Journal of Mining and Earth Sciences

Website: http://jmes.humg.edu.vn

Analyzing customer sentiments using K-means algorithm

Trung Kien Pham *, Thang Duc Nguyen, Chien Van Le, Thuong Van Nguyen

Faculty of Economics and Business Administration, Hanoi University of Mining and Geology, Vietnam

ARTICLE INFO

Article history:
Received 18th Aug. 2020
Accepted 24th Sept. 2020
Available online 31st Oct. 2020

Keywords:
Clustering algorithm,
Customer segmentation,
K-Means clustering,
Potential customer.

ABSTRACT

Customer segmentation is the process of dividing customers based on common characteristics such as their behavior, buying habits and service usage,... so that companies can market for each group customers more effectively and appropriately. The paper analyzes customer cluster segmentation via the K-Means clustering methods of a business sector. The research was conducted on 272 customers with characteristics of age, income and expense score. The research results are divided into 2 target customer clusters, promising to help care and marketing customers more effectively; Help business units to have appropriate marketing strategies to reduce costs and increase efficiency.

*Corresponding author
E-mail: phamkientrung@humg.edu.vn
DOI: 10.46326/JMES.KTQT2020.19

Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất
Trang điện tử: http://tapchi.humg.edu.vn

Ứng dụng thuật toán K-Means trong phân cụm khách hàng mục tiêu

Phạm Kiên Trung *, Nguyễn Đức Thắng, Lê Văn Chiến, Nguyễn Văn Thưởng

Khoa Kinh tế và Quản trị kinh doanh, Trường Đại học Mỏ - Địa chất, Việt Nam

THÔNG TIN BÀI BÁO

Quá trình:
Nhận bài 18/8/2020
Chấp nhận 24/9/2020
Đăng online 31/10/2020

Từ khóa:
K-Means clustering,
Khách hàng mục tiêu,
Phân cụm khách hàng,
Thuật toán phân cụm.

TÓM TẮT

Phân cụm khách hàng (customer segmentation) là quá trình phân chia khách hàng dựa trên các đặc điểm chung như hành vi, thói quen mua sắm và sử dụng dịch vụ của họ,... để các công ty, doanh nghiệp có thể tiếp thị cho từng nhóm khách hàng một cách hiệu quả và phù hợp hơn. Bài báo nghiên cứu phân khúc cụm khách hàng thông qua phương pháp phân cụm K-Means (K-Means clustering methods) của một cơ sở kinh doanh. Nghiên cứu được thực hiện trên 272 khách hàng với các đặc điểm về độ tuổi, thu nhập và điểm chi tiêu. Kết quả nghiên cứu đã chia thành 2 cụm khách hàng mục tiêu, hứa hẹn sẽ giúp việc chăm sóc, tiếp thị khách hàng hiệu quả hơn; giúp đơn vị kinh doanh có những chiến lược marketing phù hợp giảm chi phí và tăng hiệu quả.

1. Mở đầu

Phân cụm khách hàng là quá trình phân chia khách hàng thành nhiều cụm/nhóm có chung sự tương đồng theo những tiêu chí như giới tính, tuổi tác, sở thích, thu nhập và thói quen chi tiêu, hành vi mua sắm,... để doanh nghiệp có phương thức tiếp thị hiệu quả. Khi thực hiện được phân cụm khách hàng giúp đơn vị giải quyết đúng các yêu cầu của từng khách hàng, giúp tăng lợi nhuận, giữ chân các khách hàng quan trọng, cũng như thực hiện các chiến dịch, chiến lược marketing hiệu quả hơn (Khajvand and Tarokh, 2011).

Hiện nay, có nhiều phương pháp giúp doanh nghiệp thực hiện việc phân cụm khách hàng mục tiêu dựa trên những hiểu biết về hành vi (behavior), thói quen (habits), sở thích (preferences) của khách hàng tiềm năng như K-Means, Mean-Shift, Density-Based Spatial, Expectation-Maximization, Agglomerative Hierarchical Clustering (Chen et al., 2012).

Trong phạm vi nghiên cứu, các tác giả lựa chọn phương pháp phân cụm theo thuật toán K-Means, đây là thuật toán quan trọng và được sử dụng phổ biến trong các nghiên cứu hiện nay (Chapman and Feit 2019).

Bài báo thu thập số liệu từ 272 khách hàng tại showroom ô tô với các thông tin thu thập về dòng xe quan tâm, kênh tiếp cận khách hàng, độ tuổi, thu nhập bình quân và điểm chi tiêu để thực hiện phân cụm theo thuật toán K-Means.

*Tác giả liên hệ
E - mail: phamkientrung@humg.edu.vn
DOI: 10.46326/JMES.KTQT2020.19

2. Phương pháp nghiên cứu

- Phương pháp thống kê: Thu thập và xử lý số liệu, điều tra chọn mẫu được nhóm tác giả sử dụng để có được hình ảnh tổng quát về mẫu nghiên cứu.

- Phương pháp phân cụm K-means: Thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác đinh trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất.

Thuật toán K-Means thực hiện qua các bước chính sau (Hình 1).

1. Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.

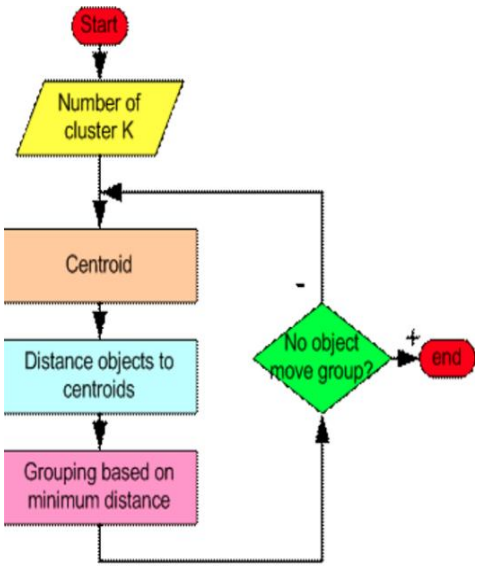
Trong nghiên cứu, để xác định được số cụm tối ưu nhóm sử dụng phương pháp Elbow. Tiến hành chạy phân cụm trên tập dữ liệu cho một phạm vi giá trị của k (k từ 1 đến 10), tại vị trí k nào tạo thành khúc cua khuỷa tay thì chọn ra k tối ưu. (Shmueli et al., 2017).

2. Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean).

3. Nhóm các đối tượng vào nhóm gần nhất.

4. Xác định lại tâm mới cho các nhóm.

5. Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng. (Zakrzewska and Murlewski, J, 2005).



Hình 1. Các bước thực hiện K-Means Clustering (Zakrzewska and Murlewski, 2005).

3. Dữ liệu nghiên cứu

Nghiên cứu thu thập thông tin của 272 khách hàng tại điểm bán hàng của công ty Trường Hải Auto, các thông tin được tập hợp gồm 6 cột: mã khách hàng ID, Chủng loại xe quan tâm, Kênh thông tin phản hồi, độ tuổi, thu nhập bình quân/tháng và điểm chi tiêu. Dưới đây là mô tả 1 phần dữ liệu.

ID	LOAIXE	KENH	AGE	TNHAP	DIEM
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>
1	...06482	MORNING-SI-AT-1.25	Showroom	20	9 55
2	...6353	CERATO-1.6-AT	Showroom	35	8.9 78
3	...6467	CERATO-1.6-AT	Showroom	33	9.7 50
4	...6486	CERATO-1.6-AT	Điện thoại	20	8.7 52
5	...6487	SEDONA-2.2-DAT	Showroom	34	9.2 53
6	...6488	SEDONA-2.2-DAT	Showroom	52	8.7 45

a, Mô tả độ tuổi của nhóm khách hàng

Độ tuổi bình quân của khách hàng là 36,1 tuổi, khách hàng có tuổi lớn nhất là 52 tuổi, nhỏ nhất là 20 tuổi, với độ lệnh chuẩn là 6,7 tuổi.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   33.00   35.00   36.06   40.00   52.00
Sd = 6.722813
```

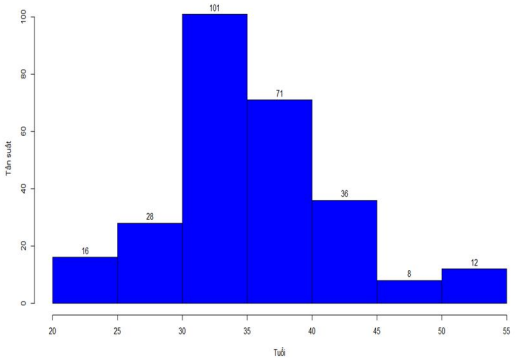
Hình 2 và 3 thể hiện phân bố độ tuổi qua biểu đồ cột và biểu đồ hộp. Với Hình 2 cho thấy độ tuổi chủ yếu là từ 33 đến 40 tuổi, Hình 3 thể hiện độ tuổi trung bình, trung vị, bách phân vị 25% và 75%, biểu đồ cho thấy có 4 giá trị ngoại vi.

b, Mô tả thu nhập của khách hàng

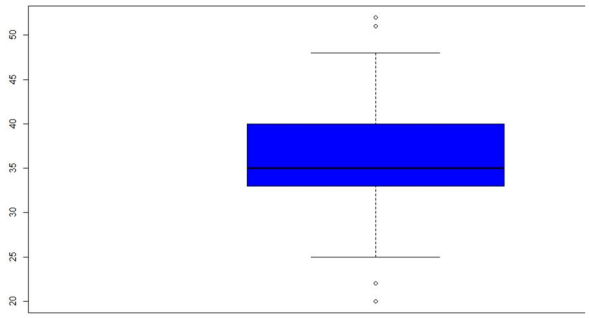
Thu nhập bình quân của khách hàng là 9,95 triệu đồng/tháng, trong đó người thấp nhất là 7,5 triệu đồng/tháng và cao nhất là 14 triệu đồng/tháng. Nhìn chung, nhóm khách hàng quan tâm đến mua xe có mức thu nhập trung bình khá trở lên. Thu nhập của khách hàng không có giá trị nào nằm ngoài khoảng bách phân vị 25% và 75% thể hiện tại Hình 4.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.500   8.800   9.500   9.952  11.200  14.000
```

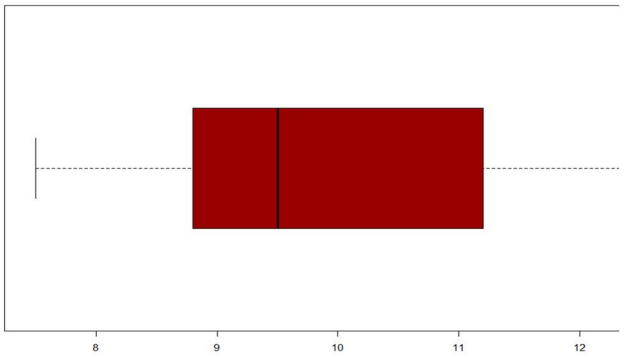
Hình 5 cho thấy rõ về phân bố thu nhập của khách hàng tập trung ở mức từ 8 triệu đồng/tháng đến mức 11 triệu đồng/tháng. Mức thu nhập trên



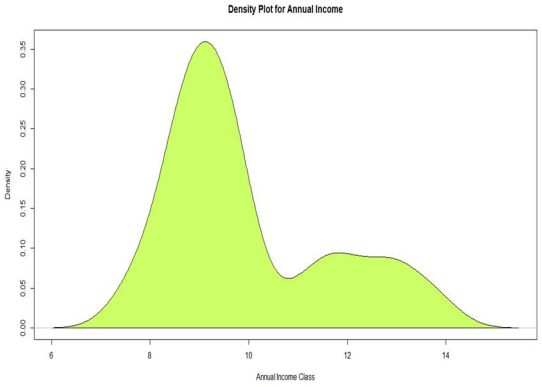
Hình 2. Biểu đồ phân bố theo độ tuổi khách hàng.



Hình 3. Biểu đồ phân bố theo độ tuổi khách hàng.



Hình 4. Biểu đồ hộp mô tả thu nhập của khách hàng.



Hình 5. Phân bố thu nhập của khách hàng.

12 triệu đồng/tháng cũng tương đối nhiều khách hàng.

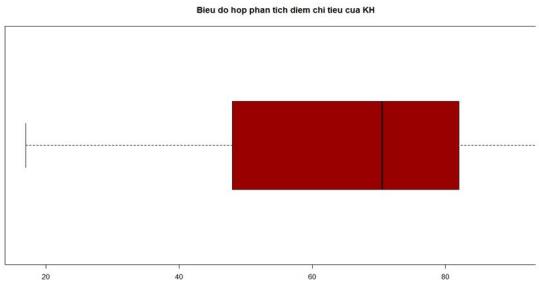
c, Mô tả điểm chi tiêu

Điểm chi tiêu cho biết mức độ chi tiêu so với thu nhập của từng khách hàng, được đánh giá từ 0 đến 100 điểm. Với dữ liệu, Hình 7 thể hiện khách hàng có điểm chi tiêu cao nhất là 95 điểm, thể hiện mức sẵn sàng chi tiêu rất cao. Khách hàng thấp nhất là 17 điểm và trung bình là 66,28 điểm, điểm trung vị là 70,5 điểm thể hiện tại Hình 6. Nhìn chung, nhóm khách hàng có điểm chi tiêu ở mức trên trung bình so với thu nhập bình quân chung.

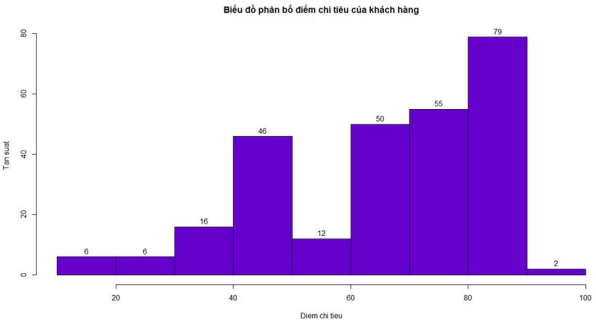
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.00   48.00   70.50   66.28   82.00   95.00
```

4. Kết quả nghiên cứu

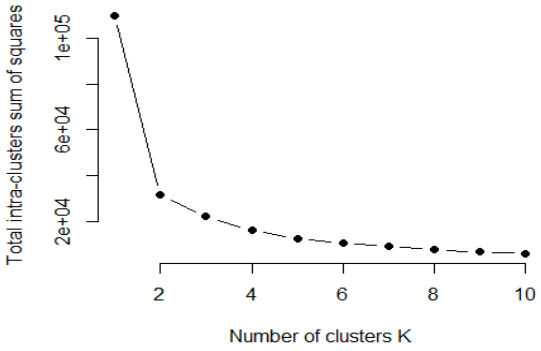
1. Bằng phương pháp Elbow Method: Nghiên cứu xác định số cụm tối ưu để phân bổ khách hàng là 2 cụm Hình 8a và 8b. Đây là số cụm nên phân bổ theo phương pháp này (Shmueli et al., 2017). Tuy nhiên, nếu cần doanh nghiệp có thể phân cụm với k=3, k=4,...



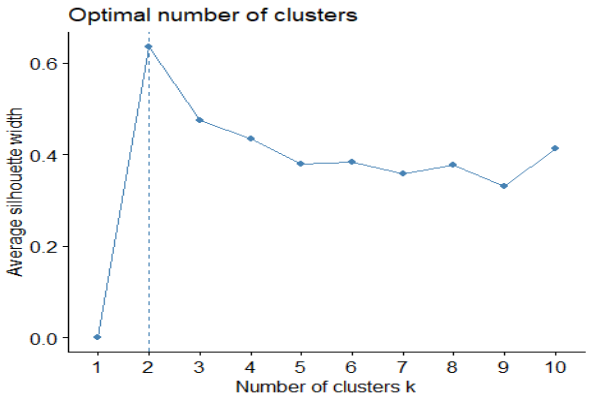
Hình 6. Biểu đồ hộp mô tả điểm chi tiêu của khách hàng.



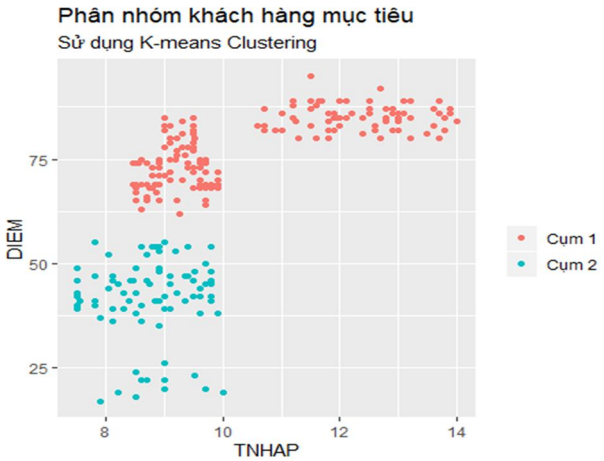
Hình 7. Biểu đồ cột mô tả điểm chi tiêu của khách hàng.



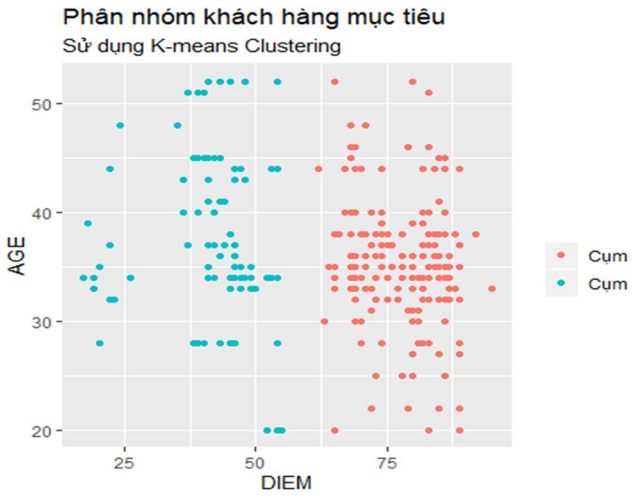
Hình 8a. Xác định số lượng cụm tối ưu theo phương pháp Elbow.



Hình 8b. Xác định số lượng cụm tối ưu theo phương pháp Elbow.



Hình 9. Phân cụm khách hàng theo điểm chi tiêu và thu nhập.



Hình 10. Phân cụm khách hàng theo điểm chi tiêu và độ tuổi.

2. Sau khi xác định được số lượng cụm tối ưu là 2, nhóm nghiên cứu thực hiện phân vùng ngẫu nhiên khác nhau 50 lần (Chapman and Feit, 2019).

3. Thực hiện số lần lặp 100 lần để chọn kết quả tốt nhất. Cụ thể:

K-means clustering with 2 clusters of sizes 86, 186

Cluster means:

	AGE	TNHAP	DIEM
1	37.63953	8.753607	41.41860
2	35.32258	10.506385	77.76882

Kích thước cụm 1 là 186 đối tượng và cụm 2 là 86 đối tượng quan sát.

Tâm điểm cụm 1 (centroid cluster 1): độ tuổi 37,6 tuổi; thu nhập 8,75 triệu đồng/tháng; điểm chi tiêu 41,4 điểm.

Tâm điểm cụm 2 (centroid cluster 2): độ tuổi 35,3 tuổi, thu nhập 10,5 triệu đồng/táng; điểm chi tiêu 77,7 điểm.

Within cluster sum of squares by cluster:

[1] 13458.24 18036.95

(between_SS / total_SS = 71.3 %)

Như vậy, 71,3% sự khác biệt của khách hàng có thể được giải thích bằng sự khác biệt trong mỗi nhóm.

4. Mô phỏng kết quả phân cụm

Qua Hình 9 cho thấy 2 cụm khách hàng khách nhau về thu nhập và điểm chi tiêu:

Cụm 1: Cụm khách hàng màu đỏ thuộc nhóm khách hàng có điểm chi tiêu cao (trên 60 điểm) và có thu nhập từ 7,5 triệu đồng đến 14 triệu đồng/tháng.

Cụm 2: Cụm khách hàng màu xanh thuộc nhóm có điểm chi tiêu thấp (dưới 60 điểm) và có thu nhập tập trung từ 7,5 đến 10 triệu đồng/tháng.

Hình 10, nhóm tác giả phân 2 cụm khách hàng theo tiêu thức điểm chi tiêu và độ tuổi.

Cụm 1: Cụm khách hàng màu đỏ thuộc nhóm khách hàng có điểm chi tiêu cao (trên 60 điểm) và độ tuổi không tập trung

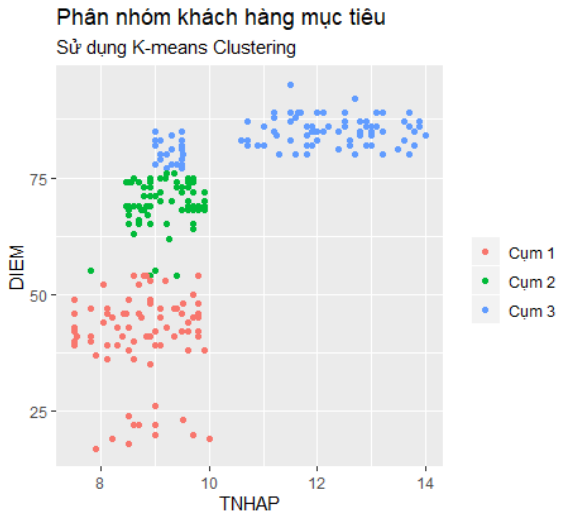
Cụm 2: Cụm khách hàng màu xanh thuộc nhóm có điểm chi tiêu thấp (dưới 60 điểm) và độ tuổi không tập trung.

Thực tế tại đơn vị kinh doanh này, việc phân cụm khách hàng thường được phân loại thành 3 loại: khách hàng nóng, khách hàng ấm, khách hàng lạnh. Nhóm nghiên cứu tiếp tục tiến hành thử phân cụm với k=3, dù không đồng nhất với các phân loại của đơn vị, cũng cho công ty này cái nhìn tốt hơn, Hình 11.

Như vậy, với các đặc điểm của nhóm khách hàng, thì việc phân cụm theo điểm chi tiêu và thu nhập cho doanh nghiệp thấy rõ ràng hơn cụm khách hàng mục tiêu, và theo thuật toán K-Means thì việc phân thành 2 cụm khách hàng là tối ưu.

5. Kết luận

Với sự trợ giúp của việc phân cụm, chúng ta có thể hiểu các thông tin khách hàng tốt hơn nhiều, giúp bộ phận chăm sóc khách hàng đưa ra quyết định cẩn thận. Ngoài ra, với việc xác định khách hàng, các công ty có thể đưa ra các sản phẩm và dịch vụ nhằm mục tiêu khách hàng dựa trên một số thông số như thu nhập, tuổi tác, mô hình chi tiêu,...



Hình 11. Phân cụm khách hàng theo điểm chi tiêu và thu nhập với k=3.

Tuy nhiên, việc phân cụm theo thuật toán K-Means cần xác định rõ số lượng cụm cần phân bố ngay từ ban đầu, đây cũng gây khó khăn khi thực hiện phương pháp này.

Bên cạnh đó, nghiên cứu sẽ đầy đủ hơn nếu được thu thập các thông tin về hành vi, thói quen và sở thích của khách hàng.

Những đóng góp của tác giả

Xây dựng ý tưởng, Lựa chọn đối tượng nghiên cứu, phương pháp nghiên cứu, viết bài báo: Phạm Kiên Trung; Phân tích dữ liệu: Nguyễn Đức Thắng; Phân tích dữ liệu kiểm chứng dữ liệu thu thập và kết quả nghiên cứu: Lê Văn Chiến; Thu thập, phân nhóm và tổng hợp số liệu: Nguyễn Văn Thưởng.

Tài liệu tham khảo

Chapman, C., & Feit, E. M, (2019). *R for marketing research and analytics*. New York, NY: Springer.

Chen, D., Sain, S. L., & Guo, K, (2012). *Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining*. Journal of Database Marketing & Customer Strategy Management, 19(3), 197-208.

Khajvand, M., & Tarokh, M. J, (2011). *Estimating customer future value of different customer segments based on adapted RFM model in retail banking context*. Procedia Computer Science, 3, 1327-1332.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C, (2017). *Data mining for business analytics: concepts, techniques, and applications in R*. John Wiley & Sons.

Zakrzewska, D., & Murlewski, J, (2005). *Clustering algorithms for bank customer segmentation.* In 5th International Conference on Intelligent Systems Design and Applications (ISDA'05) pp. 197-202. IEEE.