Tạp chí Phát triển Khoa học và Công nghệ – Kinh tế-Luật và Quản lý, 6(1):2005-2015

Open Access Full Text Article

Bài Nghiên cứu

Một nghiên cứu liên ngành giữa phân tích phân khúc khách hàng trong marketing và phương pháp học máy

Hồ Trung Thành*, Nguyễn Đăng Sơn

Use your smartphone to scan this QR code and download this article

TÓM TẮT

Phân tích phân khúc khách hàng (Customer Segmentation) là một trong những vấn đề quan trọng trong việc quản lý khách hàng và xây dựng các chiến lược marketing phù hợp. Việc phân chia thành các nhóm khách hàng sẽ giúp những người quản lý nắm bắt rõ những đặc điểm của khách hàng hay hành vi tiêu dùng của họ, từ đó tiếp cận đúng khách hàng mục tiêu, giữ chân được khách hàng (Customer Retention), gia tăng được doanh thu và lợi thế cạnh tranh của doanh nghiệp. Tuy nhiên, phân tích để tìm ra đúng nhóm khách hàng là một vấn đề đặt ra mà doanh nghiệp cần giải quyết dựa trên cơ sở vững chắc và đáng tin cậy. Cùng với sự hỗ trợ từ các giải pháp công nghệ hiện nay như quản lý quan hệ khách hàng (Customer Relationship Management) và sự phát triển mạnh mẽ của công nghệ Khoa học dữ liệu, việc áp dụng các thuật toán, các phương pháp bao gồm cả định tính và định lượng nhằm giúp phân chia các nhóm khách hàng trong phân tích marketing. Bài báo này tập trung thực hiện một nghiên cứu liên ngành kết hợp giữa phương pháp RFM (Recency, Frequency, Monetary) và học máy (Machine Learning) để phân tích phân khúc khách hàng. Nghiên cứu được thực hiện thông qua phương pháp thực nghiệm trên tập dữ liệu (dataset) với 541,909 giao dịch của cửa hàng bán lẻ trực tuyến đã gom cụm được 5 phân khúc khách hàng với những đặc trưng của từng cụm được kiểm định chất lượng đã cho thấy tính hiệu quả và khả năng ứng dụng của nghiên cứu.

Từ khoá: Phân khúc khách hàng, RFM, học máy, phân cụm, tỷ lệ duy trì khách hàng

GIỚI THIỆU

Trong phân tích marketing hay các công việc liên quan đến quản lý, phục vụ, chăm sóc khách hàng, việc thấu hiểu khách hàng, cố gắng đem đến những sản phẩm, dịch vụ, trải nghiệm tốt nhất luôn là mục tiêu mà mọi doanh nghiệp hướng đến. Tuy nhiên hành trình này sẽ luôn chứa đựng nhiều vấn đề hay bài toán thậm chí là không dễ dàng để có được câu trả lời. Một sản phẩm hay một chương trình khuyến mãi khi tung ra thị trường khó có thể đáp ứng được hết nhu cầu của tất cả khách hàng. Chính vì vậy các doanh nghiệp đã chuyển dần sang việc phân chia khách hàng thành các nhóm riêng – được gọi là phân khúc khách hàng, nhằm tập trung hóa và chăm sóc khách hàng tốt hơn dựa trên những đặc trưng riêng của từng nhóm khách hàng.

Với sự phát triển mạnh mẽ của công nghệ khoa học dữ liệu hiện nay, việc thu thập và lưu trữ dữ liệu về khách hàng là nguồn tài nguyên mang nhiều giá trị tiềm năng đang chờ khai phá và cũng là cơ sở thuận lợi để áp dụng các mô hình toán học, thuật toán, phương pháp học máy trong việc khai thác và giải quyết các vấn đề kinh doanh. Từ việc phân tích dữ liệu, các quyết định của người quản lý có tính khách quan và đa chiều hơn. Các quyết định dựa trên dữ liệu (Data-driven decision making) được đưa ra sẽ giảm bớt được sự cảm tính vốn khó đo lường được. Việc kết hợp phân tích dữ liệu dựa trên các phân khúc khách hàng đã góp một phần vào sự thành công trong từng chiến lược marketing hay chính sách chăm sóc khách hàng nói riêng và duy trì được sự tồn tại, phát triển của doanh nghiệp nói chung trong bối cảnh thị trường chung có rất nhiều sự cạnh tranh khốc liệt.

Để giải quyết được vấn đề trên, trong nghiên cứu này sẽ tập trung vào bài toán phân khúc khách hàng với các mô hình, phương pháp phân tích dựa trên sự kết hợp hai nền tảng kinh doanh (Business) và công nghệ thông tin (Information Technology). Từ đó giúp cung cấp những chứng cứ về kết quả từ tổng quan đến chi tiết về tình hình vận hành kinh doanh và các chính sách với từng phân khúc khách hàng được phân tích. Một trong những lợi ích lớn nhất của phân tích phân khúc khách hàng là giúp doanh nghiệp quản trị khách hàng hiệu quả hơn. Khi doanh nghiệp phân khúc khách hàng thành những nhóm khác nhau (Hình 1) dựa trên nhân khẩu học, sở thích, hành vi mua sắm sẽ giúp doanh nghiệp có được chiến lược phù hợp để đồng hành cùng những nhu cầu mua sắm hay sử dụng dịch vụ của khách hàng và từ đó có thể phản hồi kịp

Trường Đại học Kinh tế - Luật, ĐHQG-HCM, Việt Nam

Liên hệ

Hồ Trung Thành, Trường Đại học Kinh tế - Luật, ĐHQG-HCM, Việt Nam

Email: thanhht@uel.edu.vn

Lịch sử

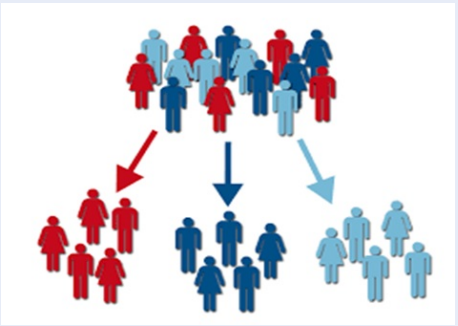
- Ngày nhận: 08/6/2021
- Ngày chấp nhận: 20/8/2021
- Ngày đăng: 04/9/2021

DOI : 10.32508/stdjelm.v6i1.850

Check for updates

VNU-HCM Press

Trích dẫn bài báo này: Thành H T, Sơn N D. **Một nghiên cứu liên ngành giữa phân tích phân khúc khách hàng trong marketing và phương pháp học máy**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 6(1):2005-2015.



Hình 1: Minh họa phân khúc khách hàng. (Nguồn: Subiz)[a]

[a] *P. Dung, "Phân khúc khách hàng để Marketing hiệu quả," Subiz, ngày 27/3/2017 tại https://subiz.com.vn/blog/phan-khuc-khach-hang.html . [Ngày truy cập lần cuối 28/06/2021].*

Hình 3: Một hóa đơn bán hàng tại cửa hàng bán lẻ tại Việt Nam

thời với những nhu cầu này.

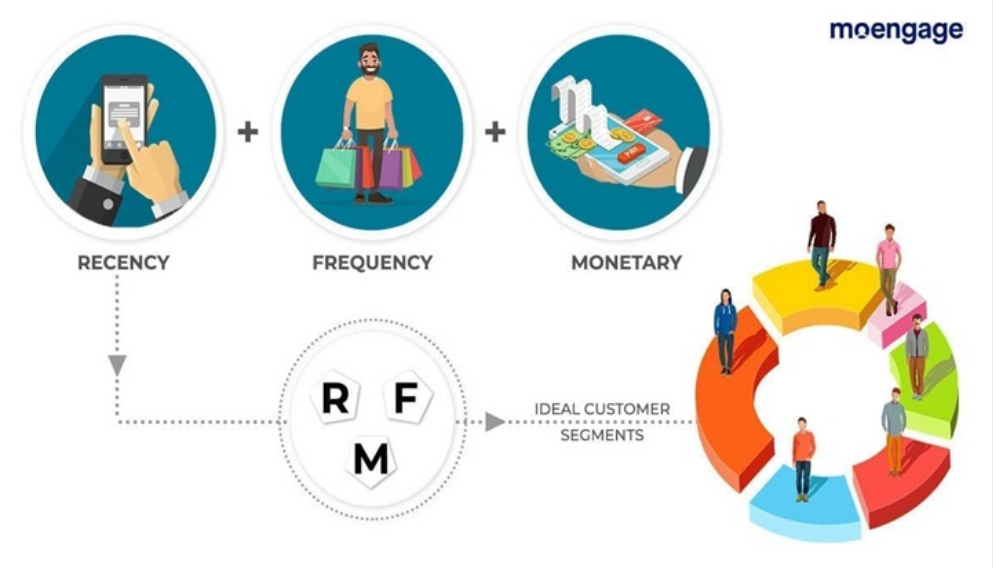
Nội dung tiếp theo của bài báo là phần 2 gồm cơ sở lý thuyết và các nghiên cứu liên quan, nhằm định hình, xác định các mô hình, thuật toán phù hợp với mục tiêu đặt ra. Các vấn đề liên quan và quá trình thực nghiệm được mô tả trong phần 3 - phương pháp và quy trình thực hiện nghiên cứu. Sau quá trình thực nghiệm, kết quả và đặc điểm của các phân khúc khách hàng được tìm ra được đề cập trong phần 4 và thảo luận kết quả. Phần cuối cùng là kết luận và hướng phát triển của nghiên cứu.

CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

Phương pháp RFM thường được sử dụng trong việc phân chia các nhóm khách hàng và tìm ra đặc điểm của từng phân khúc khách hàng. Trong Hình 2, phương pháp RFM được biết đến như một bản tóm tắt lại các giao dịch của khách hàng dưới ba yếu tố [1], bao gồm: Recency được xem là lần cuối gần nhất mà khách hàng đã mua hàng (khoảng cách giữa ngày tiến hành áp dụng phương pháp và ngày gần nhất khách hàng mua hàng); Frequency là tần suất mua hàng của khách hàng hay khách hàng đã mua hàng bao nhiêu lần; Monetary là tổng lượng tiền mà khách hàng đã chi tiêu cho toàn bộ hoạt động mua sắm.

Trong những giai đoạn đầu tiên, sau khi thiết lập được phương pháp RFM, mỗi yếu tố Recency, Frequency và Monetary của mỗi khách hàng thường được xếp hạng theo thứ bậc (ranking) với thang điểm thường từ 1 đến 5. Trong bài báo của John R. Miglautsch [2], tác giả đã xếp hạng các khách hàng bằng việc sử dụng nhãn nhóm khách hàng (Customer quintiles).

Tuy nhiên, về sau với nhu cầu của con người ngày càng phát triển, số lượng giao dịch, hàng hóa cũng tăng cao. Lượng khách hàng và các giao dịch ở mỗi doanh nghiệp cũng có sự thay đổi khác nhau và mang các đặc thù không giống nhau. Điều này dẫn đến việc vận hành phương pháp RFM cũng có sự thay đổi so với trước. Trong những nghiên cứu sau này, các nhà phân tích số liệu đã ứng dụng và cải tiến trong việc phân chia các nhóm khách hàng bằng việc sử dụng các thuật toán, phương pháp trên nền tảng toán học trong lĩnh vực học máy. Đây là một trong những lĩnh vực trong trí tuệ nhân tạo, lĩnh vực đang phát triển rất mạnh mẽ song song với ngành khoa học dữ liệu. Cụ thể, trong nghiên cứu của tác giả Palaksha Anitha và Malini Mrityunjay Patil [3] đã sử dụng phương pháp phân cụm (clustering) K-means – một phương pháp trong mô hình học không giám sát (Unsupervised Machine Learning) nhằm phân chia các nhóm khách hàng dựa trên ba yếu tố trong phương pháp RFM. Mỗi một phân khúc khách hàng lúc này được xem như là một cụm (cluster) trong K-means.



Hình 2: Minh họa phương pháp RFM (Nguồn: Moengage, 2021)[a]

[a] *Aditya, Predictive Segments using RFM Analysis: An In-Depth Guide [Updated], Moengage, ngày 22/2/2021, tại https://www.moengage.com/blog/rfm-analysis-using-predictive-segments/ . [Ngày truy cập cuối 28/6/2021].*

Điểm nổi bật khi sử dụng phương pháp K-means hay các phương pháp học máy nói chung đó là khả năng "tự học" của phương pháp. Các phương pháp trong học máy là tập hợp các bước xử lý dữ liệu dựa trên nền tảng toán học và thống kê. Đây cũng là điều giải thích cho sự khác biệt vì sao các phương pháp học máy nói chung khác với việc xử lý dữ liệu bằng phương pháp lập trình truyền thống. Chính vì vậy, với phương pháp có chất lượng càng tốt thì hiệu quả xử lý và thao tác trên những tập dữ liệu khổng lồ của các phương pháp học máy càng mạnh mẽ cũng như kết quả sau quá trình "tự học" dữ liệu sẽ tạo ra các quyết định và dự đoán tốt hơn [4,5].

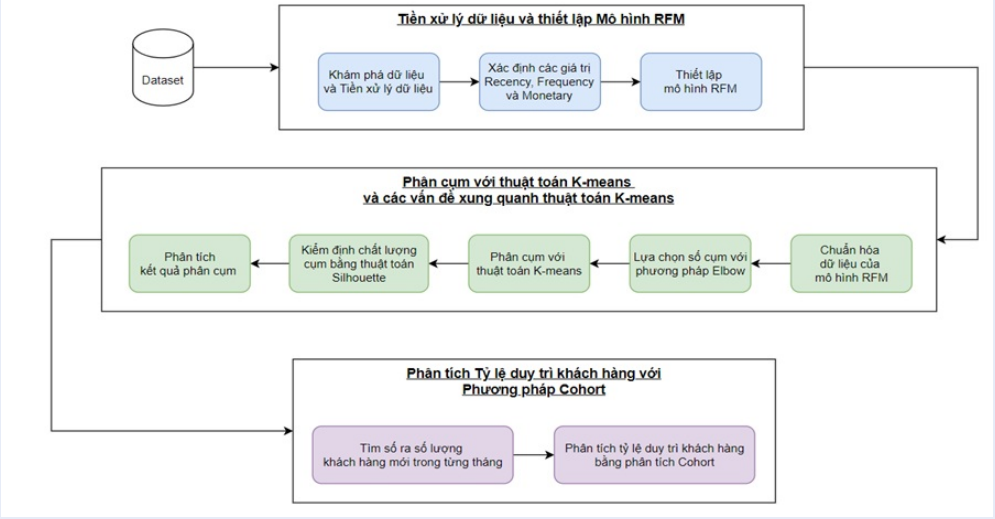
Trong nghiên cứu của nhóm tác giả [3] trên đã thực hiện phân cụm hai lần và chọn ra kết quả tốt nhất. Lần đầu tiên được thực hiện giữa Recency và Monetary và lần sau cùng được thực hiện giữa Frequency và Monetary. Trong nghiên cứu [1], bên cạnh việc sử dụng phương pháp K-means, tác giả cũng đã so sánh độ hiệu quả khi phân cụm trên các phương pháp Fuzzy C-means và RM K-means. Kết quả của nghiên cứu đã chỉ ra sự hiệu quả khi sử dụng các phương pháp phân cụm trong học máy cũng như cung cấp dữ liệu về đặc điểm hành vi khách hàng trong từng phân khúc.

Trong nghiên cứu của bài báo, nhóm tác giả khai thác những điểm mạnh của các nghiên cứu trước và từ đó đề xuất phương pháp nghiên cứu liên ngành kết hợp giữa phân tích phân khúc khách hàng trong marketing. Trong đó, bài báo tập trung xây dựng mô hình dữ liệu RFM dựa trên dữ liệu giao dịch với những tham số đặc trưng và cấu trúc tương đồng có thể tìm thấy trên các hóa đơn bán hàng trong bất kỳ cửa hàng tại các nước trên thế giới cũng như tại Việt Nam (Hình 3) và áp dụng phương pháp học máy không giám sát để phân tích phân khúc khách hàng và tìm ra những giá trị thật sự (insight) có khả năng tác động, ảnh hưởng tới hành vi và quyết định mua hàng của khách hàng. Bên cạnh đó, để đảm bảo chất lượng của kết quả nghiên cứu so với các nghiên cứu trước, bài báo sử dụng phương pháp Elbow với chỉ số kiểm định Silhouette để tối ưu số cụm khách hàng, hệ số chuẩn (Z-score) và Quy tắc kiểm chứng (Empirical Rule) được áp dụng để xử liệu các dữ liệu bất thường (Outlier) và phương pháp Cohort để phân tích tỷ lệ duy trì khách hàng kết hợp biểu đồ nhiệt trên phân phối ma trận.

PHƯƠNG PHÁP VÀ QUY TRÌNH THỰC NGHIỆM NGHIÊN CỨU

Phương pháp nghiên cứu

Hình 4 trình bày quy trình nghiên cứu với 4 giai đoạn chính như sau: 1) Giai đoạn 1 từ dữ liệu đầu vào là tập dataset được khảo sát và tiền xử lý (Data Preprocessing) nhằm tìm ra những đặc điểm không phù hợp. Sau đó, các đặc trưng cần thiết từ hành vi tiêu dùng của khách hàng tiềm ẩn trong dữ liệu được lựa chọn phù hợp với việc tính toán các giá trị Recency, Frequency, Monetary và cuối cùng là hoàn chỉnh mô



Hình 4: Phương pháp, quy trình nghiên cứu và thực nghiệm

hình dữ liệu RFM; 2) Giai đoạn 2 là giai đoạn chiếm tỷ trọng lớn cũng như có mức độ phức tạp nhất trong toàn bộ nghiên cứu. Từ việc khám phá dữ liệu ở Giai đoạn 1, các vấn đề và những đặc điểm liên quan đến các giá trị trong mô hình RFM cũng được tìm ra và chính điều này có làm ảnh hưởng đến dữ liệu đầu vào cho phương pháp K-means cũng như bảo đảm tính chính xác ở kết quả phân cụm khi phương pháp được thực thi. Do đó, trong giai đoạn 2, nghiên cứu sẽ lựa chọn các phương pháp và mô hình phù hợp với đối tượng dữ liệu nhằm giải quyết việc chuẩn hóa dữ liệu đầu vào và phương pháp kiểm định liên quan đến phương pháp K-means để đạt kết quả tốt nhất và phân tích các nhóm khách hàng, ra quyết định lựa chọn các nhóm khách hàng dựa trên kết quả phân tích từ phương pháp lai; 3) Giai đoạn 3 khai thác dữ liệu có được từ mô hình RFM, nghiên cứu sẽ tiến hành phân tích Cohort tìm ra số khách hàng mới mỗi tháng và tính được tỷ lệ duy trì theo biểu đồ nhiệt trên phân phối theo ma trận.

Thực nghiệm phân tích phân khúc khách hàng

Tiền xử lý dữ liệu và thiết lập mô hình dữ liệu RFM

Các phương pháp sau đây được thực hiện dựa trên tập dữ liệu (dataset) của một cửa hàng bán lẻ trực tuyến quà tặng và phụ kiện (có trụ sở đặt ở Vương quốc Anh)[6]. Tập dữ liệu chứa 541,909 giao dịch của một cửa hàng bán mặt hàng về quà tặng và phụ kiện. Trong đó có nhiều khách hàng của cửa hàng là nhà bán lẻ. Các giao dịch này xảy ra trong hai năm từ 2010 đến 2011.

Trên mỗi đơn hàng trên Hình 5, nghiên cứu sẽ tập trung khai thác các thuộc tính, bao gồm: thuộc tính CustomerID mỗi một số hóa đơn chỉ được thuộc về một khách hàng; thuộc InvoiceNo (số hóa đơn), mỗi một đơn hàng sẽ có mã hóa đơn riêng và mỗi số được phân biệt với các hóa đơn khác nhau. Một số hóa đơn xuất hiện nhiều bản ghi (record) trong dữ liệu và được hiểu là nhiều mặt hàng được mua trên cùng một hóa đơn. Thuộc tính này dùng để tính giá trị Frequency; thuộc tính Quantity (số lượng mỗi mặt hàng) đã mua mỗi hóa đơn; UnitPrice (đơn giá của mặt hàng). Với công thức *Quantity x Price* có thể xác định được tổng số tiền trên mỗi món hàng trong hóa đơn và từ đó xác định được thành tiền của mỗi đơn hàng. Các thuộc tính này dùng để tính giá trị Monetary; thuộc tính InvoiceDate (ngày mua hàng) dùng để tính giá trị Recency bằng cách chọn ra InvoiceDate mới nhất (gần nhất) trong toàn bộ hóa đơn (InvoiceNo) của từng khách hàng.

Sau quá trình khảo sát và tiền xử lý cũng như loại bỏ các giá trị không cần thiết và giữ lại các giá trị phù hợp, mô hình dữ liệu RFM được thiết lập với kết quả được trình bày trên Hình 6.

Chuẩn hóa dữ liệu mô hình RFM

Quay trở lại với mô hình RFM, khi quan sát các giá trị của Recency, Frequency và Monetary, có thể nhận thấy sự không tương đồng nhau về đơn vị và độ chênh lệch phạm vi giá trị quá lớn giữa ba yếu tố F, R và M khi xét đến tứ phân vị thể hiện trên Hình 7.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0

Hình 5: Một phần tập dữ liệu đầu vào.

CustomerID	Recency	Frequency	Monetary
12346.0	328	1	77183.60
12747.0	5	11	4196.01
12748.0	3	209	33719.73
12749.0	6	5	4090.88
12820.0	6	4	942.34
...	...	...	...
18280.0	280	1	180.60
18281.0	183	1	80.82
18282.0	10	2	178.05
18283.0	6	16	2094.88
18287.0	45	3	1837.28

Hình 6: Kết quả mô hình RFM sau khi được tiền xử lý và thiết lập.

	Recency	Frequency	Monetary
count	3920.000000	3920.000000	3920.000000
mean	94.742092	4.246429	1864.385601
std	99.533485	7.199202	7482.817477
min	3.000000	1.000000	3.750000
25%	20.000000	1.000000	300.280000
50%	53.000000	2.000000	652.280000
75%	145.000000	5.000000	1576.585000
max	376.000000	209.000000	259657.300000

Hình 7: Mô tả tứ phân vị trong dữ liệu RFM.

Giá trị Recency trải dài từ 3 đến 376 (ngày mua hàng gần nhất), Frequency trải dài từ 1 đến 209 (lần mua). Đặc biệt, Monetary là giá trị có miền giá trị lớn nhất từ 3.75 đến 259657.3 (đơn vị tiền tệ). Khi nhìn vào phân phối của tứ phân vị trong Monetary cũng đã có thể thấy Monetary có giá trị lớn hơn rất nhiều so với hai yếu tố còn lại.

Chính vì sự phân bố giá trị của các yếu tố trong tập dữ liệu và các ảnh hưởng của outlier đến kết quả phân cụm, giải pháp được ra đó là quy đổi các giá trị trên về cùng một đơn vị với phương pháp phân phối theo hệ số chuẩn (standard score) hay còn được gọi với tên gọi khác là Z-score[7]. Với điểm Z-score này sẽ giúp chúng ta hình dung được độ xa của một điểm dữ liệu so với điểm dữ liệu trung bình (điểm chuẩn). Công thức để quy đổi các giá trị theo Z-score như sau:

$$Z = \frac{x - mean}{std} \quad (1)$$

Trong đó với x là giá trị của điểm dữ liệu, mean là giá trị trung bình của tập dữ liệu, std (standard deviation) là độ lệch chuẩn của tập dữ liệu. Sau khi thực hiện đồng nhất lại giá trị và đơn vị dữ liệu RFM với kết quả như Hình 8:

CustomerID	Recency	Frequency	Monetary	Re_zs	Fre_zs	Mon_zs
12346.0	326	1	77183.60	2.343811	-0.451000	10.066906
12747.0	3	11	4196.01	-0.901742	0.938220	0.311637
12748.0	1	209	33719.73	-0.921838	28.444775	4.257675
12749.0	4	5	4090.88	-0.891694	0.104688	0.297586
12820.0	4	4	942.34	-0.891694	-0.034234	-0.123237
...	...	...	...	...	...	...

Hình 8: Minh họa kết quả Z-Score của Frequency.

Với phương pháp tính đơn giản nhưng lại mô tả lại được chính xác và gần hơn giá trị thực ban đầu của dữ liệu, điều này làm giảm đi khoảng cách chênh lệch lớn giữa các yếu tố trong phương pháp RFM và không làm thay đổi ý nghĩa ban đầu của dữ liệu. Giải thích cho kết quả này: trung bình tần suất mua hàng trên mỗi khách hàng là 4.24 lần. Khi đối chiếu với Fre_zs của khách hàng 12346 và 12748: Khách hàng 12346 có số lần mua hàng ít hơn so với mặt bằng chung (trung bình) là 0.45 lần. Đây là lý do giải thích cho sự xuất hiện của dấu âm trong giá trị này; Khách hàng 12748 có tần suất mua hàng cao hơn và nhiều hơn trung bình hơn 28 lần (28.44).

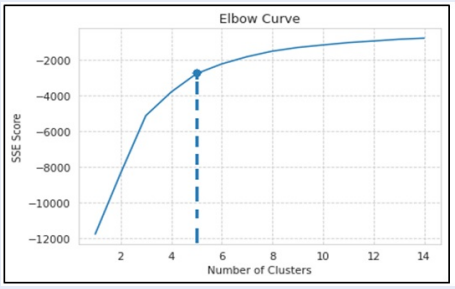
Lựa chọn số cụm tối ưu cho phương pháp K-means

Phương pháp Elbow được minh họa dưới dạng đồ thị đường cong với trục hoành là số K cụm (nghĩa là số

phân khúc khách hàng dựa trên giá trị từ mô hình dữ liệu RFM), trục tung là chỉ số SSE (Sum of Errors) – tức chỉ số đo lường sự khác biệt giữa các điểm trong cụm. SSE được tính bằng tổng các khoảng cách tính từ điểm dữ liệu trong cụm đến tâm cụm và lặp lại trên toàn bộ các cụm [8]. Công thức của SSE:

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_j} d\left(x_{ij}, m_i\right)^2 \quad (2)$$

với x là điểm dữ liệu, m là tâm cụm và k là số cụm. Tiến hành thực hiện phương pháp Elbow có số cụm từ 1 đến 20 trên mô hình RFM thu kết quả như sau:



Hình 9: Kết quả đồ thị đường SSE trong phương pháp Elbow (khuỷu tay).

Với đường SSE giống hình khuỷu tay, ta có điểm gập khuỷu tay với K = 5 (điểm giữa 4 và 6 trên trục hoành) sẽ là số cụm thích hợp. Giải thích cho điều này, khi càng tăng số cụm, giá trị của đường SSE cũng gần như tăng đều, nghĩa là sự khác biệt các điểm trong cụm gần như không có sự thay đổi. Hay nói cách khác đường SSE có xu hướng giảm dần độ dốc sau điểm "khuỷu tay" và ngay vị trí này trên đường SEE được xem như điểm tối ưu cho tham số đầu vào trong phương pháp gom cụm K-means.

Kiểm định chất lượng cụm với chỉ số Silhouette

Để đảm bảo được số nhóm khách hàng được phân tích là 5 từ phương pháp Elbow là tốt nhất, nghiên cứu tiến hành đo lường chỉ số Silhouette trên số cụm K=5 thu được kết quả trên Hình 10, với điểm số trung bình thu được là khoảng 0.6008 và cao nhất đối với tất cả số cụm trong khoảng từ 3 đến 9.

Điều này giải thích rằng, với số cụm là 5, khoảng cách từ các đối tượng trong cụm đến tâm cụm đã được tối ưu và không xảy ra hiện tượng lệch tâm cụm cho ảnh hưởng bởi giá trị Monetary như đã đề cập trước đó. Bên cạnh đó khi số cụm tăng dần từ 5 đến 9, đặc biệt là khi tăng dần từ 7, điểm Silhouette trung bình đã có sự giảm dần, điều này có điểm tương đồng với đường SSE tại Hình 9. Theo nghiên cứu của tác giả [9], kết quả

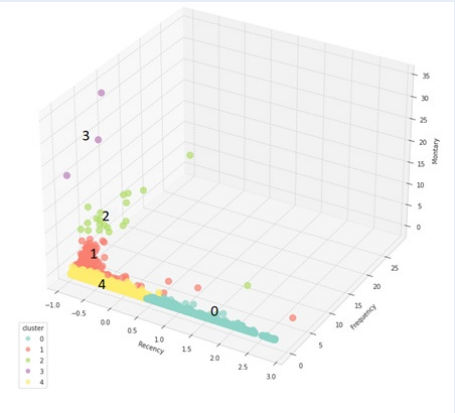
```
n_clusters = 2 Average silhouette_score: 0.8928856885439436
n_clusters = 3 Average silhouette_score: 0.5806043286271576
n_clusters = 4 Average silhouette_score: 0.595159847239341
n_clusters = 5 Average silhouette_score: 0.6008093327750194
n_clusters = 6 Average silhouette_score: 0.5963898159824265
n_clusters = 7 Average silhouette_score: 0.514184499119894
n_clusters = 8 Average silhouette_score: 0.483438600709789
n_clusters = 9 Average silhouette_score: 0.4845078649560471
```

Hình 10: Kết quả Silhouette trung bình với số cụm từ 2 đến 9.

xác định số nhóm khách hàng như trên có thể được đưa vào phân tích thực tế để tìm ra đặc điểm của các phân khúc khách hàng.

Gom cụm phân khúc khách hàng và trực quan hóa kết quả phân tích

Phân tích và trực quan kết quả phân cụm với biểu đồ phân tán (scatter) trên không gian ba chiều trên Hình 11. Kết quả thể hiện 5 cụm phân khúc khách hàng với đặc trưng có trong mỗi cụm.



Hình 11: Biểu đồ phân tán (Scatter plot) các nhóm khách hàng trên không gian ba chiều.

cluster	
0	962
1	289
2	20
3	3
4	2646

Hình 12: Số lượng phần tử (khách hàng) trong từng cụm (cluster).

Kết quả phân cụm được trực quan trên Hình 11 và Hình 12, với mật độ các điểm của cụm 0 và 4 là ổn

định nhất, tiếp theo đó là cụm 1 có độ ổn định thấp hơn với một vài điểm nằm khá xa tâm cụm. Riêng cụm 2 và cụm 3, thứ nhất là hai cụm này có số lượng phần tử cụm ít hơn tương ứng 20 và 3; thứ hai là khi xét đến một đặc điểm khác như tọa độ theo Frequency (đối với cụm 2) và Monetary (đối với cụm 3) có giá trị dương rất cao (lớn hơn 3), nên đây được xem là các dữ liệu ngoại lai (outlier) theo Quy tắc kiểm chứng với ba độ lệch chuẩn "68-95-99.7"[7]. Kết hợp hai điều kiện trên, ta có thể nhận định được là khó có thể gán nhãn nhóm khách với cụm 2 và 3. Trong kết quả dưới đây sẽ tập trung vào phân tích đặc điểm của các cụm tương ứng các nhóm khách hàng. Các tên gọi được gán nhãn cho các nhóm (phân khúc) khách hàng dưới đây dựa trên đặc điểm mô tả tứ phân vị và là nhãn mô tả một cách tổng quan nhất đặc điểm từng phân khúc khách hàng. Chi tiết đặc điểm từng nhóm khách hàng được gán nhãn và phân tích trong phần 4.

KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

Phân tích nhóm khách hàng trung thành (cụm 0)

	cluster	Recency	Frequency	Monetary
count	289.0	289.000000	289.000000	289.000000
mean	1.0	17.044983	17.169550	7710.260727
std	0.0	32.829654	7.361851	6932.397471
min	1.0	1.000000	3.000000	1296.440000
25%	1.0	3.000000	12.000000	3984.220000
50%	1.0	9.000000	15.000000	5528.480000
75%	1.0	18.000000	20.000000	8676.850000
max	1.0	373.000000	55.000000	50491.810000

Hình 13: Mô tả tứ phân vị của nhóm khách hàng trung thành.

Theo mô tả tứ phân vị trên Hình 13, phân khúc khách hàng này có số lượng là 289 khách hàng, chiếm 7.4% tổng số khách hàng. Từ kết quả trên có thể rút ra một số đặc điểm của nhóm khách hàng này. Trong đó, ngày mua hàng gần nhất nằm trong nhóm tốt nhất. Trung bình nhóm khách hàng này thường có số ngày mua gần nhất là 17 ngày; Tần suất mua hàng trung bình đạt 17 lần cao hơn rất nhiều so với hai nhóm còn lại; Và nhóm khách hàng có thể sẵn sàng chi nhiều tiền cho hoạt động mua sắm.

Trực quan hóa dữ liệu với biểu đồ trên Hình 14, biểu đồ cột khu vực bên trái thể hiện tỷ lệ phần trăm theo doanh thu (Monetary) và số lượng khách hàng ở từng phân khúc; khu vực bên phải là số liệu chi tiết với dạng cột (trục dọc trái) minh họa cho doanh thu (Monetary) và biểu đồ đường (trục dọc phải) là số khách hàng. Với biểu đồ mô tả ở Hình 14 đã cho thấy tổng lượng Monetary nhóm này đứng thứ 2 trong tất cả các phân khúc với khoảng 30.5% doanh thu.

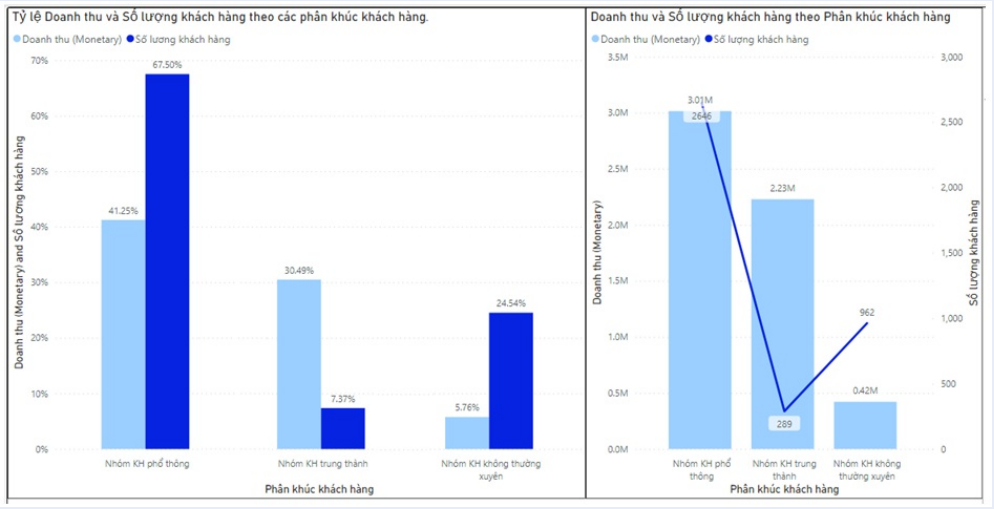
Với các đặc điểm về Recency, Frequency và Monetary ta có thể nhận thấy, không những đây là nhóm khách hàng trung thành và thậm chí có thể nhóm khách hàng mang lại tiềm năng lớn đối với doanh nghiệp. Mặc dù nhóm khách hàng này chỉ chiếm 7.37% nhưng doanh thu họ đem lại chiếm 30.49% và thường xuyên mua hàng trong năm (khoảng 17 lần/năm tức khá đều đặn hàng tháng). Cộng thêm một lợi thế đó chính là Recency của cụm này thấp tức họ vẫn có xu hướng quay lại vào các lần mua sắm tiếp theo.

Phân tích nhóm khách hàng phổ thông (cụm 4)

	cluster	Recency	Frequency	Monetary
count	2646.0	2646.000000	2646.000000	2646.000000
mean	4.0	45.334089	3.328042	1139.251294
std	0.0	36.127702	2.328849	1159.475694
min	4.0	1.000000	1.000000	6.200000
25%	4.0	17.000000	1.000000	356.737500
50%	4.0	34.000000	3.000000	748.155000
75%	4.0	67.000000	5.000000	1537.345000
max	4.0	158.000000	11.000000	12393.700000

Hình 15: Mô tả tứ phân vị của nhóm khách hàng phổ thông.

Đây là nhóm khách hàng có số lượng đông đảo nhất với tỷ lệ cao nhất 67.5%. Trong đó, theo như kết quả phân tích trên Hình 15, mức chi tiêu không quá cao và thấp hơn Nhóm khách hàng trung thành nhưng chiếm khá cao với 41.3% doanh thu; Recency và Frequency duy trì ở mức độ ổn định hơn, 50% nhóm khách hàng này có lượt mua hàng khoảng 3 lần trên 1 năm và 75% số lượng khách mua hàng 5 lần trong năm. Với nhóm khách hàng này, doanh nghiệp có thể tiếp tục cải thiện các chính sách bán hàng hiện tại để giữ chân nhóm khách hàng chủ lực này. Bên cạnh đó tìm ra những khách hàng tiềm năng trong nhóm này và thúc đẩy họ trở thành những khách hàng trung thành. Thêm vào đó, có một điểm cần được quan tâm với yếu tố Recency trong nhóm khách hàng phổ thông, đó là Recency tăng từ 34 lên 67 khi xét từ 50% lượng khách của nhóm này lên 75%.



Hình 14: Tỷ lệ khách hàng và doanh thu tương ứng theo từng phân khúc.

Phân tích nhóm khách hàng không thường xuyên (cụm 1)

Một số đặc điểm trong nhóm khách hàng này rất đáng được quan tâm so với hai nhóm khách hàng còn lại thể hiện trên Hình 16. Trong đó, mức độ chi tiêu là thấp nhất trong tất cả các phân khúc, chiếm khoảng 5.8% doanh thu; Tần suất mua hàng rất ít và có xu hướng duy trì thấp, cụ thể Rencency trung bình rất cao, đã hơn 247 ngày tương đương khoảng hơn 8 tháng không có hoạt động mua sắm; Frequency trung bình là rất thấp, khoảng 1.5 lần trong năm, thậm chí trong đó 75% khách hàng ở nhóm này chỉ mua sắm tối đa 2 lần trong năm. Đây có thể xem là nhóm khách hàng mang lại nhiều rủi ro cũng như những thách thức cho doanh nghiệp. Sự đóng góp giá trị của nhóm khách này là không cao và không nổi bật, nhưng lại chiếm $\frac{1}{4}$ số lượng khách hàng của cả doanh nghiệp.

	cluster	Recency	Frequency	Monetary
count	962.0	962.000000	962.000000	962.000000
mean	0.0	247.623701	1.559252	437.937402
std	0.0	65.800947	1.092555	522.775637
min	0.0	144.000000	1.000000	3.750000
25%	0.0	190.000000	1.000000	163.225000
50%	0.0	242.500000	1.000000	305.765000
75%	0.0	297.000000	2.000000	508.300000
max	0.0	374.000000	12.000000	7832.470000

Hình 16: Mô tả tứ phân vị của nhóm khách hàng không thường xuyên.

Phân tích tỷ lệ duy trì khách hàng (Customer Retention)

Phương pháp phân tích Cohort hay còn được hiểu là phân tích theo nhóm một cách tuần tự theo khoảng thời gian. Phương pháp phân tích này thường được ứng dụng để đo lường mức độ tương tác của người dùng theo thời gian[10]. Cụ thể trong bài toán phân tích tỷ lệ duy trì khách hàng (ký hiệu r) này, Cohort sẽ giúp tìm ra những khách hàng mới trong những tháng mới đối với từng tháng trong toàn bộ chu kỳ kinh doanh. Sau khi xác định được số lượng khách hàng trong từng chu kỳ mới ứng mỗi mốc thời gian (trong bài toán này, mỗi mốc thời gian và mỗi chu kỳ được tương ứng với mỗi tháng) kết hợp với công thức tỷ lệ duy trì thu được kết quả.

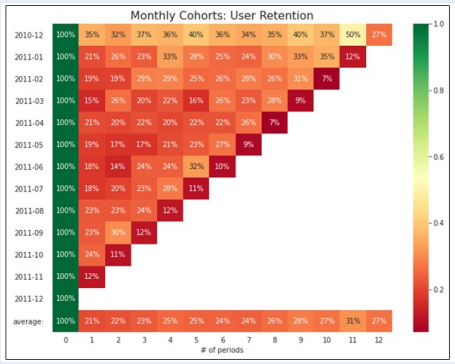
r = (Số khách hàng trong mỗi tháng tiếp theo)/(Tổng số khách hàng ban đầu) (3)

Với kết quả phân tích tỷ lệ duy trì khách hàng dưới dạng ma trận và biểu đồ nhiệt trên Hình 17, bao gồm:

- Quan sát theo chiều ngang biểu đồ, tỷ lệ duy trì khách hàng tính theo mốc thời điểm đầu tiên tháng 12/2010, lượng khách hàng đã sụt giảm mạnh ngay sau tháng đầu tiên và không có sự thay đổi đáng kể ở những tháng tiếp theo. Điểm nổi bật là ở tháng thứ 11 đã có sự tăng mạnh lên đến 50%. Tương tự cho những mốc thời gian khác, chúng ta hoàn toàn có thể kiểm tra lại tính khách quan ở những thời điểm khác nhau trong năm.
- Quan sát ở một khía cạnh khác đó là chiều dọc của biểu đồ, ta thu được tỷ lệ duy trì trung bình sau mỗi một chu kỳ (một tháng) với giá trị trung

bình (average). Cứ sau chu kỳ một tháng tính từ mọi mốc thời gian, ta chỉ duy trì được 21% khách hàng và giá trị này không có xu hướng tăng trong những chu kỳ tiếp theo (trung bình đạt 25% trên 12 tháng).

- Nhìn chung, tỷ lệ duy trì này chưa tốt. Tuy nhiên, một điểm sáng nhỏ là từ tháng thứ 7 trở đi, tỷ lệ này đã có sự cải thiện nhỏ. Trung bình tăng khoảng 4% so với tháng thứ 7, và tăng cao nhất tháng thứ 11 (hơn 7% so với tháng thứ 7). Như vậy nếu các chính sách hiện tại đang có dấu hiệu tốt, doanh nghiệp có thể duy trì. Bên cạnh đó, kết hợp với các kết quả phân khúc khách hàng trên, nhà quản lý có thể tăng cường thêm các chương trình chăm sóc khách hàng mới nhằm cải thiện cả hai kết quả và chỉ số này.



Hình 17: Trực quan hóa tỷ lệ duy trì khách hàng dưới dạng ma trận và biểu đồ nhiệt.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Mô hình nghiên cứu liên ngành được đề xuất với phương pháp RFM đã được thực nghiệm đầy đủ các bước với dữ liệu mua hàng lịch sử của khách hàng bao gồm ba yếu tố Recency, Frequency và Monetary được quan tâm. Nhằm khai thác hiệu quả mô hình dữ liệu RFM, phương pháp K-means được áp dụng kết hợp với phương pháp RFM để phân tích phân khúc khách hàng. Các yếu tố trong phương pháp RFM có sự liên kết lẫn nhau và thể hiện những ý nghĩa ở các khía cạnh khác nhau của khách hàng. Từ đó giúp chúng ta dễ dàng tìm ra các phân khúc khách hàng có hành vi mua sắm tương đồng nhau.

Với việc áp dụng các phương pháp, thuật toán như Silhouette, Z-Score, Quy tắc kiểm chứng giúp kết quả phân tích dữ liệu đảm bảo được độ tin cậy và chính xác cũng như có thể phát hiện ra những điểu bất thường (outlier) trong tập dữ liệu. Khi loại bỏ được những outlier sẽ làm cho kết quả cuối cùng tối ưu hơn. Từ những kết quả trên có thể thấy được vai trò của quá trình tiền xử lý dữ liệu là nhiệm vụ then chốt khi phân tích dữ liệu. Với các kết quả nghiên cứu đạt được từ bài báo đã giới thiệu không chỉ một mô hình nghiên cứu liên ngành mà còn được xem như các nguồn tham khảo trên nhiều góc nhìn, khía cạnh để giúp người quản lý có một bức tranh tổng quan nhiều chiều hơn với tình hình hiện tại của doanh nghiệp và giúp nhận diện rõ được khả năng của nghiên cứu liên ngành trong phân tích marketing nói riêng và trong lĩnh vực phân tích dữ liệu và khách hàng nói chung với các phương pháp học máy.

Bên cạnh đó, bộ dữ liệu đang được sử dụng để thực nghiệm mô hình trong nghiên cứu này là từ một cửa hàng bán lẻ ở Anh trong khoảng thời gian 2010-2011. Tuy nhiên, theo khảo sát bộ dữ liệu này về cấu trúc có sự tương đồng so với bộ dữ liệu bán lẻ tại các cửa hàng, doanh nghiệp bao gồm cả doanh nghiệp thương mại điện tử tại Việt Nam. Trong đó bao gồm đầy đủ các biến đặc trưng của dữ liệu giao dịch cần thiết cho mô hình nghiên cứu như để cập trong phần 2 và phần 3. Trong xu thế hiện nay ở các doanh nghiệp Việt Nam đã và đang sẵn sàng chuyển đổi số với lượng dữ liệu ngày càng tăng cao. Các hệ thống quản lý khách hàng ngày càng được tự động hóa. Tuy nhiên, hệ thống chủ yếu là ghi nhận dữ liệu giao dịch và thực hiện những thống kê định kỳ theo phương pháp truyền thống dẫn đến kết quả chưa đảm bảo được tính khách quan, chính xác và khó phân tích được hành vi mua sắm của khách hàng để có cơ sở xây dựng chiến lược tiếp cận khách hàng và bán hàng hiệu quả hơn. Vì vậy, bên cạnh đóng góp một nghiên cứu liên ngành trong bài báo, kết quả nghiên cứu còn giới thiệu một giải pháp phân tích với dữ liệu giao dịch lớn giúp tối ưu hiệu quả trong việc ra quyết định ở cấp quản lý.

Tuy nhiên, với kết quả phân cụm có được dựa trên yếu tố kỹ thuật, doanh nghiệp và người quản lý cần xác thực lại kết quả trên với góc nhìn của kinh doanh, kinh tế và thực tế để có thể ra quyết định tối ưu nhất. Một phương pháp, thuật toán hay một mô hình có thể chưa khái quát được toàn bộ những tổng quan trong doanh nghiệp hiện tại. Doanh nghiệp cần kết hợp nhiều hơn các phương pháp, mô hình phân tích khác để có sự hiểu biết sâu sắc về hành vi khách hàng để xây dựng những chiến lược tiếp cận và kinh doanh phù hợp. Từ dữ liệu về các phân khúc khách hàng và kết hợp với các nghiên cứu khác có thể xây dựng các chiến lược marketing và chăm sóc khách hàng riêng cho từng nhóm cũng như nguồn dữ liệu cho Bộ phận nghiên cứu và phát triển sản phẩm (R&D).

DANH MỤC TỪ VIẾT TẮT

Machine Learning: Phương pháp học máy.
K-means: Một trong những thuật toán được sử dụng trong lĩnh vực Machine Learning thuộc mô hình Học không giám sát.
Cluster: Cụm hay nhóm, gồm các điểm dữ liệu trong phân tích cụm.
Outlier: Dữ liệu ngoại lai.
RFM: Mô hình được cấu thành từ ba yếu tố Recency – Frequency – Monetary.
Recency: Thời gian của lần cuối gần nhất mà khách hàng đã mua hàng.
Frequency: Tần suất mua hàng của khách hàng.
Monetary: Tổng lượng tiền mà khách hàng đã chi tiêu cho toàn bộ hoạt động mua sắm.
Z-Score: Phép đo số mô tả mối quan hệ của giá trị với giá trị trung bình của một nhóm giá trị. Z-Score được hoạt động dựa theo độ lệch chuẩn so với giá trị trung bình.

XUNG ĐỘT LỢI ÍCH

Nhóm tác giả xin cam đoan rằng không có bất kì xung đột lợi ích nào trong công bố bài báo.

ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Toàn bộ nội dung bài báo chỉ do nhóm tác giả thực hiện. Các tác giả có đóng góp như nhau trong quá trình nghiên cứu về ý tưởng, mục tiêu, phương pháp nghiên cứu, đề xuất mô hình, phân tích dữ liệu, đánh giá và thảo luận kết quả.

TÀI LIỆU THAM KHẢO

1. Christy AJ, et al. RFM ranking - An effective approach to customer segmentation. Journal of King Saud University - Computer and Information Sciences; 2018 p1-7;Available from: https://doi.org/10.1016/j.jksuci.2018.09.004.
2. Miglautsch JR. Thoughts on RFM scoring. Journal of Database Marketing. 2000; 8(1):67-72;Available from: https://doi.org/10.1057/palgrave.jdm.3240019.
3. Anitha P, Patil MM. RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud University - Computer and Information Sciences; 2019. p.1-8;Available from: https://doi.org/10.1016/j.jksuci.2019.12.011.
4. Alpaydın E. Introduction to Machine Learning (Adaptive Computation and Machine Learning series). 2nd ed. Cambridge: The MIT Press; 2009. p.1-19;.
5. Muller A, Guido S. Introduction to Machine Learning with Python: A Guide for Data Scientists. 3rd ed. Boston: O'Reilly Media; 2017. p.170-183;.
6. Chen D, Sain SL, Guo K. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing and Customer Strategy Management; 2012. 19(3). p.198-208;Available from: https://doi.org/10.1057/dbm.2012.17.
7. Salkind NJ. Statistics for People Who (Think They) Hate Statistics. 6th ed. Los Angeles: SAGE Publications, Inc; 2016. p.202-220;.
8. Patel E, Kushwaha DS. Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. Procedia Computer Science; 2020. 171(2020). p.158-167;Available from: https://doi.org/10.1016/j.procs.2020.04.017.
9. Larose DT. Data Mining and Predictive Analytics (Wiley Series on Methods and Applications in Data Mining). 2nd ed. Hoboken: John Wiley & Sons; 2015. p.582-589;.
10. Scroll A, Yoskovitz B. Lean Analytics: Use Data to Build a Better Startup Faster. 1st ed. Treseler M, editor. Cambridge: O'Reilly Media, Inc.; 2013. p.24-26;.

Science & Technology Development Journal – Economics - Law and Management, 6(1):2005-2015

Open Access Full Text Article

Research Article

An interdisciplinary research between analyzing customer segmentation in marketing and machine learning method

Ho Trung Thanh*, Nguyen Dang Son

ABSTRACT

Customer segmentation is one of the key factors in managing customers and building the appropriate marketing strategies. Segmenting customer groups will help managers understand the characteristics of their customers or consumer behaviors, thereby reaching the right target customers, retaining customers (Customer Retention), increasing revenue and competitive advantages of the business. However, finding the right customer groups is a challenge that businesses need to solve on a solid and reliable basis. Along with the support from current technology solutions such as Customer Relationship Management (CRM) and the application of algorithms and methods including both qualitative and quantitative research to enable businesses to cluster customer groups in marketing analysis. This article concentrates on introducing a hybrid model that combines RFM (Recency, Frequency, Monetary) and Machine Learning to analyze customer segmentation. The study was conducted through an empirical method on a dataset with 541,909 transactions of online retail stores, clustering 5 customer segments with the characteristics of each cluster being tested for quality demonstrating the effectiveness and applicability of the study.

Key words: Customer segmentation, RFM, Machine Learning, clustering, customer retention rate

University of Economics and Law, VNU-HCM, Vietnam

Correspondence

Ho Trung Thanh, University of Economics and Law, VNU-HCM, Vietnam

Email: thanhht@uel.edu.vn

History

- Received: 08/6/2021
- Accepted: 20/8/2021
- Published: 04/9/2021

DOI : 10.32508/stdjelm.v6i1.850

Check for updates

VNU-HCM Press

Cite this article : Thanh H T, Son N D. **An interdisciplinary research between analyzing customer seg-mentation in marketing and machine learning method**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 6(1):2005-2015.