

L'analyse discriminante est menée avec comme jeu de données d'entraînement les observations de juin 2014 et leur groupe associé. Les variables utilisées sont les mêmes que celles retenues à la section 4.1.2. Le tableau 6, construit de la même manière que le précédent, croise le groupe d'appartenance des succursales en mars et en juin 2014. Il montre que cette méthode a tendance à classer plusieurs observations des groupes 1 et 2 dans le groupe 5, ce qui est surprenant étant donné que l'interprétation faite de ces groupes est assez différente. La proportion de succursales ayant été classées dans le même groupe aux deux temps d'observations est de 64 %. Elle est donc inférieure à celle obtenue en reconstruisant les groupes.

Tableau 6 : Tableau croisant le groupe d'appartenance des observations de juin 2014 (obtenu par classification non supervisée) et le groupe d'appartenance des observations de mars 2014 (obtenu par analyse discriminante), en pourcentage (les cases ombragées correspondent au pourcentage de succursales classées dans le même groupe d'un temps à l'autre)

		Groupes en mars 2014								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**
Groupes en juin 2014	**1**	21	<1	5	2	13	0	0	0	**42**
	2	2	9	0	2	6	0	0	0	**19**
	3	1	0	13	0	0	<1	0	0	**15**
	4	<1	0	<1	9	2	0	0	0	**12**
	5	0	0	0	0	7	0	0	0	**7**
	6	0	0	0	0	0	2	0	0	**2**
	7	0	0	0	0	0	0	2	0	**2**
	8	0	0	0	0	0	0	0	1	**1**
	Total	**25**	**10**	**18**	**13**	**28**	**2**	**2**	**1**	**100**

[9] Une macro permettant d'effectuer ce test a été fournie par le service de consultation statistique de l'Université Laval

[10] La règle du pouce est de prendre k proche de la racine carrée du nombre d'observations

4.3 Discussion sur l'objectif

Il est difficile d'évaluer la performance de la méthode présentée pour prédire le groupe d'appartenance d'une succursale. En effet, les erreurs de classification proviennent à la fois de la méthode utilisée et des variations du profil des succursales dans le temps. Cependant, l'analyse discriminante semble être moins fiable que la reconstruction des groupes, car elle a tendance à classifier beaucoup d'observations dans le groupe 5. Toutefois, les résultats présentés dépendent évidemment des données et il est possible qu'ils soient très différents lorsqu'un autre ensemble de données sera utilisé.

En somme, l'utilisation de cet outil de classification peut s'avérer intéressante à l'étape exploratoire et comme complément à l'analyse des données socio-économiques concernant les succursales afin de leur associer un profil d'épargne. Il sera toutefois nécessaire de reconstruire les groupes régulièrement et de garder en tête les problèmes soulevés.

5 Objectif 3 : Prédiction de l'atteinte d'un objectif annuel de *ventes-retraits*

Il est important pour l'équipe d'optimisation de la performance d'identifier les succursales auprès desquelles elle doit intervenir. Généralement, cette tâche est réalisée en observant et en analysant leur profil à un moment fixe du temps. À l'aide d'analyses statistiques et de tableaux de bord, les gestionnaires ont une bonne idée des succursales en difficulté. Ces méthodes donnent de bons résultats, mais il est tout de même intéressant d'essayer de les complémenter.

Une approche alternative proposée par les conseillères en intelligence d'affaires est de tenter de prédire l'atteinte de l'objectif à partir des données disponibles au cours de l'année à l'aide d'un modèle statistique. Le dernier objectif de ce projet vise donc à ajuster un certain nombre de modèles aux données observées en juin pour prédire si une succursale atteindra son objectif annuel de *ventes-retraits* en décembre.

Tout d'abord, une méthode simple est détaillée. Il s'agit de prédire l'atteinte de l'objectif annuel de *ventes-retraits* en se fiant à l'atteinte de l'objectif partiel en juin. Ensuite, un modèle de régression logistique est présenté. Celui-ci permet de prédire la probabilité qu'une succursale atteigne son objectif annuel selon ses caractéristiques en juin. Enfin, le dernier modèle proposé est un modèle de régression linéaire. Il permet quant à lui de prédire le montant cumulatif annuel de *ventes-retraits* d'une succursale afin de le comparer à l'objectif fixé.

Les variables utilisées afin de déterminer l'atteinte de l'objectif sont les suivantes : le montant cumulatif de *ventes-retraits* et l'objectif de *ventes-retraits* fixé au préalable. Il s'agit des deux variables de la septième catégorie de la section 2.3. Pour ces deux variables, les observations de juin et décembre des années 2013 et 2014 sont disponibles. L'atteinte de l'objectif a été définie par un montant cumulatif de *ventes-retraits* supérieur à l'objectif fixé. Une nouvelle variable donnant l'atteinte de l'objectif d'épargne

a donc été créée en comparant les observations disponibles pour chaque succursale et ce, pour chaque temps d'observation.

Les modèles de régression ont été ajustés avec les données de l'année 2014, puis validés avec celles de l'année 2013. Il est important de noter qu'en 2014, entre juin et décembre, quelques paires de succursales ont été fusionnées. Au moment d'ajuster les modèles de régression, ces quelques succursales ont été retirées du jeu de données puisque la relation entre le profil de la succursale en juin et l'atteinte de l'objectif en décembre était confuse. De plus, sous les conseils des superviseures de ce projet, une succursale a été retirée puisqu'elle a eu, en 2014, une entrée d'argent exceptionnelle. Aucune observation n'a été retirée des données de 2013.

5.1 Stratégie simple : tableau de fréquences croisées

Il est intuitif de penser qu'une succursale ayant atteint son objectif en juin l'atteindra aussi en décembre. La stratégie proposée ici est de prédire l'atteinte de l'objectif en décembre en se fiant uniquement à l'atteinte de l'objectif en juin. Cette approche a été étudiée sur les données de l'année 2013. Elle est présentée afin d'être comparée avec les modèles développés aux sections 5.2 et 5.3.

Le tableau 7 montre que 21 % des succursales sont mal prédites par cette approche. Ce tableau permet aussi de calculer la sensibilité et la spécificité de la méthode de prédiction. La sensibilité, c'est-à-dire la probabilité de déclarer qu'une succursale n'atteindra pas son objectif et qu'en réalité elle ne l'atteigne pas, est de 83 %. La spécificité, c'est-à-dire la probabilité de déclarer qu'une succursale atteindra son objectif et qu'en réalité elle l'atteigne, est de 73 %.

Tableau 7 : Tableau de fréquences croisant l'atteinte de l'objectif de juin et l'atteinte de l'objectif en décembre pour les données de 2013, en pourcentage[11] (les cases ombragées représentent le pourcentage de succursales pour lesquelles l'atteinte de l'objectif est mal prédite)

		Réalité		
		Non	**Oui**	
Prédit	**Non**	49	11	**60**
	Oui	10	30	**40**
		59	**41**	**100**

Cette méthode a des limites importantes : elle ne tient pas compte des caractéristiques socio-économiques autres que l'atteinte de l'objectif en juin et elle ne permet pas d'associer à une succursale

[11] Les pourcentages sont préférés aux valeurs réelles pour des raisons de confidentialité.

une probabilité d'atteindre son objectif. Les modèles suivants permettent de remédier à ces lacunes, dans une certaine mesure.

5.2 Prédiction de la probabilité d'atteindre l'objectif d'épargne : modèle de régression logistique

Afin de prédire la probabilité qu'une succursale atteigne son objectif annuel de *ventes-retraits*, un modèle de régression logistique est ajusté. Cette section présente le modèle mathématique, les différents modèles ajustés, l'interprétation du meilleur modèle trouvé et une discussion sur le taux d'erreur mesuré sur les données de juin 2013.

5.2.1 Modèle mathématique

La fonction de lien utilisée est le logit. Le modèle est le suivant :

$$Y_i \sim Bernoulli(\pi_i) \text{ où } \pi_i = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_{57} x_{57,i}}}{1+e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_{57} x_{57,i}}}$$

- Y_i représente l'atteinte de l'objectif annuel de *ventes-retraits* en décembre pour la succursale i
- π_i représente la probabilité d'atteindre l'objectif annuel de *ventes-retraits* en décembre pour la succursale i
- $x_{1,i}, \ldots, x_{57,i}$ représentent différentes mesures prises en juin sur la succursale i (voir l'annexe 2)
- $\boldsymbol{\beta}$ est le vecteur de paramètres du modèle

5.2.2 Sélection de variables et présentation des modèles ajustés

Les variables utilisées pour construire le modèle sont les indicateurs créés à l'objectif 1, les groupes formés à l'objectif 2 et toutes les variables du jeu de données n'ayant pas été utilisées dans la création des indicateurs. Une nouvelle variable a aussi été créée pour l'ajustement de ce modèle. Il s'agit de la différence entre le montant cumulatif de *ventes-retraits* en juin et l'objectif partiel. Elle a été préférée au montant cumulatif de *ventes-retrait* en juin, car elle apporte une information supplémentaire. En effet, cette nouvelle variable permet d'évaluer où se situe une succursale au niveau des *ventes-retraits* par rapport à son objectif partiel.

Étant donné le grand nombre de variables disponibles, une sélection de variables a été effectuée. Pour ce faire, deux démarches ont été utilisées. Premièrement, une sélection de modèle a été faite à l'aide de l'option SELECTION de la procédure LOGISTIC. Les méthodes FORWARD, BACKWARD et STEPWISE ont été utilisées. Les modèles retenus par cette méthode sont les modèles 1 à 4 du tableau 8. Par la suite, la méthode *purposeful selection* proposée par Hosmer et Lemeshow[12] a été utilisée. Le modèle choisi avec cette approche est le modèle 5 du même tableau.

[12] Hosmer, Lemeshow et Sturdivant. Applied Logistic Regression, section 4.2
La sélection a été effectuée avec la macro développée par Zoran Bursac et al. en 2007

La multicolinéarité entre les variables explicatives a été vérifiée à l'aide de l'indice de conditionnement et du facteur d'inflation de la variance (VIF). Ces indices ont été obtenus avec la procédure REG en ajoutant les options VIF et COLLIN. La multicolinéarité est jugée problématique lorsqu'un des VIF est supérieur à 10 ou lorsqu'un des indices de conditionnement est supérieur à 30.

Tableau 8 : Présentation, pour les différents modèles de régression logistique ajustés, des coefficients estimés et des erreurs-type (entre parenthèses) et valeurs de p qui leur sont associées

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
Ordonnée à l'origine	0.814 (1.320) p=0.538	-1.493 (0.354) p<.0001	0.384 (1.297) p=0.767	-2.287 (0.734) p=0.002	-1.837 (1.242) p=0.139
Atteinte de l'objectif en juin		2.103 (0.251) p<0.001	2.257 (0.285) p<0.001	2.242 (0.276) p<0.001	0.974 (0.354) p=0.006
Différence entre les *ventes-retraits* et l'objectif en juin					2.09E-7 (5.60E-8) p<0.001
Valeur moyenne des propriétés financées		3.2E-5 (1.5E-5) p=0.038	7.3E-5 (2.4E-5) p=0.003	5.6E-5 (2.1E-5) p=0.009	2.4E-5 (2.0E-5) p=0.229
Montant moyen de l'épargne locale	-0.00011 (5.0E-5) p=0.034		-0.00012 (4.5E-5) p=0.007		
Montant des ventes virtuelles	3.14E-7 (1.44E-7) p=0.029				
Nombre de clients	-4.0E-5 (1.9E-5) p=0.049				
Nombre moyen d'hypothèques par client					0.309 (5.339) p=0.954
Nombre moyen de concurrents dans l'aire de diffusion des clients					2.777 (2.608) p=0.287
Pourcentage de clients faisant affaire avec les conseillers de l'équipe 1	0.076 (0.027) p=0.006				
Pourcentage de clients détenant un produit d'épargne aisés 2	0.164 (0.056) p=0.003		0.183 (0.063) p=0.004		
Pourcentage de clients détenant un prêt			-0.088 (0.040) p=0.027		

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
Pourcentage de clients détenant un compte d'épargne conventionnel					0.007 (0.025) p=0.787
Pourcentage de clients détenant le produit d'épargne fortuné 1	2.8152 (1.3768) p=0.041				
Pourcentage de clients détenant une carte de crédit de l'institution					0.00972 (0.0219) p=0.657
Score d'épargne 2					-0.1466 (0.1271) p=0.249
Score d'épargne 4	0.8386 (0.2775) p=0.003		0.7013 (0.2574) p=0.006		
Score cycle de vie 3					-0.0414 (0.1886) p=0.826
Score niveau de richesse 2					-0.1393 (0.1787) p=0.436
Sous-secteur*	2, 7, 12		2, 7, 12	2	
Groupe associé à l'objectif 2*	7				

* Les variables «sous-secteur» et «groupes associé à l'objectif 2» sont des variables catégoriques à plusieurs modalités. Les chiffres indiqués dans le tableau correspondent aux modalités dont le coefficient est significativement différent de zéro au seuil 5%. La dernière modalité est prise comme référence.

Le modèle 1 est le seul à inclure les groupes formés à la section 4.1. Dans quatre des cinq modèles, les variables correspondant à l'atteinte de l'objectif en juin et au montant moyen des prêts hypothécaires sont présentes. Il en est de même pour la variable sous-secteur dans trois des cinq modèles. Il est donc permis de croire que ces variables sont davantage associées à l'atteinte de l'objectif d'épargne en décembre que les autres. Enfin, il est intéressant de noter que pour plusieurs des variables du modèle 5, le coefficient n'est pas significativement différent de zéro.

Différentes statistiques concernant les modèles ajustés sont présentées au tableau 9. On y trouve aussi le taux d'erreur de prédiction calculé avec les données de 2013. Tous les modèles choisis présentent un bon ajustement selon la statistique de Hosmer-Lemeshow. Les critères d'information AIC et SC favorisent le modèle 5. Les statistiques reliées à la valeur prédictive du modèle, c'est-à-dire le R^2 adapté à la régression logistique et l'aire sous la courbe ROC (c), incitent plutôt à choisir le modèle 3. Pour ce qui est du taux d'erreur de prédiction associé à 2013, les résultats sont comparables, sauf pour le modèle 1 qui semble moins performant à cet égard.

Tableau 9 : Quelques statistiques concernant l'ajustement et la valeur prédictive des modèles de régression logistique ajustés et taux d'erreur mesuré sur les données de 2013 (les cases ombragées mettent en évidence les meilleurs modèles selon les critères et statistiques)

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
Hosmer-Lemeshow (p)	0.218	0.798	0.844	0.856	0.765
AIC	479.152	408.575	404.580	408.892	397.608
SC	599.188	420.191	493.638	482.462	440.202
R^2 (max-rescaled)	0.233	0.278	0.402	0.369	0.355
C	0.741	0.752	0.819	0.807	0.811
Taux d'erreur 2013 (%)	58.33	21.67	28.33	25.00	22.50

5.2.3 Interprétation du modèle retenu

Le modèle retenu est le modèle 3. Il a été préféré au modèle 5 étant donné qu'il a beaucoup moins de données influentes (voir annexe 3). Comme les données influentes ont un impact important sur l'estimation des paramètres, il est plus prudent d'utiliser le modèle 3. Des graphiques de la probabilité prédite en fonction du logit des variables explicatives ont été tracés et montrent que la fonction de lien est bien choisie.

Comme la variable sous-secteur est une variable catégorique, la dernière modalité a été choisie comme référence. Cependant, ses coefficients s'interprètent comme la comparaison entre le sous-secteur d'intérêt et tous les autres[13]. L'analyse de type III des effets du modèle indique qu'au moins un des seize coefficients associés aux modalités de cette variable est significativement différent de 0 (p=0.022).

Le tableau 10 présente les rapports de cotes associés à chaque coefficient du modèle. Ceux-ci sont plus faciles à interpréter que les coefficients eux-mêmes. Comme les rapports de cotes des sous-secteurs 2, 7 et 12 sont significativement supérieurs à 1, les succursales de ces secteurs ont une plus grande probabilité d'atteindre leur objectif de *ventes-retraits* que les autres, toutes les autres variables étant contrôlées. Le sous-secteur pour lequel cet effet est le plus marqué est le sous-secteur 2.

Pour ce qui est des variables donnant le montant moyen de l'épargne locale et la valeur des propriétés financées, le rapport de cotes est de 1. Cela s'explique par le fait que ces variables prennent de très grandes valeurs et qu'une variation d'une unité a très peu d'influence sur la cote de l'atteinte de l'objectif.

La variable « atteinte de l'objectif en juin » a un rapport de cote de 9.55. Cela signifie qu'une succursale ayant atteint son objectif en juin a 9.55 fois plus de chances d'atteindre son objectif en décembre qu'une

[13] Allison. Logistic Regression Using SAS, p.35

succursale ne l'ayant pas atteint en juin. Cela dit, l'intervalle de confiance pour cette estimation est assez large.

Tableau 10 : Rapports de cotes des coefficients du modèle 3 et intervalles de confiance de Wald à 95%

	Estimation	Intervalle de confiance
Sous-secteur 1	2.087	0.443, 9.836
Sous-secteur 2	42.067	7.403, 239.039
Sous-secteur 3	4.614	0.965, 22.061
Sous-secteur 4	1.577	0.271, 9.179
Sous-secteur 5	1.572	0.351, 7.045
Sous-secteur 6	2.198	0.575, 8.399
Sous-secteur 7	7.937	1.023, 61.560
Sous-secteur 8	2.224	0.457, 10.817
Sous-secteur 9	2.014	0.413, 9.827
Sous-secteur 10	1.723	0.390, 7.615
Sous-secteur 11	2.797	0.515, 15.191
Sous-secteur 12	5.342	1.119, 25.510
Sous-secteur 13	2.140	0.501, 9.142
Sous-secteur 14	2.880	0.627, 1.239
Sous-secteur 15	1.136	0.246, 5.246
Sous-secteur 16	3.362	0.685, 16.509
Montant moyen de l'épargne locale	1.000	1.000, 1.000
Valeur moyenne des propriétés financées	1.000	1.000, 1.000
% de clients qui détiennent un produit aisé 2	1.200	1.061, 1.358
% de clients qui détiennent un prêt	0.916	0.848, 0.990
Score d'épargne 4	2.016	1.217, 3.340
Atteinte de l'objectif en juin	9.550	5.467, 16.683

Enfin, la valeur du rapport de cotes pour la variable « pourcentage de clients détenant un produit d'épargne aisé 2 » est de 1.200. Ainsi, une augmentation de 1 % pour cette variable entraîne que la cote prédite pour l'atteinte de l'objectif annuel augmente de 20 %. Les autres variables s'interprètent de la même manière.

5.2.4 Discussion

Le modèle de régression logistique présenté dans cette section s'ajuste bien aux données et semble avoir une bonne valeur prédictive. De plus, l'interprétation de ses paramètres est cohérente dans le contexte.

Afin de le comparer à la méthode décrite à la section 5.1, le tableau 11 a été construit. La probabilité choisie pour déterminer si la succursale a atteint son objectif est 0.41, soit la proportion de succursales ayant réellement atteint leur objectif en 2013. Le taux d'erreur est de 32 %. La sensibilité est de 60 % et la spécificité est de 80 %. Comparé aux résultats de la section 5.1, la sensibilité a diminué et la spécificité a augmenté. Comme il semble plus approprié d'utiliser une méthode apte à bien prédire les succursales en difficulté, ce modèle est peut-être moins performant que la méthode simple du tableau croisé.

Néanmoins, ce modèle a l'avantage de mettre l'atteinte de l'objectif en relation avec davantage de variables. Il permet donc d'expliquer les facteurs qui influencent l'atteinte de l'objectif, comme le sous-secteur et le score d'épargne 4, par exemple. Qui plus est, le calcul de l'erreur de prévision a été fait avec un seul sous-ensemble de données. Il est donc possible que d'autres sous-ensembles donnent des résultats différents.

Tableau 11 : Tableau de fréquences croisant la prévision de l'atteinte de l'objectif effectuée à l'aide du modèle retenu et l'atteinte de l'objectif en décembre pour les données de 2013, en pourcentage (les cases ombragées représentent le pourcentage de succursales pour lesquelles l'atteinte de l'objectif est mal prédite)

		Réalité		
		Non	**Oui**	
Prédit	**Non**	36	8	**44**
	Oui	24	32	**56**
		60	**40**	**100**

5.3 Prédiction du montant de *ventes-retraits* cumulatives en décembre : modèle de régression linéaire

Dans cette section, une autre méthode pour prédire l'atteinte de l'objectif annuel d'épargne des succursales est détaillée. Il s'agit de prédire le montant de *ventes-retraits* cumulatives en décembre afin de le comparer à l'objectif fixé.

5.3.1 Modèle mathématique

Le modèle est le suivant :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_{57} x_{57,i} + \varepsilon_i$$

- $x_{1,i}, \ldots, x_{57,i}$ représentent différentes mesures prises en juin sur la succursale i (les détails se trouvent à l'annexe 2)
- y_i représente les *ventes-retraits* cumulatives en décembre de la succursale i
- $\boldsymbol{\beta}$ est le vecteur de paramètres du modèle
- $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d.

5.3.2 Sélection de modèle

Les variables explicatives de ce modèle de régression sont les variables sociodémographiques disponibles dans le jeu de données, les scores créés à l'objectif 1, les groupes créés à l'objectif 2a, ainsi que les variables représentant l'atteinte de l'objectif en juin et les *ventes-retraits* cumulatives en juin. La variable dépendante est la variable représentant le montant de *ventes-retraits* cumulatives en décembre. Il s'est avéré préférable d'utiliser cette variable plutôt que celle représentant la différence entre les *ventes-retraits* et l'objectif, puisque cette dernière contient trop de valeurs centrées autour de zéro.

La sélection de variables a été faite à l'aide de la procédure GLMSELECT avec les données de 2013 comme jeu de validation (option VALDATA=). Différentes combinaisons de méthodes de sélection des variables (option METHOD=), de critères d'entrées-sorties des variables (option SELECT=) et de méthodes de sélection du modèle final (option CHOOSE=) ont été essayées. Quatre modèles ont été choisis, principalement parce qu'ils ont été sélectionnés à plusieurs reprises par la procédure. Ils sont présentés dans le tableau 12.

La variable donnant le montant cumulatif de *ventes-retraits* en juin est présente dans tous les modèles. Il est raisonnable de croire que c'est cette variable qui explique le mieux le montant des ventes en décembre. L'atteinte de l'objectif en juin et le montant des ventes virtuelles sont aussi des variables qui semblent importantes, car elles se retrouvent dans trois des quatre modèles.

Tableau 12 : Présentation, pour les différents modèles de régression linéaire ajustés, des coefficients estimés et des erreurs-type (entre parenthèses) et valeurs de p qui leur sont associées

	Modèle 1	Modèle 2	Modèle 3	Modèle 4
Ordonnée à l'origine	811 387 (504 652) p=0.109	4 881 729 (3 177 212) p=0.125	10 253 734 (6 739 533) p=0.129	2 034 675 (616 172) p=0.001
Montant des *ventes-retraits* en juin	1.59 (0.03) p<0.001	1.24 (0.06) p<0.001	1.24 (0.06) p<0.001	1.42 (0.05) p<0.001
Atteinte de l'objectif en juin		-2 193 197 (774 853) p=0.005	-2 307 143 (771 497) p=0.003	-3 350 136 (753 595) p<0.001
Montant des ventes virtuelles		1.27 (0.27) p<0.001	1.23 (0.28) p<0.001	1.52 (0.28) p<0.001
Montant moyen de l'ensemble des prêts		-156 (67) p=0.021	-268 (104) p=0.011	
Valeur moyenne des propriétés financées		170 (60) p=0.005	200 (61) p=0.001	
Nombre moyen de concurrents dans l'aire de diffusion du client		14 840 563 (6 678 753) p=0.027	14 418 064 (6 928 322) p=0.038	
Moyenne de l'ancienneté des clients		-219 949 (128 159) p=0.087		
Score d'épargne 2		-681 601 (310 955) p=0.029	-697 718 (310516) p=0.025	
% des clients dont l'institution principale est l'institution			-160 831 (93 049) p=0.085	
% de clients qui détiennent une carte de crédit de l'institution			92 201 (70 948) p=0.195	
Nombre de clients entreprises		3 534 (850) p<0.001	3 700 (876) p<0.001	

Le **Erreur ! Référence non valide pour un signet.** montre que la statistique R^2 ajusté favorise le modèle 2. Les critères d'ajustement AIC et BIC favorisent les modèles 2 et 3, et le SBC, les modèles 2 et 4. Ces statistiques sont toutefois très similaires pour les quatre modèles. Le C_p de Mallows favorise le modèle 1. Les statistiques SSE et ASE(Validate) représentent des calculs d'erreur de prédiction du modèle[14]. La première concerne le modèle ajusté sur les données de 2014 et la seconde, les données de 2013. La statistique SSE favorise les modèles 2 et 3 et la statistique ASE(Validate), le modèle 1.

Tableau 13 : Quelques statistiques concernant l'ajustement et la valeur prédictive des modèles de régression linéaire ajustés (les cases ombragées mettent en évidence les meilleurs modèles selon les critères et statistiques)

	Modèle 1	Modèle 2	Modèle 3	Modèle 4
R2 ajusté	0.8621	0.8920	0.8916	0.8836
AIC	11 204	11 125	11 128	11 146
BIC	11 206	11 128	11 130	11 148
SBC	11 212	11 164	11 170	11 161
C_p	2	10	11	4
SSE	179E14	137E14	137E14	150E14
ASE(Validate)	1.14E14	2.50E14	2.35E14	3.60E14

La multicolinéarité a été évaluée comme à la section 5.2. Le modèle 1 ne présente évidemment aucun problème de ce côté puisqu'il n'a qu'une seule variable explicative. Les modèles 2 et 4 ne sont pas problématiques non plus. Selon les diagnostics de multicolinéarité du modèle 3, il y a un problème entre la variable mesurant la proportion de clients qui font affaire principalement avec l'institution, celle mesurant la proportion de clients qui ont une carte de crédit de l'institution et celle mesurant le montant moyen de prêts autorisés. Il est donc déconseillé d'interpréter les paramètres de ce modèle.

Le choix du modèle dépend de l'utilisation qu'on voudra en faire. Pour étudier la relation entre les caractéristiques d'une succursale et ses ventes en décembre, il serait judicieux d'analyser le modèle 2. Si l'objectif est de prédire les *ventes-retraits* en décembre, les modèles 1 ou 2 sont appropriés. Ces deux modèles sont détaillés dans la prochaine section.

[14]SAS Institute, SAS/STAT® 9.2 User's Guide, PROC GLMSELECT - Details - Criteria used in Model Selection Methods, p. 2727

5.3.3 Modèles de régression choisis

Les postulats des modèles choisis ne sont pas tous respectés. Le test de Shapiro rejette la normalité (les seuils observés sont inférieures à 0.0001 pour les deux modèles). De plus, les graphes montrant les résidus des modèles en fonction des valeurs prédites présentés à l'annexe 4 montrent que le postulat d'homogénéité de la variance n'est pas respecté. Il est difficile de trouver une transformation permettant d'améliorer ce postulat, car la variable dépendante prend des valeurs négatives. Les deux postulats sont peut-être violés en raison de la présence de données aberrantes. En effet, le modèle 1 contient 11.27 % de données aberrantes et le modèle 2, 9.30 %. Pour déterminer si une donnée est aberrante, son résidu est analysé. S'il est grand[15], la donnée est déclarée aberrante. La procédure ROBUSTREG calcule le pourcentage de données aberrantes.

Pour avoir des estimations plus robustes aux valeurs aberrantes, les paramètres ont été estimés à l'aide de la méthode M. Cette méthode est implémentée par défaut dans la procédure ROBUSTREG. Différentes fonctions de poids ont été testées (option WF=) et la fonction *fair* est celle qui donnait les meilleurs résultats pour les deux modèles selon l'AICR, le BICR et le R^2. Les tableaux Tableau **14** et Tableau 15 donnent les estimations robustes des paramètres pour les deux modèles choisis.

Tableau 14 : Coefficients estimés par la méthode M, erreurs-type de valeurs de p pour le modèle 1

	Estimation	Erreur-type	Valeur de p
Ordonnée à l'origine	566 269	275 708	0.0400
***Ventes-retraits* en juin**	1.53	0.0184	<0.0001

Pour ce qui est du modèle 1, des statistiques d'ajustement ont été calculées avec les nouveaux paramètres. Le R^2 de ce modèle est maintenant de 0.769, l'AICR de 417 et le BICR de 426. Les deux paramètres sont significativement supérieurs à zéro au seuil 5%. Ainsi, plus les *ventes-retraits* cumulatives en juin sont élevées, plus les *ventes-retraits* cumulatives en décembre le sont aussi.

Les mêmes statistiques d'ajustement ont été calculées avec les nouveaux paramètres du modèle 2. Le R^2 de ce modèle est maintenant de 0.800, l'AICR de 378 et le BICR de 422. Ainsi, toutes les statistiques d'ajustement privilégient ce modèle au précédent.

Le tableau 15 indique qu'au seuil 5 %, l'ordonnée à l'origine, le paramètre associé au montant moyen des prêts autorisés, ainsi que celui associé à la moyenne de l'ancienneté des clients de la succursale ne sont pas significativement différents de zéro. Les coefficients devant les variables ventes en juin, les ventes virtuelles, la valeur moyenne des propriétés financées, la moyenne des concurrents dans l'aire de diffusion du client et le nombre de clients entreprises indiquent que les relations entre ces variables et

[15]SAS Institute, SAS/STAT® 9.2 User's Guide, PROC ROBUSTREG - Details - Leverage Point and Outlier Detection p.5689

les *ventes-retraits* cumulatives en décembre sont positives. Pour leur part, les coefficients devant les variables donnant l'atteinte de l'objectif en juin, le montant moyen des prêts autorisés, la moyenne de l'ancienneté des clients et le score d'épargne 2 indiquent que les relations entre ces variables et les *ventes-retraits* cumulatives en décembre sont négatives.

Tableau 15 : Coefficients estimés par la méthode M, erreurs-type et valeurs de p pour le modèle 2

	Estimation	Erreur-type	Valeur de p
Ordonnée à l'origine	3 088 161	1 926 761	0.109
***Ventes-retraits* en juin**	1.21	0.0386	<0.001
Atteinte de l'objectif en juin	-1 669 207	469 895	<0.001
Ventes virtuelles	1.21	0.166	<0.001
Montant moyen de l'ensemble des prêts	-79	40	0.052
Valeur moyenne des propriétés financées	97	36	0.008
Nombre moyen de concurrents dans l'AD du client	9 526 814	4 050 205	0.019
Moyenne de l'ancienneté des clients	-142 284	77 720	0.067
Score d'épargne 2	-609 602	188 573	0.001
Nombre de clients entreprises	3 293	516	<0.001

Il est très surprenant que la relation entre l'atteinte de l'objectif en juin et les *ventes-retraits* en décembre soit négative. En effet, cela veut dire que, toutes choses étant égales par ailleurs, une succursale ayant atteint son objectif en juin a un montant de *ventes-retraits* cumulatives en décembre plus faible qu'une succursale n'ayant pas atteint son objectif. Ceci semble indiquer qu'il faut faire preuve de prudence en interprétant ce modèle.

5.3.4 Discussion

Pour calculer le taux d'erreur de prédiction de la variable représentant l'atteinte de l'objectif en décembre avec ces deux modèles, les montants de *ventes-retraits* prédits et l'objectif ont été comparés. Si le montant prédit est supérieur ou égal à l'objectif, il est prédit que la succursale atteindra son objectif. Au contraire, si le montant de *ventes-retraits* prédit est inférieur à l'objectif fixé, il est prédit que la succursale n'atteindra pas son objectif. Le tableau 16 indique que 26 % des succursales sont mal prédites avec le modèle 1 et que 55 % le sont avec le modèle 2.

Tableau 16 : Tableau de fréquences croisant la prévision de l'atteinte de l'objectif effectuée à l'aide des deux modèles de régression linéaires retenus et l'atteinte de l'objectif en décembre pour les données de 2013, en pourcentage (les cases ombragées représentent le pourcentage de succursales mal prédites avec ces modèles)

			Réalité		
			Non	Oui	
Modèle 1	Prédit	Non	52	18	**70**
		Oui	8	23	**30**
Modèle 2	Prédit	Non	9	4	**13**
		Oui	51	36	**87**
			60	**40**	

La sensibilité du modèle 1 est de 87 %, alors que sa spécificité est de 58 %. Ces nombres sont respectivement 15 % et 90 % pour le modèle 2. Il n'est pas recommandé d'utiliser le modèle 2 comme modèle de prévision, parce que son taux d'erreur est très grand, mais aussi parce que sa sensibilité est très faible. Dans le contexte d'intervention auprès des succursales en difficulté, il est préférable de choisir un modèle qui minimise les faux négatifs (déclarer qu'une succursale atteindra son objectif alors qu'elle ne l'atteint pas) plutôt que les faux positifs (déclarer qu'une succursale n'atteindra pas son objectif alors qu'elle l'atteint). En effet, il est moins dramatique d'intervenir auprès de succursales qui atteindront leur objectif que de ne pas intervenir auprès de succursales qui ne l'atteindront pas.

Il n'est pas surprenant que le modèle 1 prédise mieux l'atteinte de l'objectif avec les données de 2013, puisque la statistique ASE(Validate) du tableau 13 indique que ce modèle minimise les erreurs de prédiction des *ventes-retraits* cumulatives de décembre 2013. Toutefois, selon la statistique SSE du même tableau, le modèle 2 minimise les erreurs de prédiction des *ventes-retraits* cumulatives de décembre 2014. Ce problème embêtant provient sûrement de la grande variabilité dans les données d'une année à l'autre et suggère d'utiliser les modèles de régression avec prudence.

5.4 Discussion sur l'objectif

À la lumière des résultats obtenus, il apparaît que les différentes approches ont chacune leurs avantages et leurs inconvénients. Le tableau de fréquence est simple à utiliser et semble donner de bons résultats, mais il ne donne aucune information quant aux variables associées à l'atteinte de l'objectif. Pour sa part, le modèle de régression logistique permet d'associer une probabilité d'atteinte de l'objectif pour chaque succursale et donne une idée des variables associées à la réalisation de cet évènement. Cependant, il ne prédit pas l'atteinte de l'objectif en 2013 aussi efficacement. Enfin, le modèle de régression linéaire propose une autre approche. Par contre, certains de ses coefficients ont des signes contre-intuitifs.

Il est intéressant de noter que la proportion de succursales ayant atteint leur objectif a grandement augmenté entre 2013 et 2014, passant de 41 % à 57 %. Il est possible que cette grande variation ait eu un impact sur la qualité des prévisions effectuées avec les modèles. Il serait intéressant d'évaluer la performance de ces modèles avec les données de 2015 si la proportion de succursales ayant atteint leur objectif en décembre est alors semblable à celle observée en 2014.

6 Conclusion

L'objectif de ce projet était, d'une part, de créer un outil de classification permettant de regrouper les succursales dont les caractéristiques en lien avec l'épargne sont similaires et, d'autre part, de prédire l'atteinte d'un objectif d'épargne annuel. Pour répondre adéquatement à ces objectifs, il a été nécessaire de procéder en plusieurs étapes, chacune d'entre elles appelant à des procédures et analyses statistiques différentes.

Tout d'abord, il a fallu réduire la taille du jeu de données afin de conserver seulement les variables les plus pertinentes pour les analyses. Une fois réduit, le jeu de données contenait encore une centaine de variables et des problèmes de multicolinéarité étaient présents. Des analyses factorielles ont donc été effectuées afin de résumer l'information de certains groupes de variables en quelques scores indépendants. Ces analyses ont été faites sur un groupe de variables d'épargne et sur des groupes de *variables modalités* à caractère sociodémographique. Les scores d'épargne obtenus ont été utilisés dans la suite du projet. Du côté des différents groupes de *variables modalités*, les scores obtenus étaient interprétables et très intéressants pour être utilisés comme indicateurs résumant l'information sociodémographique. Par contre, ils ont été moins utilisés dans ce projet.

L'idée d'effectuer des analyses factorielles sur les *variables modalités* est nouvelle et pose peut-être des problèmes théoriques insoupçonnés. Il serait intéressant de l'étudier davantage dans l'avenir. Cela dit, les résultats obtenus sont convaincants et permettent de conserver l'information contenue dans les différentes variables tout en éliminant la dépendance entre celles-ci.

Pour ce qui est de la création des groupes de succursales, elle a été faite par classification non supervisée. Suite à plusieurs essais sur différents sous-ensembles de variables et avec différentes méthodes de classification, les scores d'épargne et quatre variables décrivant l'épargne ont été utilisés. Ce sous-ensemble permet d'obtenir des groupes satisfaisant les critères suggérés par les conseillères du projet. Les groupes formés présentent des caractéristiques intéressantes et permettent d'établir le profil d'épargne des succursales. Cependant, ils ne semblent pas stables dans le temps. On ignore dans quelle mesure cela est dû au fait que, d'un temps à l'autre, le profil des succursales peut changer et entraîner une classification différente. Pour sa part, l'analyse discriminante permet d'assigner un groupe à une nouvelle succursale. Elle peut être utilisée afin d'avoir une idée de son profil d'épargne.

Finalement, diverses approches permettant de prédire l'atteinte de l'objectif annuel de *ventes-retraits* des succursales ont été explorées. Une méthode intuitive utilisant un tableau de fréquences croisées a été présentée, ainsi qu'un modèle de régression logistique et un modèle de régression linéaire. Les modèles ont été ajustés sur les données de 2014 et leur valeur prédictive a été évaluée selon leur capacité à prédire l'atteinte de l'objectif sur les données de 2013. Les résultats sont satisfaisants, mais le tableau de fréquences donne de meilleurs résultats que les modèles. Cependant, ces derniers apportent des informations supplémentaires quant aux variables qui expliquent en partie l'atteinte de l'objectif. Il est à noter que les groupes formés à l'objectif 2 ne sont pas présents parmi les variables explicatives de ces modèles. Il se peut que le groupe d'appartenance d'une succursale ne soit pas associé à l'atteinte de son objectif de manière significative. Il serait aussi intéressant d'essayer de développer un modèle prédictif à l'aide de séries chronologiques. Peut-être que cette méthode donnerait des résultats intéressants.

En définitive, ce projet a permis la création d'un outil de classification des succursales d'une institution financière qu'il serait intéressant de raffiner. De plus, plusieurs avenues possibles ont été explorées pour la prévision de l'atteinte d'un objectif annuel d'épargne.

Bibliographie

Ressources électroniques

HEC Montréal. *Site du HEC Montréal*, [en ligne]. http ://www.hec.ca/ (Page consultée le 10 décembre 2014)

RBC royale. *Site de la RBC royale*, [en ligne]. http ://www.rbcbanqueroyale.com/ (Page consultée le 10 décembre 2014)

Bursac, Gauss, Williams et Hosmer. «Purposeful selection of variables in logistic regression». *Source Code for Biology and Medicine*, [en ligne]. http://www.scfbm.org/content/3/1/17 (Page consultée le 21 avril 2015)

Ouvrages

Allison, Paul D. *Logistic Regression Using SAS : Theory and Application*. Second Edition. Cary, Caroline du Nord, É-U : SAS Institute Inc, 2012, 324 pages.

Hosmer, Lemeshow et Sturdivant. *Applied Logistic Regression*. Third Edition. Hobokoken, New Jersey, É-U : John Wiley & Sons, 2013, 500 pages.

SAS Institute Inc. *SAS/STAT® 9.2 User's Guide*. Second Edition. Cary, Caroline du Nord, É-U : SAS Institute Inc, 2009, 7620 pages.

Annexe 1 : Variables du jeu de données réduit et description

Légende
nb : nombre
pc : %
mm : montant moyen
m : moyenne
* : Somme des colonnes ramenée à 100
** : Variable brute multipliée par 100

Catégorie	Nom	Description	Note
1. Temps	Temps	Représente les moments où les observations ont été effectuées	
2. Info succursale	Id		
2. Info succursale	sous_secteur		
2. Info succursale	nb_client	Nombre de clients de la succursale	
2. Info succursale	pc_eq1	Pourcentage de clients faisant affaire avec un conseiller de l'équipe 1	* **
2. Info succursale	pc_eq2	% équipe 2	* **
2. Info succursale	pc_eq3	% équipe 3	* **
2. Info succursale	nb_concur	Nombre de concurrents dans l'AD de la succursale	
2. Info succursale	nb_entr	Nombre de clients de la succursale qui sont des entreprises	
2. Info succursale	pc_entr	Pourcentage de clients qui sont des entreprises	**
2. Info succursale	mnt_vv	Ventes virtuelles nettes	
3. Épargne mnt moyen	mm_voaf	Montant moyen du volume d'affaire	
3. Épargne mnt moyen	mm_epg_loc	Montant moyen d'épargne détenue chez l'institution	
3. Épargne mnt moyen	mm_epg_est	Montant moyen d'épargne estimée	
3. Épargne mnt moyen	mm_XX	Montant moyen de produit d'épargne lié au marché 2	
3. Épargne mnt moyen	mm_XX	Montant moyen dans compte d'épargne conventionnel	
3. Épargne mnt moyen	mm_XX	Montant moyen de produit d'épargne lié au marché 1	
3. Épargne mnt moyen	mm_XX	Montant moyen de produit d'épargne aisé 2	
3. Épargne mnt moyen	mm_XX	Montant moyen de produit d'épargne aisé 1	

3. Épargne mnt moyen	mm_XX	Montant moyen de produit VM – Plein Exercice	
3. Épargne mnt moyen	mm_XX	Montant moyen de produit fortuné 1	
3. Épargne mnt moyen	mm_XX	Montant moyen de produit fortuné 2	
3. Épargne mnt moyen	mm_XX	Montant moyen de produit valeurs mobilières	
3. Épargne mnt moyen	mm_XX	Montant moyen de certificat de placement garanti (CPG)	
4. Prêt mnt moyen	mm_pr_auto	Montant autorisé moyen de l'ensemble des prêts	
4. Prêt mnt moyen	mm_pr_tot	Solde total moyen des prêts	
4. Prêt mnt moyen	mm_hyp	Valeur moyenne des propriétés financées	
3. Épargne % clients	pc_XX	Pourcentage de clients ayant un certificat de placement garanti (CPG)	**
3. Épargne % clients	pc_XX	Pourcentage de clients ayant un produit d'épargne lié au marché 1	**
3. Épargne % clients	pc_XX	Pourcentage de clients ayant un produit d'épargne aisé 2	**
3. Épargne % clients	pc_XX	Pourcentage de clients ayant un produit d'épargne aisé 1	**
3. Épargne % clients	pc_XX	Pourcentage de clients ayant un produit VM – Plein exercice	**
3. Épargne % clients	pc_XX	Pourcentage de clients ayant un produit d'épargne fortuné 1	**
3. Épargne % clients	pc_XX	Pourcentage de clients ayant un produit d'épargne lié au marché 2	**
3. Épargne % clients	pc_XX	Pourcentage de clients ayant un compte d'épargne conventionnel	**
3. Épargne % clients	pc_XX	Pourcentage de clients ayant un produit VM	**
3. Épargne % clients	pc_XX	Pourcentage de clients ayant un produit d'épargne fortuné 2	**
4. Prêt % clients	pc_pr	Pourcentage de clients ayant des prêts	**
4. Prêt % clients	pc_pr_auto	Pourcentage de clients ayant des prêts autorisés	**
5. Info clients particuliers	m_concur	Moyenne du nombre d'institutions concurrentes dans l'AD des clients	
5. Info clients particuliers	pc_ad	Pourcentage de clients de la succursale dont l'aire de diffusion est la celle de la succursale	**
5. Info clients particuliers	m_ancien	Moyenne du nombre d'années d'ancienneté des clients de la succursale	
5. Info clients particuliers	m_age	Age moyen des clients de la succursale	
5. Info clients particuliers	pc_cv0	Pourcentage des clients ayant moins de 18 ans	* **

5. Info clients particuliers	pc_cv1	% Étudiants	* **
5. Info clients particuliers	pc_cv2	% Jeunes travailleurs	* **
5. Info clients particuliers	pc_cv3	% Accédant propriété	* **
5. Info clients particuliers	pc_cv4	% Nouveaux propriétaires	* **
5. Info clients particuliers	pc_cv5	% Projets divers	* **
5. Info clients particuliers	pc_cv6	% Préparation retraite	* **
5. Info clients particuliers	pc_cv8	% Retraités	* **
5. Info clients particuliers	pc_demu	Pourcentage de clients démunis	**
5. Info clients particuliers	pc_rich0	Pourcentage des clients dont le niveau de richesse est *moins de 18 ans*	* **
5. Info clients particuliers	pc_rich1	% Utilisateurs	* **
5. Info clients particuliers	pc_rich2	% Bâtisseurs	* **
5. Info clients particuliers	pc_rich3	% Accumulateurs	* **
5. Info clients particuliers	pc_rich41	% Aisé Conventionnel	* **
5. Info clients particuliers	pc_rich42	% Aisé Investisseur	* **
5. Info clients particuliers	pc_rich5	% Fortunés	* **
5. Info clients particuliers	pc_h	Pourcentage des clients qui sont des hommes	* **
5. Info clients particuliers	pc_fr	Pourcentage des clients qui sont francophones	* **
5. Info clients particuliers	pc_ep	Pourcentage des clients ayant un emploi permanent	**
5. Info clients particuliers	nb_cc	Nombre de communautés culturelles	
5. Info clients particuliers	pc_princ	Proportion des clients dont l'institution principale est l'institution	**
5. Info clients particuliers	pc_XX	Pourcentage des clients ayant adhéré au service en ligne	**
5. Info clients particuliers	pc_XX	Pourcentage de clients qui détiennent une carte de crédit de l'institution	**
5. Info clients particuliers	m_rvbr	Revenu brut moyen des clients	
5. Info clients particuliers	m_hyp	Nombre moyen d'hypothèques détenues	
5. Info clients particuliers	m_lprod	Nombre moyen de lignes de produits utilisés par les clients	
6. Info clients entr.	m_caff_entr	Chiffre d'affaire estimé moyen des clients-entreprise	
6. Info clients entr.	nb_XX	Nombre de clients-entreprises qui utilisent les services d'aide aux entreprises	

Annexe 2 : Variables explicatives utilisées dans les régressions

$x_{1,i}$	représente la province où se trouve la succursale i (2 modalités)
$x_{2,i}$	représente la région où se trouve la succursale i (17 modalités)
$x_{3,i}$	représente le groupe d'appartenance de la succursale i (8 modalités)
$x_{4,i}$	représente le nombre de clients de la succursale i
$x_{5,i}$	représente le % de clients de la succursale i qui font affaire avec les conseillers de l'équipe 1
$x_{6,i}$	représente le % de clients de la succursale i qui font affaire avec les conseillers de l'équipe 2
$x_{7,i}$	représente le nombre de concurrents dans l'aire de diffusion de la succursale i (3 modalités)
$x_{8,i}$	représente le nombre de clients-entreprises de la succursale i
$x_{9,i}$	représente le % de clients-entreprises dans la succursale i
$x_{10,i}$	représente le montant des ventes virtuelles de la succursale i
$x_{11,i}$	représente le volume d'affaire de la succursale i
$x_{12,i}$	représente le montant d'épargne moyen investis par les clients dans la succursale i
$x_{13,i}$	représente le montant (estimé) d'épargne moyen investis par les clients hors de la succursale i
$x_{14,i}$	représente le montant moyen de l'ensemble des prêts autorisés de la succursale i
$x_{15,i}$	représente le montant moyen de l'ensemble des prêts de la succursale i
$x_{16,i}$	représente le montant moyen de l'ensemble des prêts hypothécaires de la succursale i
$x_{17,i}$	représente le pourcentage de clients de la succursale i qui ont un produit Valeurs mobilières
$x_{18,i}$	représente le pourcentage de clients de la succursale i qui ont un produit VM - Plein exercice
$x_{19,i}$	représente le pourcentage de clients de la succursale i qui ont un produit Produit d'épargne aisé 1
$x_{20,i}$	représente le pourcentage de clients de la succursale i qui ont un produit Produit d'épargne aisé 2
$x_{21,i}$	représente le pourcentage de clients de la succursale i qui ont un produit Produit d'épargne fortuné 1

$x_{22,i}$ représente le pourcentage de clients de la succursale i qui ont un produit Produit d'épargne fortuné 2

$x_{23,i}$ représente le pourcentage de clients de la succursale i qui ont un produit Produit d'épargne lié au marché 1

$x_{24,i}$ représente le pourcentage de clients de la succursale i qui ont un produit Produit d'épargne lié au marché 2

$x_{25,i}$ représente le pourcentage de clients de la succursale i qui ont un compte d'épargne conventionnel

$x_{26,i}$ représente le pourcentage de clients de la succursale i qui ont un certificat de placement garanti

$x_{27,i}$ représente le pourcentage de clients de la succursale i qui ont un prêt

$x_{28,i}$ représente le pourcentage de clients de la succursale i qui ont un prêt autorisé

$x_{29,i}$ représente la moyenne de concurrents dans l'aire de diffusion du client de la succursale i

$x_{30,i}$ représente le % de clients de la succursale i qui habitent dans l'aire de diffusion de la succursale

$x_{31,i}$ représente la moyenne de l'ancienneté des clients de la succursale i

$x_{32,i}$ représente l'âge moyen des clients de la succursale i

$x_{33,i}$ représente le % de clients démunis dans la succursale i

$x_{34,i}$ représente le % d'hommes dans la succursale i

$x_{35,i}$ représente le % de clients francophones dans la succursale i

$x_{36,i}$ représente le % de clients qui ont un emploi permanent dans la succursale i

$x_{37,i}$ représente le nombre de communautés culturelles dont proviennent les clients de la succursale i

$x_{38,i}$ représente le % de clients dont l'institution est leur institution principale dans la succursale i

$x_{39,i}$ représente le % de clients qui utilisent les services en ligne dans la succursale i

$x_{40,i}$ représente le % de clients qui utilisent la carte de crédit de l'institution dans la succursale i

$x_{41,i}$ représente la moyenne de revenu brut des clients de la succursale i

$x_{42,i}$ représente la moyenne du nombre d'hypothèques des clients de la succursale i

$x_{43,i}$ représente la moyenne du nombre de lignes de produits utilisées par les clients de la succursale i

$x_{44,i}$ représente la moyenne du chiffre d'affaire des clients-entreprises de la succursale i

$x_{45,i}$ représente le nombre de clients-entreprises de la succursale i qui utilisent les services d'aide aux entreprises

$x_{46,i}$ représente le score d'épargne 1 créé à l'objectif 1 de la succursale i

$x_{47,i}$ représente le score d'épargne 2 créé à l'objectif 1 de la succursale i

$x_{48,i}$ représente le score d'épargne 3 créé à l'objectif 1 de la succursale i

$x_{49,i}$ représente le score d'épargne 4 créé à l'objectif 1 de la succursale i

$x_{50,i}$ représente le score du niveau de richesse 1 créé à l'objectif 1 de la succursale i

$x_{51,i}$ représente le score du niveau de richesse 2 créé à l'objectif 1 de la succursale i

$x_{52,i}$ représente le score du cycle de vie 1 créé à l'objectif 1 de la succursale i

$x_{53,i}$ représente le score du cycle de vie 2 créé à l'objectif 1 de la succursale i

$x_{54,i}$ représente le score du cycle de vie 3 créé à l'objectif 1 de la succursale i

$x_{55,i}$ représente les ventes-retraits cumulatives en juin de la succursale i (régression linéaire seulement)

$x_{56,i}$ représente l'atteinte ou non de l'objectif de juin de la succursale i

$x_{57,i}$ représente la différence entre le montant des ventes-retraits cumulatives en juin et le montant prévu par l'objectif partiel en juin de la succursale i (régression logistique seulement)

Annexe 3 : Comparaison des modèles de régression logistique 3 et 5 du point de vue des données influentes

Hosmer et Lemeshow recommandent d'étudier le graphique de la statistique d'influence de la déviance afin de déceler un problème de données influentes dans l'ajustement d'un modèle de régression logistique[16]. Les données problématiques sont celles ayant une différence élevée et qui sont éloignées des autres sur le graphique.

La figure 3 montre que l'observation 200 a une très grande influence dans l'ajustement du modèle 5. En essayant de retirer cette observation, d'autres données sont problématiques.

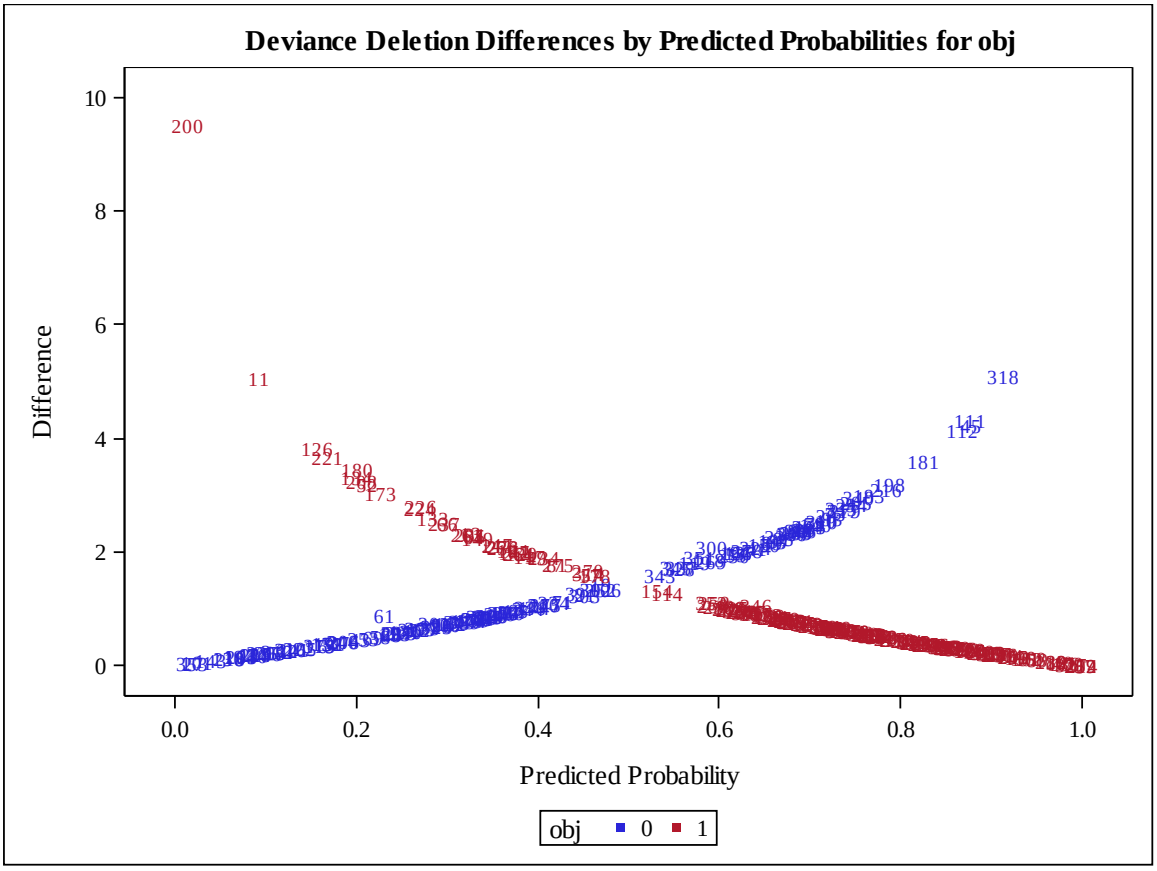

Figure 3 : Graphique de la statistique d'influence de la déviance en fonction des valeurs prédites pour le modèle 5

[16] Alisson, Logistic Regression Using SAS, p.84

La figure 4 montre que dans le cas du modèle 3, la présence de données influentes est nettement moins importante. C'est pourquoi ce modèle a été préféré au modèle 5.

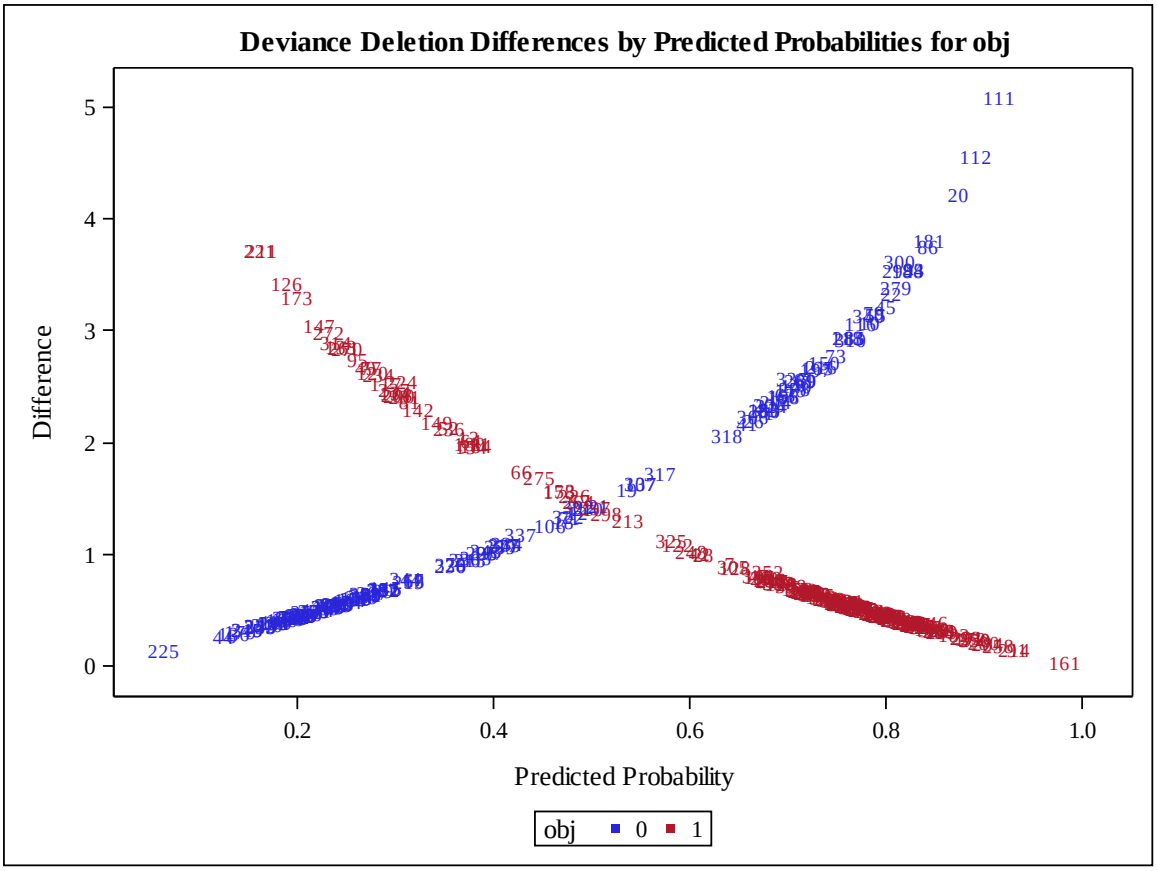

Figure 4 : Graphique de la statistique d'influence de la déviance en fonction des valeurs prédites pour le modèle 3

Annexe 4 : Vérification du postulat d'homogénéité de la variance des modèles de régression

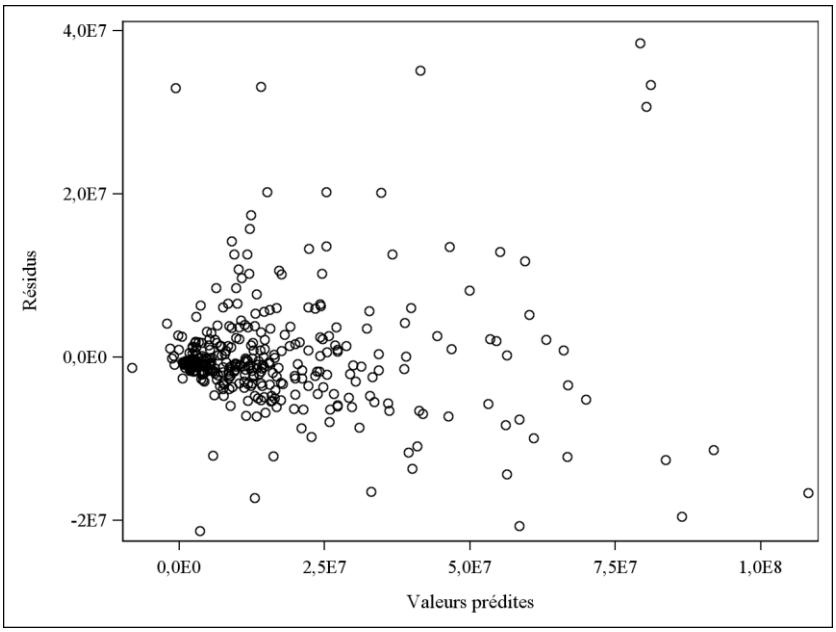

Figure 5 : Résidus en fonction des valeurs prédites pour le modèle 1

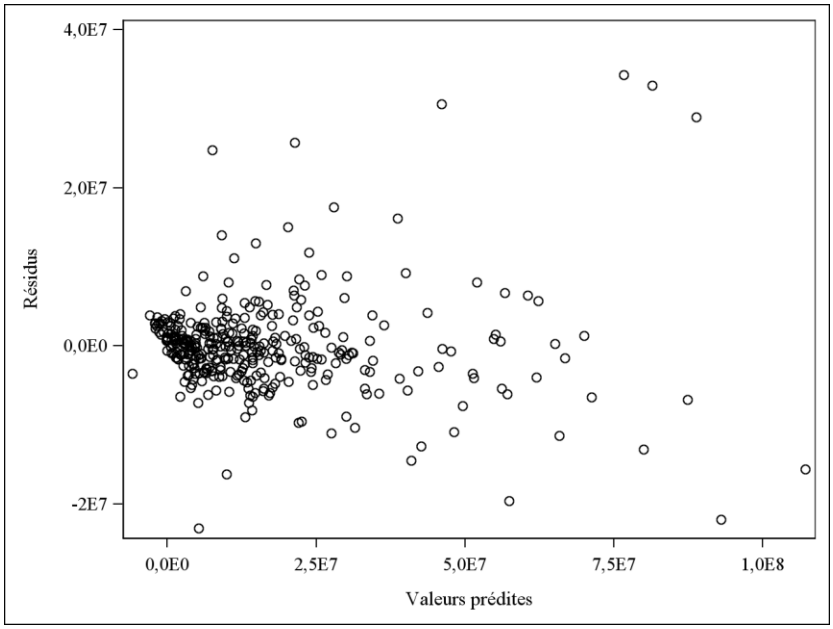

Figure 6 : Résidus en fonction des valeurs prédites pour le modèle 2