FIRST DRAFT
Preprint

Segmentation e-commerce par RFM-V : Approches mathématiques et statistiques propres au Machine Learning

Andrieu Mickaël Msc[1]*

[1]Centrale Supelec Paris, France

Correspondance
Mickaël Andrieu, 30 rue André Mureine
33130 Bègles
Email: mickael.andrieu@solvolabs.com

La segmentation e-Commerce est un domaine à l'origine d'énormément de publications dans le domaine du marketing et du machine learning. Pourtant, les conclusions tirées dans ces publications n'ont pas toujours les mêmes critères de vérification ou de qualité et ne parviennent pas aux mêmes conclusions. Pinar Ozkan et pek Deveci Kocakoç de l'Université de Dokuz Eylul ont proposé un nouveau modèle de segmentation clientèle basé sur RFM-V et une nouvelle matrice de projects des segments de clientèle. Elles concluent que ce modèle permet une meilleure segmentation et a de nombreux intérêts pour le marketing.

Nous vérifions les conclusions avancées en adoptant les approches et méthodes mathématiques propres au Machine Learning.

KEYWORDS
segmentation cliente, k-means, marketing, maching learning, matrices clientèle

1 | INTRODUCTION

Pour juger de la qualité d'une segmentation de clientèle, deux facteurs entrent en compte : la qualité intrinsèque de la segmentation - les individus d'un segment sont similaires entre eux et différents des autres - et l'interprétabilité de ces segments. Il faut que la donnée proposée soit actionnable dans le cadre d'actions marketing, être capable de mieux qualifier un individu pour lui proposer une offre de services et de produits adaptée à ces besoins.

Abbreviations: RFM-V, Récence Fréquence Montant Variété.

En 1995, un premier modèle de segmentation basée sur la récence, la fréquence et le montant fut décrit dans la littérature[1], pratique par la rapidité et la facilité de mise en oeuvre mais assez pauvre sur l'actionnabilité pour l'e-commerce.

Depuis, de nombreux papiers sont publiés et de nombreuses variations de ce modèle de segmentation existent pour s'adapter aux besoins et objectifs marketing notamment dans le cadre de l'e-commerce.

1.1 | La segmentation RFM

Pour tous les individus de l'échantillon, on considère trois métriques qualifiant leurs comportements d'achats :

- La **récence**, c'est à dire le nombre de jours depuis la date de dernier achat effectué ;
- La **fréquence**, soit le nombre de commandes effectuées ;
- Le **montant**, soit le montant total dépensé ;

En Marketing, l'exploitation de ces variables se fait à l'aide de la méthode des quantiles qui consiste à considérer un score qui va de 1 à 5 pour chacune de ces variables, et à additionner les scores de sorte que chaque individu se trouve dans 15 segments distincts (de 1 à 15) que les équipes regrouperont selon leurs besoins ou à l'aide de segments déjà connus dans la littérature ("Champions", "Lost").

1.1.1 | La segmentation RFM-V

Dans le modèle étendu proposé [2], on y introduit la notion de **variabilité**, c'est à dire le nombre de produits ou de services *distincts* achetés.

Dans l'article, il n'a pas été effectué de traitement par méthode des quantiles et donc nous ne considèrerons pas l'exploitation de scores par la méthode des quintiles.

Nous avons donc deux modèles de données : la segmentation RFM constituera notre base de référence pour évaluer les performances du modèle étendu RFM-V.

1.2 | Méthodologie et Implémentation

Le protocole de recherche a été organisée de la façon suivante :

1. Collecte des données
2. Détection et exclusion des valeurs aberrantes
3. Segmentation des individus à l'aide du Machine Learning
4. Projections des segments
5. Vérifications mathématiques et statistiques

1.2.1 | Collecte des données

Les données ont été obtenues auprès du gérant d'un site e-commerce actif réalisant un chiffre d'affaires compris entre 5 et 10 millions d'euros par an, essentiellement auprès d'une clientèle de nationalité Française.

Cette collecte des données a été organisée dans le cadre d'un partenariat explicite et dans le strict respect du règlement européen RGPD.

L'échantillon se compose de *136 637 individus* avec les variables nécessaires au projet de recherche :

- la récence
- la fréquence
- le montant
- la variété

1.2.2 | Détection et exclusion des valeurs aberrantes

Deux méthodes ont été exploitées pour détecter et retirer les valeurs aberrantes : une méthode naïve et une approche par Machine Learning non supervisé.

L'approche naïve a consisté à retirer tout individu dont l'une des variables était inférieure ou égale à 0. En effet, nous cherchons ici à regrouper des individus entre eux sur la base de leurs achats. Or, dès qu'un achat a été réalisé les variables considérées sont **nécessairement positives**.

Nous avons ensuite procédé à l'exclusion de valeurs aberrantes par détection à l'aide du modèle de Machine Learning non supervisé DBSCAN.

Cette méthode se base sur la densité de population dans l'espace multi-dimensionnel pour regrouper les individus en clusters, et considère que les individus trop éloignés de tous les autres comme des valeurs aberrantes.

Puisque nous ne pouvons pas faire d'analyse descriptive plus avancée pour savoir si un individu présente des valeurs aberrantes ou non (compte de test, bot, erreur de saisie en base ...), cette méthode de détection présente un bon compromis entre le risque de perdre des individus au comportement intéressant et exploitable des véritables anomalies qui perturberaient la qualité de la segmentation.

Pour entraîner un modèle DBSCAN, il a fallu *estimer les valeurs optimales* du nombre d'individus proches et d' ϵ.

A l'aide de la méthode dite "des plus proches voisins" et par itération successive nous avons exclus **16 individus** en utilisant comme métrique discriminante le score de silhouette.

TABLE 1 Évaluation du paramètre epsilon (variables RFM).

epsilon	segments	valeurs aberrantes	score de silhouette
0.40	8	70	0.576
0.45	7	51	0.577
0.50	7	29	0.582
0.55	7	20	0.586
0.60	6	16	0.5937
0.65	6	12	0.5938

Pour un nombre d'individus proches égal à 6

1.2.3 | Segmentation des individus à l'aide du Machine Learning

Sur l'échantillon nettoyé, nous avons appliqué un algorithme de classification non supervisé appelé K-Means pour réaliser la segmentation.

K-Means a l'avantage certain d'être très rapide à s'exécuter et simple à comprendre et donc à optimiser. Son défaut est son comportement pseudo-aléatoire à l'initiation de l'entraînement qui ne permet pas d'obtenir toujours le même résultat lors de l'entraînement et de la segmentation de données.

1. L'algorithme définit aléatoirement un ensemble de centroïdes
2. Puis, il calcule la distance moyenne dans l'hyper espace entre les individus et les centroïdes et les assigne à des clusters
3. L'algorithme itère jusqu'à ce que la distance entre les segments soit **la plus grande possible** et la distance entre individus d'un même segment **la plus petite possible**

Avant entraînement, les données ont été standardisées à l'aide de la méthode StandardScaler de la librairie Python Open Source Scikit-learn.

Ensuite, nous avons sélectionné le nombre de segments en considérant les métriques suivantes :

1. Le score de silhouette
2. Le score de Davies-Bouldin

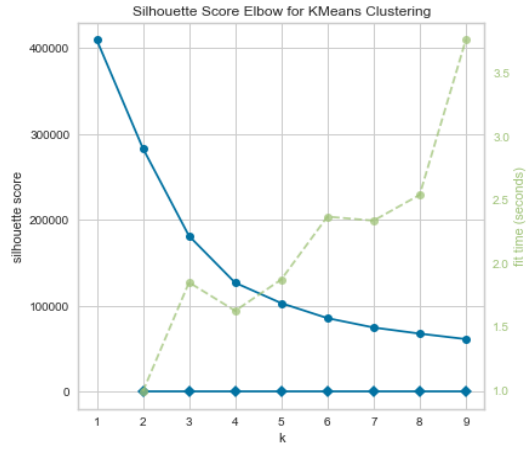

FIGURE 1 La technique "du coude" permet de choisir visuellement le nombre de segments ou de clusters : dès qu'il n'y a plus de gain effectif de score de silhouette, on prendra le nombre minimum de segments

En cas d'ambiguïté, il est recommandé de calculer le score de Davies-Bouldin pour faciliter la prise de décision.

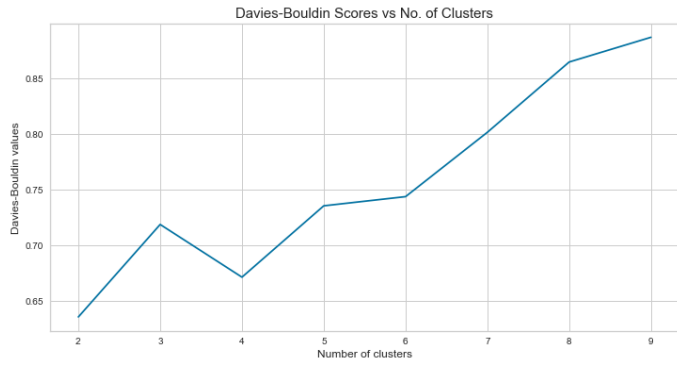

FIGURE 2 Le score de Davies-Bouldin est une métrique qui peut faciliter la prise de décision si la méthode du coude laisse un doute entre deux valeurs de segments.

Entre les modèles de segmentation RFM et RFM-V, après optimisation nous n'obtenons pas le même nombre de segments (ou clusters) :

TABLE 2 Récapitulatif segmentation : RFM RFM-V

Modèle	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
RFM	72379	55810	7162	923	114
RFM-V	74567	56700	5094	276	0

L'entraînement par K-Means produit 5 clusters pour RFM et 4 clusters pour RFM-V.

D'autre part, nous observons que le score de silhouette pour le modèle RFM-V est légèrement plus faible que celui pour le modèle entraîné en RFM :

TABLE 3 Récapitulatif métriques : RFM RFM-V

Modèle	Score de silhouette
RFM	0.53
RFM-V	0.49

RFM-V semble présenter une moins bonne segmentation que RFM.

1.2.4 | Projections des segments

Pour évaluer la qualité des segments, Pinar Ozkan and Ipek Deveci Kocakoç proposent d'exploiter la variabilité pour produire une nouvelle matrice de projection des segments de clientèle.

Originellement, les équipes Marketing produisent les matrices clientes en projetant les segments selon la fréquence et le montant d'achats. Les meilleurs segments sont isolés ("Best") et les autres segments sont travaillés selon qu'ils dépensent peu ou qu'ils achètent peu :

1. Aux clients qui dépensent peu mais souvent, on proposera des produits de meilleure qualité ("Up Selling")
2. Aux clients qui dépensent beaucoup mais rarement, on essaiera d'établir une relation plus sérieuse (mails de relance, promotions, espace privé)

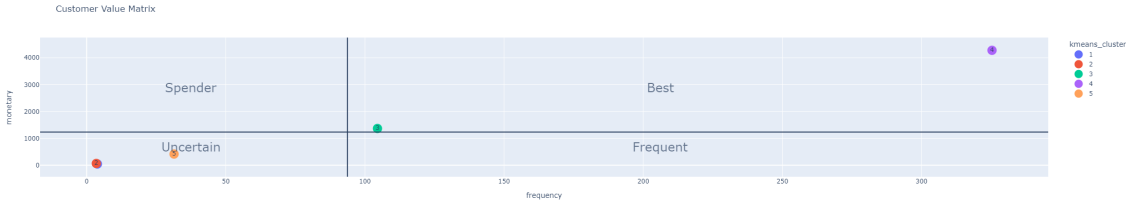

FIGURE 3 Matrice cliente de la segmentation de clientèle du modèle RFM

Visuellement, nous ne voyons pas - sur la projection classique - de différences flagrantes obtenues sur la segmentation obtenue par RFM-V qui dispose d'un segment en moins.

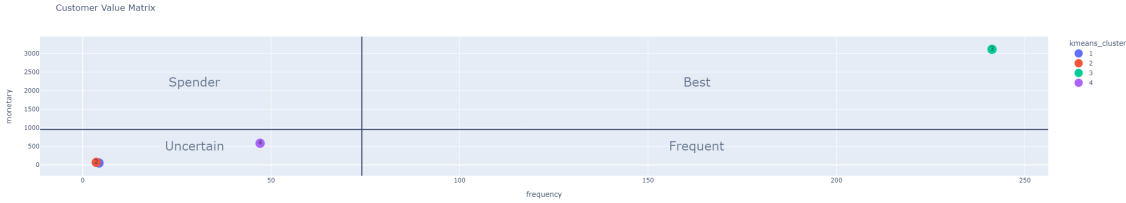

FIGURE 4 Matrice cliente de la segmentation de clientèle du modèle RFM-V

Dans cette nouvelle matrice, on fait le choix cette fois de considérer les individus en projetant le montant en fonction de la variété, ce qui va distinguer les clients qui achètent peu d'articles différents (passionnés, spécifiques) de ceux qui sont plus curieux dans leurs comportements d'achats.

Cette matrice est intéressante car elle permet aux équipes de décider quelle stratégie serait la plus efficace pour développer le chiffre d'affaires :

1. A une clientèle dont le comportement est spécifique, on proposera une montée de gamme sur le type de produits qu'il/elle achète déjà ("Up Selling");
2. A une clientèle plus curieuse, moins investie on investira plutôt sur la notion de produits liés ("Cross Selling")

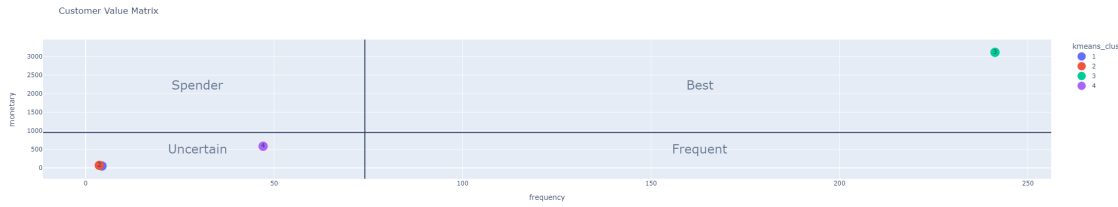

FIGURE 5 Matrice Produit de la segmentation de clientèle du modèle RFM-V

1.2.5 | Vérifications mathématiques et statistiques

En Machine Learning, nous n'évaluons pas la qualité de la segmentation ou son interprétation uniquement de façon visuelle mais à l'aide de métriques mathématiques.

Puisque nous avons deux modèles entraînés et les métriques correspondantes, on pourrait simplement comparer les deux valeurs et évaluer s'il y a une différence entre les deux.

Pourtant cette différence pourrait être due au hasard ou au fait que par son comportement pseudo-aléatoire à l'initialisation l'algorithme K-Means aurait pu produire une segmentation légèrement différente si on l'avait entraîné à nouveau.

Nous avons donc conçu un protocole de test statistique pour nous permettre de vérifier que la qualité de la segmentation RFM-V est bien supérieure à la segmentation classique RFM :

1. A l'aide d'échantillons aléatoires, nous entraînons 100 itérations RFM et 100 itérations RFM-V (10% des données d'origine)
2. Nous calculons les métriques suivantes : score de silhouette, score de Davies-Bouldin et score de Calinski Harabasz
3. Parmi ces métriques, nous prenons celle dont la distribution est normale ou pseudo normale en testant chacune des distributions
4. Puis nous appliquons les tests de Student et de Bartlett pour vérifier que les distribution des résultats obtenus par RFM et RFM-V sont significativement différents

Aucune des métriques testées n'a été démontrée normale ou pseudo normale, mais le score de silhouette présente une distribution qui semble suffisamment proche pour être utilisée :

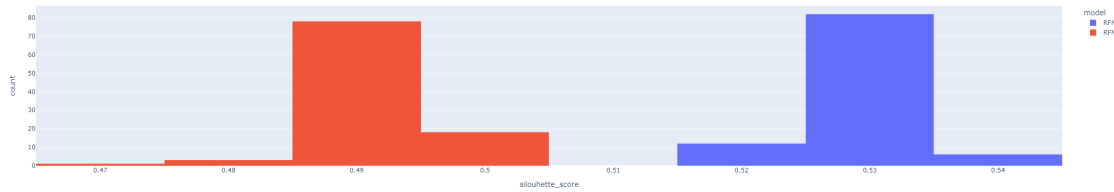

FIGURE 6 Distribution des score de silhouette obtenus au sein des 2 populations (RFM / RFM-V)

Le test de Bartlett est un test d'adéquation de variance entre des populations, on établit alors une hypothèse nulle selon laquelle les deux distribution ont la même variance.

Le test de Student est un test d'adéquation de moyenne entre des populations, on établit alors une hypothèse nulle selon laquelle les deux distribution ont la même moyenne.

Ces deux tests permettent de tester la probabilité que ce soit du au hasard.

Pour le score de silhouette, nous obtenons les résultats suivant avec un seuil à 1% :

TABLE 4 Tests statistiques

Test	p-value	Situation
Bartlett	0.167	Non rejet
Student	$6.26e^{-128}$	Rejet

Les deux distributions ont une moyenne statistiquement différentes

2 | CONCLUSIONS

L'étude statistique démontre que les variations observées entre le modèle RFM-V et RFM n'est pas le fruit du hasard et présente un score de silhouette moyen différent.

Nous avons observé dans le cadre de cette étude que pour les données choisies, le score de silhouette et donc la qualité de la segmentation sont légèrement plus faibles si l'on introduit la variabilité.

Il faudra cependant développer ce projet d'études et tester sur un grand nombre de jeux de données afin de vérifier que cette réalité s'observe de façon la plus probable et fait consensus.

REMERCIEMENTS

1. Je remercie Thomas Roux pour avoir partagé un extrait de sa base de données clientèle et autorisé l'exploitation pour analyse et pour publication en accord avec le règlement européen RGPD.
2. Je remercie Mamadou Cisse pour m'avoir accompagné et conseillé sur ces travaux de recherche, son expérience académique me fut très précieuse.
3. Je remercie l'organisme de formation OpenClassrooms qui rend possible et favorise la publication scientifique en rendant obligatoire la réalisation d'une preuve de concept académique dans le cadre de la validation du titre Ingénieur Machine Learning au sein d'associations pour le bien commun.

CONFLITS D'INTÉRÊT

Je n'ai aucun conflit d'intérêt et n'ai reçu absolument aucun financement ou avantage de nature à perturber les résultats de ce projet de recherche.

REFERENCES

[1] Bult JR, Wansbeek T. Optimal Selection for Direct Mail. Marketing Science 1995;(14):378–394.

[2] Ozkan P, Kocakoç ID. A customer segmentation model proposal for retailers 2021;.

MICKAËL ANDRIEU
Ingénieur Machine Learning Freelance spécialisé dans la segmentation e-commerce.
Anciennement membre de la Core Team en charge de la maintenance et du développement de la solution PrestaShop, il se spécialise depuis 2019 sur la valorisation de données Open Data et sociales.