Data mining vs RFM en marketing: une étude comparative dans le secteur du e-commerce

PHILIPPE ST-AUBIN[1,2], BRUNO AGARD[1,2]

[1] École Polytechnique de Montréal
Département de mathématiques et génie industriel, CP 6079, succursale Centre-Ville, Montréal, Québec, Canada
philippe.st-aubin@polymtl.ca, bruno.agard@polymtl.ca

[2] Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport (CIRRELT)

***Résumé* – Cet article présente une étude comparative d'un cas d'application d'un modèle RFM (Récence, Fréquence, Montant) et de trois techniques de data mining (clustering, règles d'association et arbre de classification) dans le contexte d'une étude de marché pour une entreprise de e-commerce. Les résultats ont montré que le modèle RFM a été en mesure de classer les clients selon leur propension à revenir faire des achats. Le data mining a, pour sa part, pu identifier des comportements d'achat différents, des liens entre les produits achetés et a pu permettre l'élaboration d'un modèle pour savoir comment prédire le retour d'un client. La comparaison a démontré que le modèle RFM permet d'identifier les meilleurs clients, mais que les résultats apportés par le data mining permettent de prédire, mais également de décrire les clients et leur comportement.**

***Abstract* - This article presents a comparative study of an application of an RFM (Recency, Frequency, Monetary) model and three data mining techniques (clustering, association rules and classification tree) used in the context of a market study for an e-commerce company. The results showed that the RFM model was able to rank customers according to their propensity to return to make purchases. The data mining was able to identify different purchasing behaviors, links between the products purchased and allowed to develop a model to know how to predict the return of a customer. The comparison showed that RFM allows to identify the best customers, but that the results provided by the data mining allow to predict, but also to describe the customers and their behavior.**

***Mots clés* – RFM, data mining, e-commerce, étude de marché.**
***Keywords* – RFM, data mining, e-commerce, market research.**

1 INTRODUCTION

Depuis quelques années, de plus en plus d'entreprises utilisent le data mining afin d'être en mesure de mieux comprendre les caractéristiques et les comportements de leur clientèle [Ngai et al., 2009]. L'utilisation de cette information peut résulter en la génération de processus nouveaux menant à une relation à long terme avec les clients [Ling et Yen, 2001].

Dans le présent article, une étude comparative sur l'utilisation d'un modèle RFM (Récence, Fréquence, Montant) et des techniques de data mining est proposée pour mener une étude de marché. L'idée de comparer les outils du data mining au modèle RFM provient d'une demande d'une entreprise à la recherche d'experts en modèle RFM afin de proposer des moyens d'étudier leur clientèle. Il semblant pertinent de montrer les possibilités offertes par les deux méthodes. L'article propose donc une comparaison des résultats d'une application des deux méthodes d'analyse sur les données de l'entreprise Tuango. L'entreprise œuvre dans le domaine du e-commerce et se spécialise dans la vente d'offres de rabais de divers produits, services et activités. Les données à disposition comptent les informations de toutes les transactions ayant eu lieu sur une période de 6 ans.

Afin d'assurer la confidentialité des données du partenaire, toutes les données financières sont absentes, l'important étant la démarche proposée.

La structure de l'article se détaille comme suit : La section 2 présente, dans l'état de l'art, une introduction aux méthodes d'analyse de données : pour une part la méthode RFM (dédiée aux études marketing), et pour une autre part le data mining. La section 3 montre le contexte du cas d'étude utilisé. La section 4 propose une comparaison d'analyses commerciales effectuées (1) avec la méthode RFM et (2) à partir de trois techniques de data mining. La section 5 présente les conclusions des résultats obtenus et les avantages de l'utilisation du data mining dans un tel contexte.

2 ÉTAT DE L'ART

2.1 Étude et analyse de marché

D'après le Réseau Entreprises Canada (2016), « Une étude de marché consiste à recueillir des renseignements qui vous aideront à mieux comprendre comment les personnes auxquelles vous espérez vendre vos produits réagiront à vos produits et services actuels et éventuels ». Toujours d'après le Réseau Entreprises Canada, une étude de marché cherche à optimiser les prises de décisions au niveau de quatre facteurs :

1. L'offre de produit : en comprenant mieux les clients, il est plus facile de développer des produits répondant à leurs besoins.
2. Les prix : l'étude du marché peut permettre d'identifier le prix que les clients sont prêts à payer, elle permet de comparer la tarification des produits à celle des concurrents. L'étude permet également d'optimiser les revenus en étudiant les marges sur les revenus.
3. La distribution : l'étude géographique du territoire

desservi permet de développer les territoires où l'offre de service est manquante. Elle permet également d'établir la valeur d'un emplacement afin de choisir les meilleurs points de distribution possible.

4. Les offres promotionnelles : une meilleure connaissance des clients permet d'identifier le meilleur canal de communication possible pour les rejoindre et les inciter à faire affaires ou continuer à faire affaires avec l'entreprise.

Il y a normalement deux types d'étude de marché. Le premier type d'étude, l'étude primaire cherche à recueillir de l'information afin de répondre à des questions de base sur le marché cible. Par exemple, qui sont les clients et quels sont les moyens pour les contacter, quels produits les intéressent et qui sont les concurrents ? La réponse aux questions est habituellement obtenue au moyen de sondages, d'expérimentations ou encore d'observations [Kotler et al., 2011]. Ainsi l'étude primaire est habituellement réalisée avant que de nouveaux produits ou services ne soient développés. L'étude sert à savoir si l'innovation ou le développement dans le domaine visé sera rentable.

Le second type d'étude, l'étude secondaire se base sur l'information que l'entreprise a déjà en sa possession. Cette information peut provenir de données de facturation, de rapports de service à la clientèle ou de toutes autres sources pouvant potentiellement répondre à une question de l'entreprise vis-à-vis son marché. L'étude secondaire complète l'étude primaire en confirmant ou précisant l'information obtenue lors de celle-ci. Elle cherche à optimiser le rendement des produits et des services existants en comprenant mieux qui sont les consommateurs et comment adapter l'offre pour mieux répondre à leurs besoins.

L'étude présentée ici se concentre sur l'extraction d'information pour les études secondaires. Dans le cas de la présente étude, les informations à disposition sont les données de facturation, l'étude réalisée est donc une étude de marché secondaire. Nous allons comparer les performances de deux méthodes, tout d'abord la méthode RFM, une méthode de segmentation pour faire de la prédiction basée sur trois critères : la Récence, la Fréquence et le Montant [Blattber et al., 2008]. RFM est une méthode traditionnellement utilisée en marketing, et ses résultats seront comparés à ceux obtenus à partir du data mining.

2.2 RFM

La méthode RFM est une méthode utilisée depuis le début des années 60. Le modèle a été développé à la base pour optimiser les retours du marketing direct. Il est basé sur trois variables : la Récence qui détermine le temps écoulé depuis la date du dernier achat, la Fréquence qui détermine le nombre d'achats effectués dans une période de temps donnée et finalement le Montant, qui indique le montant des achats effectués durant une période de temps donnée (différente ou pas de la période précédente) [Bult et Wansbeek, 1995].

Il existe plusieurs variantes et améliorations au modèle. La technique de base consiste à séparer les clients en segments pour les trois paramètres. C'est à dire qu'on attribue pour chaque observation trois notes selon la valeur obtenue pour chaque paramètre. Par exemple, en vérifiant si la relation est en général positive ou négative, on sépare l'ensemble des résultats en groupes et on attribue la meilleure note au groupe le plus enclin à répondre à l'offre, puis on diminue la note accordée jusqu'au dernier groupe. On peut multiplier ensuite ensemble les trois notes accordées pour R, F et M et ainsi obtenir des segments de clients plus enclin à répondre. [Miglautsch, 2000] a également proposé la possibilité de pondérer chacun des paramètres afin de donner plus ou moins d'importance à une des variables.

Les groupes sont ordinairement séparés en quintiles, 5 pour la Récence, 5 pour la Fréquence et 5 pour le Montant [Huges, 1996], [Jo-Ting et al., 2010]. RFM est un modèle très utilisé, mais certains auteurs critiquent le fait que la division des groupes en quintiles est arbitraire [Wheaton, 1996], [Yang, 2004]. Dans son article, [Yang, 2004] propose d'ailleurs une méthode basée sur la détection automatique d'interaction par chi-carré (CHAID) pour comparer la performance des clients afin de regrouper les clients sans grande différence de performance.

Plusieurs auteurs utilisent des modèles RFM et d'autres méthodes statistiques ou de data mining pour estimer la valeur des clients [Cheng et Chen, 2009], [Fader et al., 2005], [Shorabi et Khanlari, 2007]. Par exemple, [Shorabi et Khanlari, 2007] utilisent le clustering combiné au RFM afin de modéliser la valeur du temps de vie des clients d'une banque.

Finalement, [Colombo et Jiang, 1999] proposent un modèle RFM stochastique pour lequel le nombre de réponses d'une sollicitation suit une distribution binomiale et dont la probabilité de réponse de chaque individu suit une distribution beta, étant donné que la probabilité de réponse de chaque individu diffère.

La littérature montre donc que les recherches plus récentes utilisent les modèles RFM comme point de départ à l'élaboration de modèles plus complexes et raffinés faisant l'utilisation d'autres méthodes statistiques ou d'apprentissage.

2.3 Data mining

[Berry et Linoff, 2004] définissent le data mining comme étant un processus d'extraction et de détection d'informations et de modèles cachés dans de grandes bases de données. Sept types de modélisations sont définis par [Turban, et al., 2007] comme étant les recherches d'associations, la classification, la segmentation (clustering), la prédiction, la régression, la découverte de séquences et la visualisation.

Trois de ces types de modélisation ont été utilisés dans cet article afin d'obtenir de l'information additionnelle et différente de celle proposée par les méthodes traditionnelles d'étude de marché. Les trois méthodes utilisées sont présentées dans l'ordre d'utilisation :

2.4 Clustering

L'analyse de cluster cherche à détecter des groupes d'objets homogènes. Il s'agit pour y arriver d'évaluer la distance entre ces objets. La définition de distance et d'homogénéité change selon le contexte d'utilisation et selon les structures à détecter. On retrouve de nombreuses applications de cette méthode dans divers domaines. Par exemple, en biologie, imagerie, psychiatrie, psychologie, géologie, marketing, etc. [Jain et al., 1999]. Il existe de très nombreuses méthodes de segmentation. Une des principales est la méthode k-means [MacQueen, 1967]. Par contre, la faiblesse de cette méthode est qu'il faut donner des points de départ pour faire la segmentation. Dans des cas où la segmentation compte beaucoup de dimensions ou beaucoup de points, le mauvais choix des points de départ peut mener la méthode à converger vers des optimums locaux. Ce qui a motivé l'élaboration de nombreuses variantes et améliorations de la méthode. Plusieurs d'entre elles sont présentées dans le livre de [Kaufman, Rousseeuw, 1990]. Par exemple, la méthode CLARA (CLustering LArge Application), méthode qui est particulièrement utile pour les grandes bases de données. Elle fonctionne en tirant

aléatoirement plusieurs échantillons sur lesquels la méthode PAM est appliquée. Les k medoids du sous échantillon obtenant le meilleur résultat sont ensuite utilisés pour calculer le partitionnement sur l'ensemble de l'échantillon.

2.5 Association

Ce type de méthode vise à établir les liens existants entre des items dans un enregistrement. [Agrawal, et al., 1993] ont posé les bases de l'algorithme Apriori [Agrawal et Srikant, 1994], algorithme le plus largement utilisé dans la recherche de règles d'utilisation. Les règles d'association sont la plupart du temps utilisées dans des contextes d'analyses « market-basket » et dans l'élaboration de programmes de ventes croisées. Le principe de l'algorithme est de générer tour à tour des ensembles d'items fréquents contenant de plus en plus d'items. On détecte ainsi les associations d'items les plus fréquents. La fréquence d'un item (son support) est calculée comme le nombre d'occurrences de celui-ci. Toutes les combinaisons de deux items parmi les items obtenant un support supérieur à un certain seuil sont générées. Les combinaisons fréquentes (supérieure au seuil) sont ensuite conservées pour générer un ensemble contenant k+1 item où k est le nombre d'items dans l'ensemble. Les ensembles sont générés jusqu'à ce que plus aucun ensemble ne dépasse le seuil. On obtient de cette manière toutes les combinaisons d'objets les plus fréquentes. Trois indicateurs permettent d'indiquer la qualité et l'importance d'une règle : le *Support*, la *Confiance* et la *Conviction (Lift)*. Le *Support* indique la probabilité de retrouver l'ensemble antécédents dans la base de données. La *Confiance* indique la probabilité que la règle soit respectée. Il s'agit de calculer le ratio du *Support* du conséquent sur le *Support* des antécédents. La *Conviction* fait le ratio du support des ensembles contenant les antécédents et les conséquents sur le support des antécédents et des conséquents pris séparément. Il s'agit donc d'un indicateur qui compare la fréquence de respect de la règle à la fréquence de respect de la règle si elle n'était dû qu'au hasard.

2.6 Arbre de classification

Les arbres de classification tentent de prédire une variable en fonction d'autres variables comprenant des classes prédéfinies. De nombreuses méthodes de construction d'arbres existent. [Breiman, et al., 1983] ont développé la méthode CART (Classification And Regression Tree) pour développer un arbre binaire. Puis [Quinlan, 1983] a développé un modèle pour les arbres de type n-aire avec son modèle ID3. La construction d'un arbre se fait par découpages successifs en cherchant à chaque étape le critère qui donne le meilleur tri pour une branche donnée. Les algorithmes varient ensuite principalement par le critère de sélection de la meilleure découpe et par les critères d'élagages pris en compte.

3 CAS D'ÉTUDE

3.1 Contexte

Quelques outils présentés dans l'état de l'art ont pu être testés pour apporter de nouvelles connaissances à une entreprise faisant principalement affaire au Québec. Pour accéder à son offre de produits, il est nécessaire d'être abonné à la compagnie. Elle peut ainsi profiter de son grand nombre de clients, pour offrir des rabais sur des produits, services ou activités. L'entreprise concentre son offre sur 7 catégories de produits soient : *Restaurants*, *Arts & Entertainment*, *Shopping*, *Beauty & Spas*, *Healt & Fitness*, *Travel*, *Food & Drink*. La catégorie *Arts & Entertainment* est composée de réductions pour des événements culturels ou des événements de divertissement. Les offres classées dans la catégorie *Healt & Fitness* concernent des cours de sports ou des réductions sur des services de santé. *Travel* est surtout composé d'offres sur des hôtels, mais également sur des services de transports. Finalement *Food & Drink* représente les offres de commerces ou services en lien avec la nourriture.

Tous les produits sont exclusivement offerts sur le web ou par une application mobile. Chaque jour, un courriel est envoyé aux membres. Ce courriel présente les nouvelles offres du jour, ainsi que les plus populaires.

Les offres sont donc pour la plupart uniques. Comme l'offre et les prix sont très diversifiés, la clientèle est difficile à identifier. Les efforts marketing de l'entreprise sont donc orientés à toucher le plus large public possible.

3.2 Analyse descriptive

Dans cette section, une analyse descriptive a été effectuée dans le but d'obtenir une vue d'ensemble des données et pour pouvoir vérifier les informations contenues dans les données. Ceci s'est réalisé en étudiant les valeurs limites, les moyennes et les statistiques de base des données de transactions dont dispose l'entreprise. Il a été possible de confirmer que l'information extraite des données était juste, en vérifiant des informations simples comme le chiffre d'affaires, les ventes moyennes, etc.

En tout, l'information de plusieurs millions de transactions était disponible pour analyse. Les transactions avaient toutes eu lieu sur une période de 6 ans. Les données comprenaient, entre autres, l'identifiant du client (un code unique par client, l'identifiant du produit (un code unique par produit), la quantité de produits achetés, le montant déboursé, la date de l'achat, la catégorie du produit, la taxonomie (sous-catégorie) du produit, la date de l'achat, etc. (voir Tableau 1)

Tableau 1. Données de transaction

DEALID	[redacted]USER_ID	ORDER_QTY	ORDER_AMOUNT	ORDER_DATE	taxo_ID	no_categories
855352	42440132	3	XX.XX	2014-05-13	172	1
849161	42378871	1	XX.XX	2014-05-01	32	1
1190252	42392402	2	XX.XX	2014-10-28	5502	17
1320062	42475672	2	XX.XX	2015-02-15	712	3
1387302	42486502	2	XX.XX	2015-05-06	512	1
2394862	42425902	2	XX.XX	2015-11-20	332	1
1340002	42427772	1	XX.XX	2015-02-21	5942	17
873032	42517362	1	XX.XX	2014-06-12	5502	17

Avec des manipulations de données et en appliquant divers filtres sur celles-ci, des structures permettant de visualiser les données ont été obtenues. Le but de ces manipulations étant de trouver des facteurs pouvant influencer les ventes. Les effets de trois phénomènes ayant possiblement un effet sur les ventes ont donc été étudiés :

1. L'évolution dans le temps
2. Les catégories d'offres
3. Le prix des offres

Ces trois analyses permettent d'aider la compagnie à répondre à plusieurs questions d'une étude de marché. Notamment sur l'offre de produit avec le point 2, les prix avec le point 3 et sur les offres promotionnelles avec le point 1.

3.3 Évolution dans le temps

L'évolution dans le temps cherche à identifier si certaines périodes de l'année sont caractérisées par des variations significatives des ventes.

Pour ce faire, on a regroupé les transactions en prenant la somme quotidienne des ventes pour tous les jours depuis 2006. La variation mensuelle des ventes annuelles (Figure 1) a révélé

que celles-ci étaient toujours minimales en début d'année, soit durant le mois de janvier, maximale en décembre et qu'un minimum local s'observait au mois de mai et dans les environs du mois d'août.

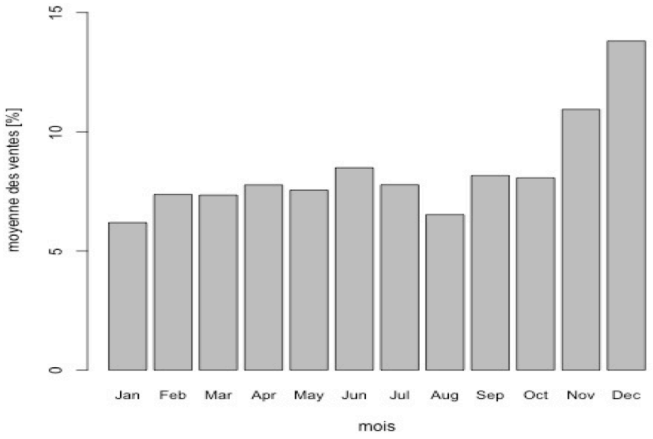

Figure 1. Moyenne mensuelle du pourcentage des ventes annuelles

Selon les types d'offres, on observe qu'il existe de légers effets de saison. L'effet de janvier et de décembre s'explique par la période entourant les Fêtes. Les clients dépensent donc plus en décembre et inversement dépensent moins dans le mois suivant cette période. Les minimums de mai et d'août pourraient s'expliquer par les périodes d'activité scolaire. Le mois d'août se caractérise par la période de retour en classe et même plus généralement de fin des vacances pour les travailleurs. Tandis que le mois de mai est généralement la période des examens. Donc, une recommandation possible pour augmenter les ventes à ces périodes de l'année, serait de proposer des produits en lien avec les activités des clients à ce moment de l'année. Par exemple, en mai, durant les examens, offrir des rabais sur les services de tutorat, et en août sur le matériel scolaire et de bureau.

Par la suite, en analysant les ventes quotidiennes, des pics particuliers de ventes ont été identifiés. Ces pics ont pu être reliés à des promotions ou à des offres particulièrement populaires.

L'étude de l'évolution des ventes dans le temps permet donc d'identifier des variations dans les ventes et d'identifier des évènements ponctuels.

3.4 Catégories d'offres

Afin de pouvoir améliorer leur offre, il est important de bien comprendre comment se distribuent les ventes dans les différentes catégories de produits. Pour ce faire, la somme des ventes des taxonomies (sous-catégories) a été calculée pour chaque mois de chaque année. À partir de ces données, les calculs des proportions en volume et en montant de ventes sont effectués (Tableau 2).

Tableau 2. Proportion des ventes des principales catégories

Catégorie	*Volume [%]*	*Montant [%]*
Restaurants	23,3	16,6
Arts & Entertainment	21,0	11,4
Shopping	14,4	14,2
Beauty & Spas	12,0	17,9
Healt & Fitness	8,2	7,9
Travel	7,2	22,5
Food & Drink	4,3	2,0
mineur	4,1	2,5
Professional Services	3,3	1,9

Le Tableau 2 montre les proportions en volume et en montant des ventes qu'occupe chacune des catégories. Les valeurs sont calculées à partir de toutes les transactions contenues dans les données. On observe des choses traditionnelles qui montrent que ce ne sont pas les catégories les plus vendues qui apportent le plus de revenus. Il est donc pertinent d'analyser plus en détails les données disponibles.

4 ÉTUDE COMPARATIVE

4.1 RFM

Cette section présente une application de la méthode RFM. Une segmentation sur les clients servira à établir quelles sont les catégories de produits les plus susceptibles d'intéresser les groupes de clients les plus potentiellement rentables selon le modèle. Ensuite, le taux de réponse sur les données de l'année suivant l'année de l'analyse sera calculé pour tous les différents groupes de segmentation afin de tester l'efficacité du modèle.

Dans notre cas, la récence est définie comme le nombre de jours entre la date du dernier achat et la date de l'analyse. Étant donné que l'on cherche ensuite à mesurer l'efficacité du modèle, la date de l'analyse a été choisie comme étant le premier jour de l'année sur laquelle la mesure de l'efficacité sera basée.

On mesurera l'efficacité du modèle en comparant le ratio retour (achat dans l'année suivant l'analyse) sur le nombre de membres d'un groupe. RFM est un modèle qui cherche à donner l'inclinaison d'un client à revenir faire affaire avec la compagnie. Donc, pour vérifier si le modèle fonctionne bien, le ratio nombre de retour sur nombre d'individus dans le groupe devrait être croissant avec le score.

La fréquence est définie comme étant le nombre total d'achats effectués par le client depuis son premier achat. Finalement, le montant est défini comme la valeur totale du montant dépensé par le client pour tous ses achats.

Les valeurs obtenues pour chaque paramètre sont ensuite divisées en intervalles. Pour F et M nous utilisons 5 intervalles, le cinquième intervalle, composé des valeurs les plus élevées, se voit octroyer la note de 5. Le premier intervalle, composé des valeurs les plus basses, se voit octroyer la plus basse note, soit 1. Pour la récence (R) nous utilisons 10 intervalles. Le premier reçoit la meilleure note, un 10 et le dernier, la plus basse note, soit 1. Les notes sont ensuite multipliées entre elles, pour chaque client, afin d'obtenir une note globale.

L'échelle de la récence utilise plus d'intervalles afin d'augmenter le poids de ce critère dans le score des acheteurs récents et donc de pénaliser les acheteurs inactifs depuis un certain temps. La méthode relègue habituellement les acheteurs inactifs depuis une période de temps choisie au dernier rang ; la fréquence et le montant étant habituellement calculés sur une période de temps fixée et récente. Dans le cas présent, il faut tenir compte du type d'offres de la compagnie [Lumsden et al., 2008] en notant que les membres font en moyenne 2 achats par année. Donc, si on calcule le montant et la fréquence sur une période de temps, par exemple un an, la différence entre les résultats possibles sera mince. Un membre n'ayant pas fait d'achats durant l'année obtiendra une fréquence de 0 tandis que l'acheteur moyen obtiendra une fréquence de 2. Le calcul de la fréquence et du montant sur une période de temps diminue donc la variance autour de la moyenne. La différence entre les groupes pour ces paramètres devient donc moins significative et plus difficile à détecter.

Prendre l'ensemble des achats sur toute la vie du client,

désavantage toutefois les nouveaux clients pour le score de la fréquence et du montant. Cet effet est quelque peu compensé par la pondération plus importante sur la récence.

4.1.1 Segmentation selon RFM

La division en 10 groupes pour la récence puis en 5 groupes pour la fréquence et le montant donne 60 groupes de notes différentes possibles par les combinaisons des multiplications entre les nombres de 1 à 10 avec les nombres de 1 à 5 puis de 1 à 5 encore.

Avec chacun des groupes de score, on calcule le nombre d'achats effectués dans les différentes catégories pour tous les membres d'un groupe. On divise ensuite le nombre d'achats de chaque catégorie par le nombre d'offres dans ces catégories. On obtient de cette manière le nombre d'achats par offre pour chaque catégorie d'offres dans chaque groupe. On classe finalement les catégories en ordre de nombre d'achats par offre décroissant pour tous les groupes (Tableau 3).

Tableau 3. Catégories favorites selon le score

score	1	2	3	4	5	6	7
250	Arts and Entertainment	Restaurants	Shopping	Travel	Health & Fitness	Beauty & Spas	Food & Drink
225	Shopping	Restaurants	Beauty & Spas	Food & Drink	Arts and Entertainment	Travel	Health & Fitness
200	Arts and Entertainment	Shopping	Health & Fitness	Beauty & Spas	Food & Drink	Restaurants	Travel
180	Food & Drink	Health & Fitness	Beauty & Spas	Restaurants	Arts and Entertainment	Shopping	Travel
175	Arts and Entertainment	Restaurants	Shopping	Health & Fitness	Beauty & Spas	Travel	Food & Drink

De cette manière, on peut observer quels produits sont les plus consommés par les meilleurs clients identifiés par RFM.

4.1.2 Résultats RFM

Pour vérifier que la méthode arrive bien à identifier les groupes les plus enclins à racheter pour le cas d'application, on a noté un échantillon de données ayant effectué des achats avant l'année **n**. Le ratio de réponse à l'année **n**+1 pour chaque groupe sera évalué. Comme la méthode vise l'identification des meilleurs clients, le ratio de réponse devrait croitre en fonction du score. Si tel est le cas, cela signifie que la méthode permet de bien segmenter les clients en fonction de leur probabilité de retour.

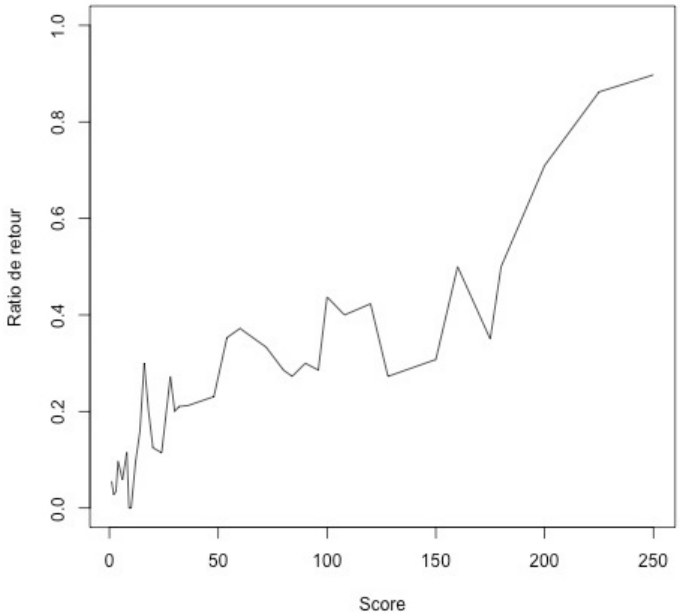

Figure 2. Proportion des clients ayant acheté des produits en fonction de leur score

En ne conservant que les groupes contenant au moins 1% de l'échantillon de clients, on obtient la Figure 2. On constate que la segmentation RFM permet bien d'identifier les meilleurs clients et qu'en général le retour des clients est croissant en fonction de leur note.

Ainsi, en appliquant la méthode on est en mesure d'identifier des groupes de clients plus enclins à continuer à faire des affaires avec la compagnie. Il est également possible d'identifier quels produits sont les plus achetés par les meilleurs groupes.

Ceci nous permet de développer une stratégie marketing optimisée pour ces clients plus enclins à revenir. En utilisant cette information pour développer des promotions, on peut maximiser le retour. Aussi, en suivant l'évolution de l'appartenance des clients à un groupe, il est possible de prévoir son temps de survie.

4.2 Data Mining

Dans cette section, des techniques de data mining sont utilisées pour trouver de l'information non triviale sur les clients [Anand et Büchner, 1998]. Elle permet, dans un premier temps, d'identifier des segments de clients en utilisant des techniques de segmentation [Carrier & Povel, 2003]. Elle permet aussi de détecter les ventes croisées avec les règles d'association. Finalement, un arbre de classification est développé dans le but de permettre aux gestionnaires de l'entreprise de déterminer quels facteurs permettent d'identifier les acheteurs potentiels de l'année suivante. Le choix de ces méthodes en particulier s'est fait sur la base qu'elles couvraient assez bien le spectre d'analyse du data mining. Tout en permettant d'obtenir des résultats immédiatement utilisables pour des prises de décision stratégiques.

4.2.1 Segmentation de la clientèle

La segmentation de la clientèle permet d'identifier les clients présentant des caractéristiques semblables. Dans le cas présent, la segmentation effectuée regroupe les clients présentant le même comportement d'achat.

Pour y parvenir, on construit une table pour laquelle chaque enregistrement représente un client et où chaque champ représente une catégorie de produits (Cf Table 3).

Tableau 4. Structure client-catégorie

ID_clients	Health...Fitness	Restaurants	Beauty...Spas	Arts.and.Entertainment	Food...Drink	Shopping	Travel
26428951	1	1	0	2	1	17	4
26428971	0	0	0	0	0	0	0
26429141	0	0	0	1	0	0	0
26429201	1	2	2	0	0	2	1
26429231	0	2	0	0	0	0	0
26429251	0	1	0	0	0	1	0
26429271	0	0	0	1	0	3	0

Les nombres inscrits dans les champs catégorie donnent le nombre de produits de cette catégorie achetés par le client.

Afin de réduire la dimensionnalité de la segmentation, seuls les sept catégories présentant le plus grand volume de ventes ont été incluses dans les variables.

Une fois la table construite, compte tenu de la grande dimension de la base de données, on applique la méthode CLARA.

Cependant, pour appliquer cette méthode, il faut déterminer le meilleur nombre de groupes k. Celui-ci est estimé de manière expérimentale, en utilisant une méthode par agglomération qui suit une première agglomération grossière (k initial = 50). Le dendrogramme obtenu est présenté à la Figure 3 :

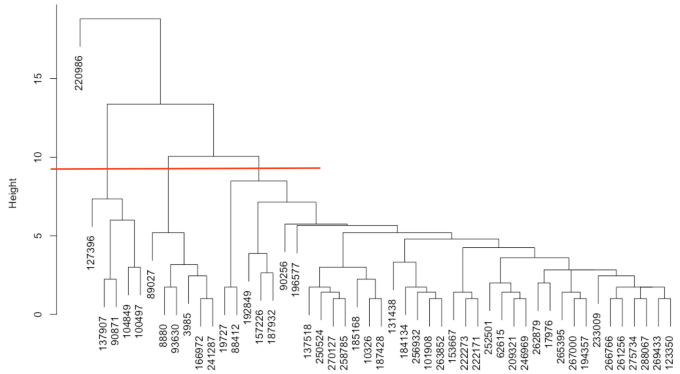

Figure 3. Dendrogramme avec choix du nombre de groupe

Le nombre de groupes optimal pour une distance donnée s'obtient en coupant le dendrogramme. Or, en observant la distance entre les groupes, on remarque qu'elle augmente rapidement jusqu'à 4 groupes puis augmente beaucoup plus lentement par après. Un bon candidat pour la segmentation semble donc 4 groupes, car ce choix maximise l'hétérogénéité entre les groupes.

Le Tableau 5 montre un descriptif général des groupes obtenus. Le groupe 1 représente 54% des clients. Ce groupe est responsable de 35% des achats effectués. Les membres de ce groupe font en moyenne deux achats.

Tableau 5. Description des groupes de segmentation

Groupe	*Client [%]*	*Achat [%]*	*Nombre moyen d'achats*
1	54	35	2
2	13	12	2,84
3	19	25	4,02
4	14	28	6,29

De manière plus détaillée, il est possible de voir quelles sont les catégories privilégiées par chaque groupe de clients :

Tableau 6. Nombre d'achats moyen par catégorie

Groupe	Health&Fitness	Restaurants	Beauty&Spa	Arts&Ent	Food&Drink	Shopping	Travel
1	0,1	0,3	0,4	0,1	0,1	0,7	0,3
2	1,9	0,3	0,2	0,1	0,1	0,1	0,2
3	0,1	0,3	0,3	2,8	0,1	0,2	0,2
4	0,3	4,2	0,4	0,4	0,3	0,4	0,3

Le Tableau 6 montre le nombre moyen d'achats de chaque catégorie pour un membre « moyen » de chaque groupe.

On observe que les groupes 2 à 4 sont des groupes qui concentrent leurs achats dans une catégorie. Environ 70% des achats de ces groupes se font dans une seule catégorie. Tandis que les membres du groupe 1 font des achats un peu plus distribués entre les diverses catégories.

En effectuant également la méthode avec 3 groupes, on obtient des résultats semblables. Par contre, avec 3 segments, le groupe 2 qui achète surtout des produits de la catégorie *Healt & Fitness* se redistribue dans les autres groupes.

Donc, la méthode proposée dans cette section permet d'identifier des groupes d'acheteurs et leur importance. Il devient de cette manière plus facile d'identifier les meilleurs clients et de savoir de quelle manière améliorer l'offre de service pour mieux répondre à leur besoin.

4.2.2 Associations entre les ventes

Cette section, présente les manipulations nécessaires à la découverte d'affinités entre les produits vendus au cours d'une année. Ces affinités sont considérées comme des achats de produits conjoints fréquents.

Pour y parvenir, l'algorithme Apriori [Agrawal et Srikant, 1994] est utilisé.

Pour appliquer cet algorithme, on construit, pour tous les clients, un vecteur contenant tous les achats qu'il a effectués au cours d'un laps de temps (dans le cas présent : une année).

Il faut toutefois considérer dans l'analyse la nature des données de l'entreprise, nous possédons le nom du produit, la catégorie et la taxonomie auxquels ils sont identifiés. Des règles d'associations tenteront donc d'être trouvées pour ces trois niveaux de structure. Considérant que la plupart des produits vendus sont uniques, il est attendu que peu de règles soient retrouvées à ce niveau.

Trois tables de données sont donc construites avec la même structure, mais pour trois niveaux différents :

Tableau 7. Ensemble des catégories pour chaque client

TID	Item
26508181	{Beauty & Spas,Arts and Entertainment,Food & Drink,...
26661641	{Travel,Arts and Entertainment,Beauty & Spas,Restaur...
28250691	{Others,Health & Fitness,Restaurants,Shopping}
27483251	{Beauty & Spas,Restaurants,Arts and Entertainment,Pr...
26928211	{Shopping,Professional Services,Restaurants,Others}

À partir de tables similaires, mais avec des taxonomies ou des produits, on génère les règles pour les deux autres niveaux de structures.

L'application de cette technique a permis de révéler quelques relations d'achats :

Tableau 8. Règles d'associations sur les catégories

Règle	Lhs_1	Lhs_2	Rhs	Support	Confiance	Lift
1	Arts and Entertainment	Food & Drink	Restaurants	0,013	0,59	1,7
2	Food & Drink	Shopping	Restaurants	0,013	0,57	1,6
3	Beauty & Spas	Travel	Restaurants	0,013	0,51	1,5
4	Arts and Entertainment	Travel	Restaurants	0,016	0,51	1,5

Le Tableau 8 montre quatre règles d'associations sur les catégories. Les colonnes *Lhs_x* représentent les antécédents de la règle, tandis que la colonne *Rhs* représente le conséquent. Les colonnes *Support*, *Confiance* et *Lift* donnent les indicateurs de qualité de la règle.

Les quatre règles extraites ici indiquent le même résultat : un client qui achète parmi deux catégories différentes achètera aussi de la catégorie *Restaurants*.

Tracer les règles tel que présenté à la Figure 4. Liaisons entre les catégoriesFigure 4, permet de visualiser les liens entre les catégories. Les flèches montrent de quelle manière les catégories sont reliées. Les nœuds reliant les antécédents au conséquent ont une taille proportionnelle au support de la règle. La couleur du nœud quant à elle tend vers le rouge à mesure que le *Lift* de la règle augmente. Être en mesure de visualiser ainsi les liens entre les produits permet une conception plus efficace d'offres promotionnelles.

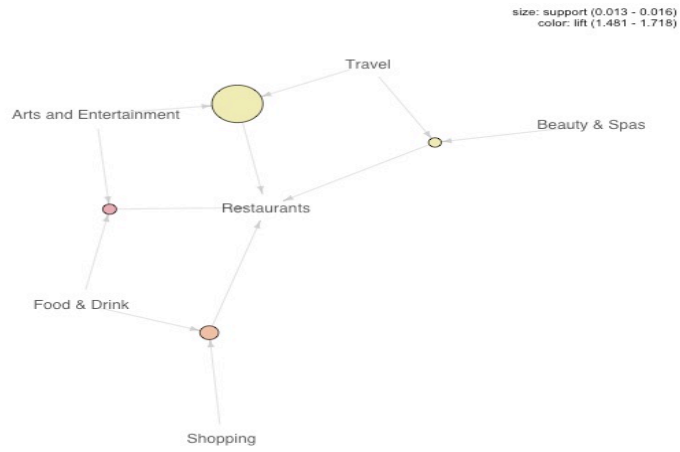

Figure 4. Liaisons entre les catégories

Des règles au niveau des taxonomies et des produits ont également été découvertes. Celles-ci ont permis de conclure que les liens sont plus forts à l'intérieur des catégories qu'entre elles car les *Lifts* pour ces règles étaient plus grands. Les règles sur les produits eux-mêmes ont révélées que les offres de réduction répétées étaient rachetées une seconde fois dans 30% des cas. Le *Lift* dans ces cas-là était le plus important de tous. Ce qui signifie que le lien entre ces produits est très fort.
L'analyse des règles d'associations dans une étude de marché donne donc beaucoup d'informations sur les ventes-croisées. Ce qui offre un bon point de départ pour la conception d'offres promotionnelles. Par exemple, offrir une promotion sur une offre de restaurants aux clients qui achètent des offres de deux catégories ou plus.

4.2.3 Catégorisation

[Chu et.al, 2007] a énoncé qu'il était entre cinq et dix fois moins cher de mettre des efforts à conserver les clients qu'à tenter d'en acquérir de nouveaux. Sachant cela, les compagnies sont prêtes à mettre beaucoup d'efforts pour savoir quelles décisions stratégiques peuvent avoir une influence sur la rétention des clients.
Une avenue de solution possible à ce problème est proposée ici. En utilisant les informations de transactions des clients et en proposant un modèle d'arbre de classification, des variables permettant de prédire le retour d'un client dans un laps de temps donné peuvent être identifiées. À partir de ces informations, des décisions concernant les offres de produits et les promotions peuvent être prises.
Dans les étapes qui suivent, un modèle d'arbre de classification est développé dans le but de trouver les variables permettant de prédire le retour d'un client l'année suivante.
Pour ce faire, une table dont chaque entrée représente un client ayant fait au moins un achat dans les 5 premières années de l'historique à disposition et dont les attributs sont ceux présentés dans le Tableau 9 est construite. Le but est de prédire l'achat (ou non) à l'année 6.

Tableau 9. Variables de classification

Variables	Description
Freq	Nombre d'achats total dans les 5 premières années
Montant_total	Montant total des achats dans les 5 premières années
nombre_cato	Nombre de catégories différentes achetées dans les 5 premières années
nbj	Nombre de jours entre le dernier achat et le premier jour de l'année 6
m_année1 ; … m_année5	Montant des achats dans chaque année
q_ année1 ; … q_année5	Nombre d'achats dans chaque année
cat_ année1 ; … cat_année5	Nombre de catégories différentes dans chaque année

Un attribut binaire : achat_année 6 est ajouté pour chaque entrée. C'est ce dernier attribut que l'arbre tente de prédire avec tous les autres.

Pour construire l'arbre, un échantillon de clients est tiré aléatoirement. Cet échantillon $\mathbf{D_T}$ sera celui utilisé pour entraîner l'arbre. Une fois l'arbre généré, on le teste avec le reste des données ou sur un autre sous-échantillon sélectionné aléatoirement $\mathbf{D_V}$.
La génération de l'arbre fonctionne selon le principe CART (Classification And Regression Trees) [Breiman, et.al, 1983]. Ainsi, la construction d'un arbre de classification commence avec l'ensemble des données dans un groupe. Ensuite, elles sont partionnées selon l'attribut le plus discriminant. Il n'y a pas de limite au nombre de fois qu'une variable peut être sélectionnée pour diviser un nœud. Les objets cessent d'être partitionnés lorsqu'ils sont assignés à une classe homogène ou qu'il n'y a plus d'attributs permettant d'améliorer la partition.
L'arbre obtenu prend la forme suivante :

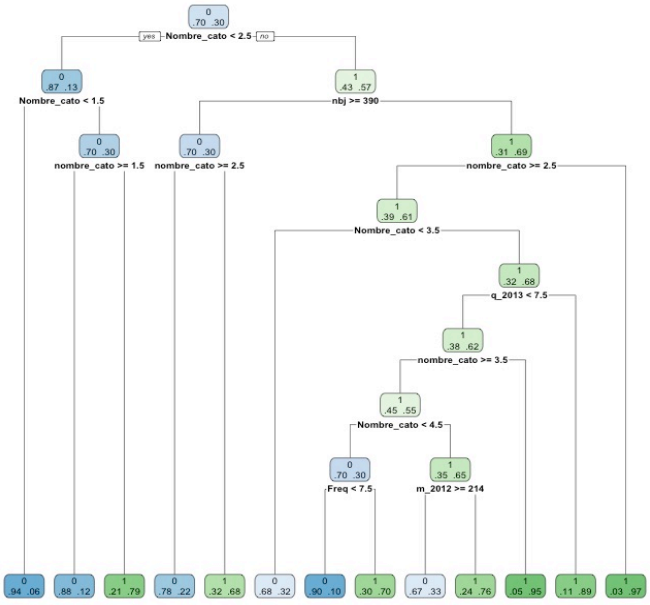

Figure 5. Arbre de classification

L'arbre débute par un nœud qui représente l'ensemble des clients. Il se divise ensuite en fonction d'un critère sur une variable. Les branches de gauche signifient que le critère est respecté (« oui ») et la branche de droite qu'il ne l'est pas (« non »).
En appliquant le modèle sur un échantillon de validation $\mathbf{D_V}$ composé de 1083 entrées sélectionnées aléatoirement dans la base, on obtient une précision de prédiction de 87,6%. Le Tableau 10 présente les résultats dans une table de confusion :

Tableau 10. Table de confusion

		Réelles		
		0	1	Erreur
Prédites	0	657	83	11,2%
	1	41	302	12,0%
	Erreur	5,8%	21,6%	87,6%

On observe avec la table de confusion (Tableau 9) que le modèle donne de meilleurs résultats pour prédire les clients qui n'achèteront pas que pour prédire les futurs acheteurs. En effet, le taux d'erreur pour la prédiction d'un achat futur est de 21,6% versus 5,8% pour prédire un non achat.
On constate sur l'arbre que le nombre de catégories de produits différents achetés revient à plusieurs reprises dans l'arbre. Il semble donc s'agir d'une variable importante pour fidéliser la clientèle. Cette information est donc cruciale dans l'élaboration de promotions. Elle permet de conclure qu'une manière de fidéliser les clients est de les convaincre d'acheter de plusieurs catégories de produits différentes.

4.3 Comparaison RFM – Data mining

D'après les résultats obtenus par les deux méthodes, on constate qu'elles ne fournissent pas le même type de résultats.
Dans un premier temps, la méthode RFM identifie des segments de clients plus enclins à demeurer client avec la compagnie. Ainsi, la méthode est à un niveau plus global et vise à concentrer les efforts marketing sur les meilleurs segments de clients.
Un avantage de l'utilisation du modèle RFM est qu'il est simple d'application. Il peut rapidement permettre d'obtenir des résultats facilement interprétables.
Le data mining, pour sa part, situe ses résultats à un niveau plus micro. Les résultats permettent d'identifier des types de clients. Par exemple, l'identification de comportements d'achat obtenu à partir du clustering ou les règles d'association peuvent servir à maximiser les retours pour l'ensemble des clients. En effet, cet outil permet d'apporter une description du phénomène des ventes croisées ou des comportements d'achat. Cet outil est donc mieux adapté pour chercher à quel part de marché s'adresser en présence d'un nouveau produit.
Ainsi, dans un contexte de marketing de e-commerce, le data mining peut mener à des outils de personnalisation du contenu et des offres.
Donc en résumé, le niveau de détails sur les résultats est plus important dans le cas du data mining. Les informations tirées des techniques permettent de mieux répondre aux besoins des clients. Par contre, ces méthodes sont plus complexes à appliquer et plus lourdes à calculer. Le Tableau 11 compare les résultats obtenus pour les deux méthodes dans le présent cas d'utilisation.

Tableau 11. Comparaison des résultats obtenus

RFM	Data mining
Identifie les segments ayant la plus grande probabilité de réponse selon 3 paramètres	Identifie des segments en fonction du comportement d'achat
	Découvre les relations entre les ventes
	Identifie la probabilité de réponse en fonction de critère sur tous les paramètres

5 CONCLUSION

Deux méthodes d'analyses utiles au domaine du marketing ont été appliquées sur les données d'une entreprise de e-commerce. Le but de cette application était de comparer les résultats afin de pouvoir tirer le meilleur de chaque méthode.
La première méthode appliquée : un modèle RFM a permis de conclure quels clients avaient le plus de chances de transiger de nouveau avec l'entreprise. En étudiant le taux de retour des clients en fonction de leur score global, on a pu conclure que la méthode permettait effectivement de bien identifier les clients avec la plus grande probabilité de retour.
La seconde méthode utilisait quelques-uns des outils du data mining. Leur application a permis d'identifier des segments de clients en fonction de leurs comportements d'achat grâce au clustering. Ensuite, des règles d'associations sur les achats ont fait découvrir les liens d'achats entre les données. Finalement, un arbre de classification a permis d'identifier quels paramètres permettent de prédire le retour d'un client l'année suivante.
L'analyse de la comparaison des résultats pour les deux méthodes amène à conclure que la stratégie développée est différente selon la méthode. En effet, les résultats du RFM sont globaux et permettent d'identifier les meilleurs clients. D'autre part, les résultats fournis par les méthodes de data mining permettent d'enrichir les données en décrivant les données et en distinguant des informations cachées sur les clients et les achats, tout en permettant de savoir comment prédire. Les informations obtenues peuvent donc servir à maximiser la stratégie marketing pour tous les clients, puisque les résultats sont descriptifs et permettent d'inclure tous les clients.
L'étude de marché réalisée sur les données est donc bien plus enrichie par l'utilisation du data mining. Ainsi, le problème d'identification de sources pour trouver de l'information nécessaire à la réalisation d'une étude de marché, identifié par le réseau entreprise Canada, se résout de lui-même par l'utilisation du data mining sur les données de facturation de l'entreprise elle-même.

6 REMERCIEMENTS

Les auteurs de l'article souhaitent remercier Tuango ayant donné accès à ses données afin de permettre la réalisation de cette étude.

7 RÉFÉRENCES

Agrawal, R., Imielinski, T., & Swami, A., (1993) Mining Association Rules between Sets of Items in Large Databases. Acm sigmod record, ACM., 22(2), pp. 207-216.

Agrawal, R., & Srikant R., (1994) Fast Algorithms for Mining Association Rules in Large Databases. Proc. 20th int. conf. very large data bases, VLDB., 1215, pp. 487-499.

Anand, S. S., & Büchner, A. G., (1998) Decision Support Using Data Mining. Financial Times Pitman: London.

Blattberg, R. C., Kim, B. D., & Neslin, S. A., (2008) Database Marketing, 1st Ed., Springer: New-York.

Blur, J. R., & Wansbeek, T., (1995) Optimal selection for direct mail. *Marketing Science* (1986-1998), 14(4), pp. 378-394.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J., (1983) Classification and Regression Trees. Wadsworth International Group, Belmont, Ca.

Carrier, C. G., & Povel, O., (2003) Characterising data mining software. Intelligent Data Analysis, 7, pp. 181-92.

Cheng, C. H., & Chen, Y. S., (2009) Classifying the segmentation of customer value via RFM model and RS theory. *Expert systems with applications*, *36*(3), pp. 4176-4184.

Chu, B. H., Tsai, M. S., & Ho, C. S., (2007) Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems*, 20, pp. 703-718.

Colombo, R., & Jiang, W., (1999) A stochastic RFM model. *Journal of Interactive Marketing,* 13(3), pp. 2-12.

Fader, P. S., Hardie, B. G., & Lee, K. L., (2005) RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, 42(4), pp. 415-430.

Hughes, A. M., (1996) Boosting Response with RFM. *Marketing Tools*, 3(3), pp. 4-5.

Jain, A. K., Murty, M. N., & Flynn, P. J., (1999) Data Clustering: A Review, *ACM Computing Surveys (CSUR)*, 31(3), pp. 264-323.

Kotler, P., & Keller, K., (2011) Marketing management (14th Edition), Upper Saddle River, NJ, Prentice-Hall.

Kaufman, L., & Rousseeuw, P. J., (2009) Finding groups in data: an introduction to cluster analysis*,* 344. John Wiley & Sons.

Ling, R., Yen, D. C., (2001) Customer relationship management: An analysis framework and implementation strategies. Journal of Computer Information Systems, 41, pp. 82-97.

Lumsden, S. A., Beldona, S., & Morrison, A. M., (2008) Customer Value in an all-inclusive travel vacation club: An application of the RFM framework. Journal of Hospitality & Leisure Marketing, 16(3), pp. 270-285.

Lumsden, S. A., Beldona, S., & Morrison, A. M., (2008) Customer Value in an All-Inclusive Travel Vacation Club: An Application of the RFM Framework. *Journal of Hospitality Marketing & Management*, 16(3), pp. 270-285.

MacQueen, J., (1967) Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*,* 1, pp. 281-297.

Miglautsch J. R., (2000) Thoughts on RFM scoring. *The Journal of Database Marketing*, 8(1), pp. 67-72.

Mitra, S., Pal, S. K., Mitra, P., (2002) Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, 13, pp. 3-14.

Ngai, E.W.T., Xiu, Li, Chau D.C.K., (2009) Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2), pp. 2592-2602.

Quinlan, J. R., (1990) Probabilistic Decision Trees. *Machine Learning: An Artificial Intelligence Approach*, 3, pp. 140-152.

Shorabi, B., Khanlari, A., (2007) Customer Lifetime Value (CLV) Measurement Based on RFM Model. *Iranian Accounting & Auditing Review*, 14(47), pp. 7-20.

Turban, E., Aronson, J. E., Liang, T. P., (2005) Decision support and business intelligence systems. 8th Ed., Pearson Education : USA.

Westphal, C., & Blaxton, T., (1998) Data mining solutions : methods and tools for solving real-worl problems. Wiley : New-York.

Wheaton, J., (1996) The Superiority of Tree Analysis over RFM: How It Enhances Regression. *DM News,* pp. 21-23.

Yang, A., (2004) How to Develop New Approaches to RFM Segmentation. *Journal of Targeting, Measurement and Analysis for Marketing,* 13(1), pp. 50-60

Réseau Entreprises Canada (2016), Guide pour l'étude et l'analyse des marchés, http://www.entreprisescanada.ca/fr/planification/etudes-de-marche-et-statistiques/effectuer-une-etude-de-marche/guide-pour-letude-et-lanalyse-des-marches/, (consulté le 20 octobre 2016)