Customer Segmentation Based on RFM Model Using K-Means, Hierarchical and Fuzzy C- Means Clustering Algorithms

Research · August 2019

DOI: 10.13140/RG.2.2.15379.71201

CITATIONS
0

READS
1,999

2 authors, including:

Surefunmi Idowu
National College of Ireland
5 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Customer Segmentation Based on RFM Model Using K-Means, Hierarchical and Fuzzy C-Means Clustering Algorithms

Oluwasurefunmi Idowu
National College of Ireland
MSc. Data Analytics
x18158188

Adithya Annam
National College of Ireland
MSc. Data Analytics
x18134963

Eashwar Rangarajan
National College of Ireland
MSc. Data Analytics
x18140386

Srivatsav Kattukottai
National College of Ireland
MSc. Data Analytics
x18145922

***Abstract* - E-commerce is the way of purchasing products via an online platform. With tons of technological improvements, e-commerce industry has seen a vast improvement in recent years. Firms in this domain aim at increasing profit by analyzing the previous customer purchase data. Therefore, by deploying a data mining technique, customer data can be analyzed which helps an organization understand their customers and take necessary decisions. This study presents the analysis of three clustering algorithms - K-means Clustering, Fuzzy C-means Clustering, and Hierarchical Clustering, which were built using R programming language to understand customer behavior and segment them based on the purchase pattern through Recency, Frequency and Monetary (RFM) model. The analysis was done using the KDD methodology approach and all three algorithms were evaluated internally and externally using various validation measures like Silhouette Width, Dunn Index, Adjusted Rand Index and Variation Index. The k-means and fuzzy c-means results were similar as five clusters were realized and named, with the largest cluster being the high spending customers. On the other hand, the hierarchical produced two clusters. With a Dunn Index of 1.58, it was concluded that the hierarchical model performed best.**

Keywords: **Customer Segmentation, Clustering, RFM, K-means, Hierarchical, Fuzzy C-means.**

I. INTRODUCTION

Business is always an outcome of supply and demand. Any industry runs around its customers and consumers. Hence, it is necessary for firms to understand their customers and decipher their wants and needs. The competitiveness in every industry makes the organization to stay up-to-date and provide the best possible service to its customers. Over the past few decades, with advancements in technology, the digitalization of businesses has grown manifold. This has paved way for the availability of large volumes of data, thereby allowing analysis to be performed by categorizing and grouping the customers and in turn classifying their needs. A lot of technological transformations have happened in the last three decades, causing the internet to serve as a platform for running businesses. This gave birth to the term e-commerce. While analyzing this data is an interesting topic, it is more fascinating to develop generic models which would perform this operation on any similar data. The e-commerce industry is likely to expand dramatically in the coming years. By developing a machine learning model that can help us understand customers and their purchase pattern, organizations can make critical business decisions to attract more customers and serve them better, thereby gaining a competitive advantage. Consequently, it has attracted a lot of research by marketers and data analysts who have discovered several machine learning models to aid profit making in this domain. One of the widely known techniques is clustering, which is simply defined by [19, p. 286] as an unsupervised classification model which finds patterns in a dataset. These patterns can be found using some analysis like basket market analysis [11] and RFM model [7].

This analysis is inspired from [2] which develops a k-means clustering model to understand the customer purchase pattern and sequence. Besides, the research tries to find the characteristics of customers by segmenting them based on a phenomenon called Recency, Frequency and Monetary model (RFM). By using the cluster node in Statistical Analysis Systems (SAS) Enterprise Minor, the k-means clustering algorithm was employed. It aims to develop three different clustering algorithms: k-means, fuzzy c-means and hierarchical clustering for analyzing customer purchase pattern. These customers will be clustered based on the RFM model and a comparison will be done, choosing the best technique.

The next section of this paper highlights the research question and objectives. Afterwards, the work is justified based on numerous literatures published in past and recent times. Subsequently, the methodology and necessary steps are explained, followed by the evaluation metrics. Conclusively, the paper is summarized, and possible future work is proposed.

II. RESEARCH QUESTION

"Can hierarchical clustering model perform better than k-means and fuzzy c-means algorithm when analyzing customer purchase pattern using RFM analysis?"

Hence, the objectives of this research are the following:

- To implement k-means, fuzzy c-means and hierarchical clustering in analyzing customer purchase pattern using RFM model.
- To evaluate each model internally and externally.
- To derive customer segments using the mentioned clustering techniques.
- To pick the best performing clustering model.

III. LITERATURE REVIEW

A. Introduction

Clustering technique is generally used to find similarity in a dataset [1]. It has been applied by so many authors through different methods and in a variety of fields. For instance, in the medical field, doctors use it to diagnose patients with similar diseases thereby prescribing same treatment for them. In the marketing field, retailers use it to understand customer buying behavior in order to enhance their offers and make more profit. In the banking field [13], it is used to gain key insights into customer value and determine strategies for customer segments. These days, companies don't mind investing a large amount of money in developing marketing strategies, which aim towards customer segmentation, thereby maximizing customer satisfaction and retention. In order to do this, they analyze big customer data to give them insights, which enables them to make stronger business decisions, thereby gaining competitive advantage in the market [14]. For customer segmentation to be useful, it must be easily understood, identifiable, relevant, significant and reachable[1]. Customers could be segmented based on their visits [8][11], purchase history and details [15], payment details and so on.

B. Clustering Techniques

There are two main methods of clustering; hard or crisp clustering and soft clustering [1]. Hard clustering includes k-means (used for numeric data) and k-mode (for categorical data), while soft methods include fuzzy c-means [2], possibilistic c-means [10], and evidential c-means [3]. K-means clustering is one of the most popular clustering techniques. It is only useful for a dataset with numeric variables. Since it was specially designed that way, it cannot cluster objects with categorical variables. Although k-means is easy to implement and obtains good results, it has some limitations highlighted by [19, p. 289] in the sense that it does not have the tendency to find the optimum number of clusters because it uses "an element of random chances." It was also stated that it is not ideal for clusters which have considerable differences in density. More so, research has proven that k-means clustering model doesn't work well with clusters of varying size and density [3]. Therefore, in cases where we cannot define a k-value, hierarchical clustering model can be used, which provides more intuitive results. Subsequently, researchers [5] have tried to combine k-means with other clustering algorithms in order to yield better performance. According to [5, p. 609], clustering algorithms can be classified into four parts, which includes k-means, hierarchical, density based and self-organization maps. Furthermore, in the literature, a hybrid clustering technique was implemented, which is a combination of k-means and hierarchical clustering, aiming to organize large amounts of text data into significant clusters. It was realized that in hierarchal clustering, prototype vectors spotted out and were discovered to hide the noise from the technique, but no comparison was found within the techniques. However, some have done the work to compare and contrast several clustering algorithms [4], the unique combination of k-means, fuzzy c-means and hierarchical clustering have not been evaluated together.

One important issue in clustering is the fact that the data is not completely understood. This phenomenon is called imperfect knowledge of the dataset. To handle these imperfections, several uncertainty theories have been recommended, one of which is fuzzy sets [16]. Taking a further look into clustering methods, [10] proposed a possibilistic meta-clustering model where two granules are created for customers and the products bought frequently. Data inputs were derived from the database of retail store by segmentation process of customers and their purchasing pattern and the model was implemented with the help of k-modes clustering algorithm thus making better decision-making results. Although, [5] evaluated clustering techniques using precision, accuracy and recall, [16] used kappa index, rand index. Some papers [11] use market basket analytics method for the analysis of customers shopping preferences by segmenting customers' visits with help of feature selection approach, collaborative filtering and association rules [8].

C. Recency Frequency and Monetary (RFM) Model

Originally known as RFM analysis, is a widely known standard technique used to evaluate customer lifetime value, especially in the retail industry. It was first proposed by [17]. Recency stands for the last purchase date within a specific period, Frequency stands for the number of purchases within a specific period, and Monetary is the value of purchases within a specific period. The RFM model is calculated thus:

$$RFM\ score = (rs \times rw) + (fs \times fw) + (ms \times mw)$$

where:
rs = recency score and rw = recency weight
fs = frequency score and fw = frequency weight, and
ms = monetary score and mw = monetary weight

The model has been used by several authors [7][20]. Although [7] developed a clustering model to identify customer segments of one of Turkey's largest sport retail stores using two-step cluster analysis and k-means clustering, [20] used a combination of regression and k-means. Subsequently in [7], the RFM values were used as indicators to cluster customers. For two-step

[1] https://askpivot.com/blog

clustering, the number of clusters was not fixed, whereas in k-means clustering, the best model was built when the value of k was 4. The same approach has been used in other retail stores, but in different countries like India [12] and Romania [6].

D. Conclusion

In this paper, we compare three clustering techniques such as k-means and hierarchical clustering approach to determine customer segments using data of a Brazilian retail store, sourced from Kaggle. The clusters are evaluated internally and externally, and the best algorithm is picked based on the different parameters.

IV. METHODOLOGY AND IMPLEMENTATION

The methodology followed in this research is KDD (Knowledge Discovery of Databases). KDD is chosen, as this research follows a data-driven approach and this study aims at acquiring knowledge from database which is to classify the customers based on purchase pattern. The purchase pattern is analyzed using Clustering models by identifying the Recency, Frequency and Monetary value of customers. This research is implemented through various stages that are explained below:

- Data Collection.
- Data pre-processing.
- Data transformation.
- Data mining.
- Evaluation.

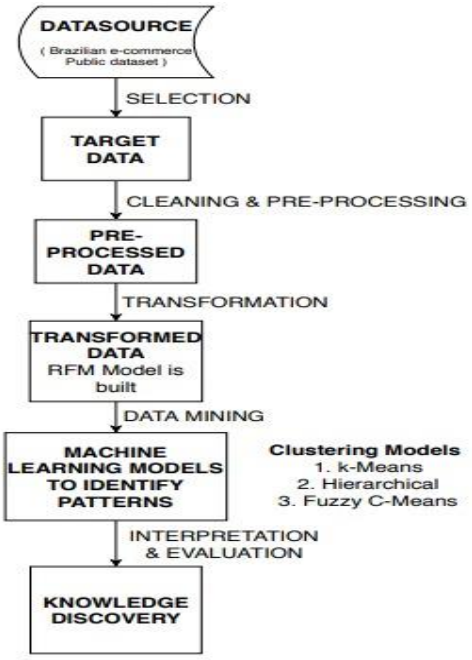

Fig. 1. KDD Process Flow

A. Data Selection:

To analyze the purchase pattern of customers in retail shops, dataset of a Brazilian retail store (Olist) has been downloaded from Kaggle that contains customer location, products ordered and payment related information. Since the purchase pattern can be analyzed by considering how recent and frequent customers have placed orders, as well as the price of the product purchased, only 3 datasets were used for segmentation purpose that contains customer details, orders and payment data.

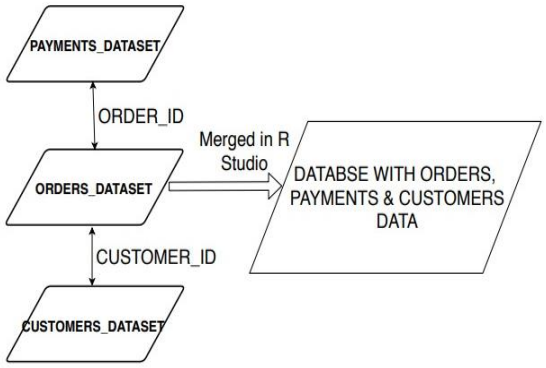

Fig. 2. Data Architecture

Figure 2 explains the merging of the individual datasets to form the database, which is done in R Studio. These tables were joined based on order_id and customer_id.

B. Data Cleaning and Pre-processing:

The target data is cleaned before proceeding further. Cleaning here includes, handling missing values, detecting and handling outliers etc. The merged database has 103,886 records. This database does not contain any missing values. Outliers are detected once RFM values are calculated as all fields will be numeric and the missing values are checked once again. As part of data pre-processing, the date was extracted from the "order_approved_at" column which consists of the time stamp of the invoice generation, thereby splitting it to get only the dates. Thus, the pre-processed data generated.

C. Data Transformation:

Cleaned and pre-processed data is transformed in order to build machine learning models on them. In transformation stage, customers are grouped, and Recency, Frequency and Monetary values are calculated for each customer. Recency is calculated by identifying the difference between current date and the latest invoice date. Frequency is calculated by identifying the number of orders by a customer. Monetary is calculated by the total amount of all orders by the customer. Thus, the transformed

dataset contains Customer ID, Recency (in days), Frequency, and Monetary. This transformed dataset is checked for any missing values & outliers. The below plot displays the ratio of missing entries. As the ratio of missing values is less than 0.1%, we can ignore the missing entries and consider only the complete cases for further analysis.

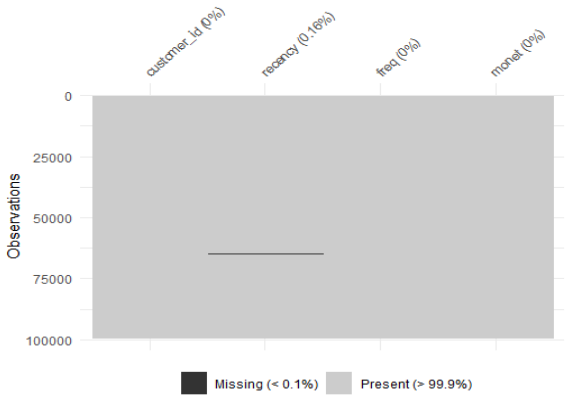

Fig. 3. Missingness Map

Thus, the data is transformed and is made ready for the models to be implemented. Outliers are detected and removed accordingly.

D. Data Mining:

(i) Elbow Method for K-means:

Elbow method is commonly used to identify the optimal number of clusters. For a range of k-value, k-Means is executed and the value of total within cluster Sum of squares is plotted. The value of k at which the plot bends is the optimal number of clusters. As shown in the plot, after k=5, value of total WSS doesn't improve further. Hence, number of clusters will be chosen as 5. The same number of clusters are used to develop Fuzzy C-means model as well.

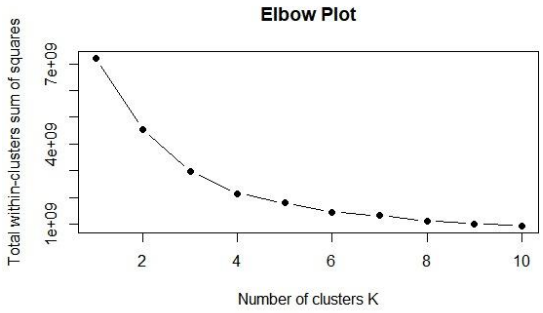

Fig. 4. Elbow Plot to identify optimal number of clusters

(*ii) Dendrograms for Hierarchical Clustering:*

Dendrograms contain the memory of hierarchical clustering algorithm. Each vertical line represents the distance or dissimilarity thresholds and the longest vertical line is selected that is not crossed by any horizontal lines. Then a cut off horizontal line is drawn to find the optimal number of clusters where it is seen from the below dendrogram plot that the number of clusters obtained is 2.

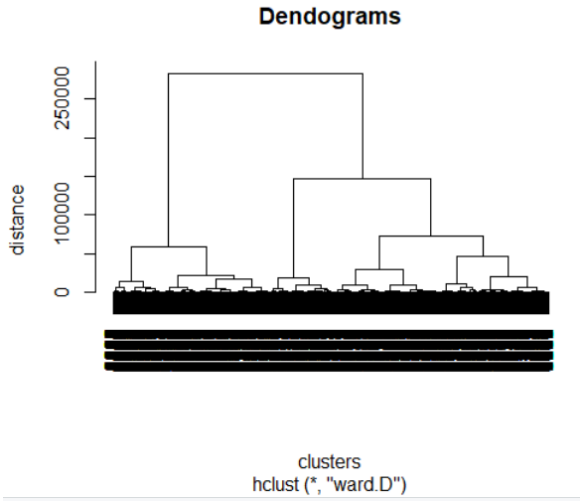

Fig. 5. Dendrogram plot to identify the optimal clusters

E. Machine Learning Models:

(i) K-means Clustering:

K-means is executed with number of clusters as 5. The output obtained is presented in figure 6 below.

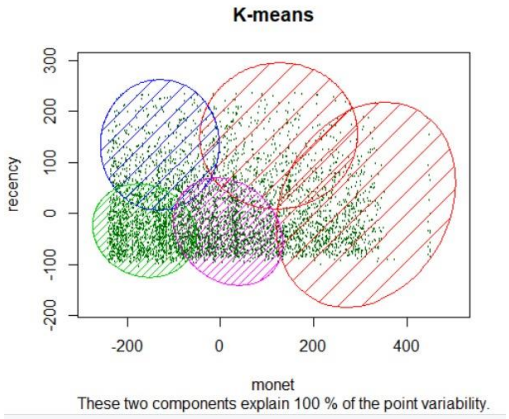

Fig. 6. K-means output

5 clusters represent 5 groups of customers as below:

1. High spending customers
2. Less Spending, recent customers
3. Average spending recent customers
4. Less spending, less recent customers
5. Average spending, less recent customers

This k-Means model is evaluated, and various parameters are discussed in section v.

(ii) Fuzzy c-means:

Fuzzy C-means was built to identify 5 clusters. It is soft clustering unlike k-means, and it categorizes as:

1. High Spending customers,
2. Average Spending customers,
3. Less spending, Less recent customers,
4. Less spending, recent customers,
5. Least spending, recent customers.

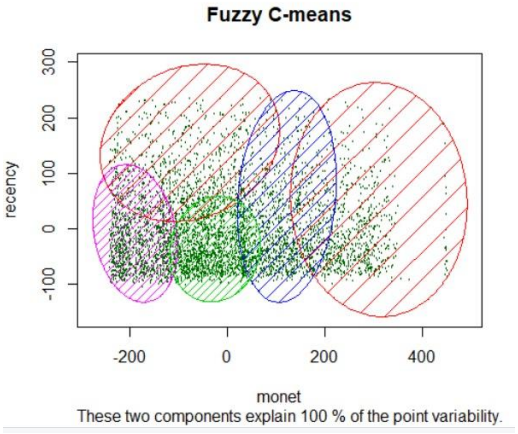

Fig. 7. Fuzzy c-means output

Various performance metrics are calculated and discussed in the next section.

(iii) Hierarchical Clustering

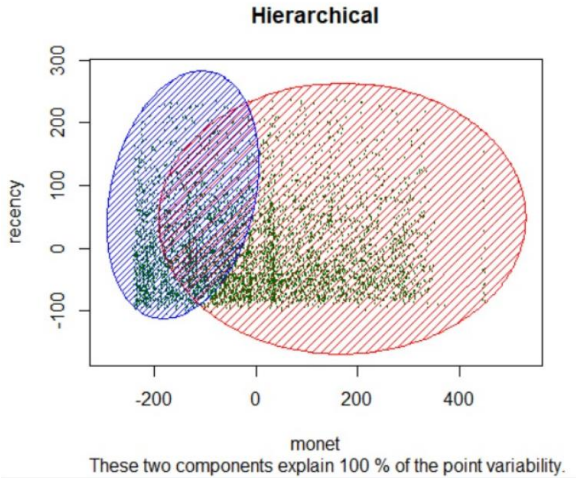

Fig. 8. Hierarchical output

By considering the dendrogram, we can see the optimal number of clusters is 2 which groups the customers as below:

1. Low spending customers and
2. High spending customers.

The performance of this model can be measured by various cluster metrics which are explained in below section.

V. RESULTS AND DISCUSSION

The performance of cluster models is validated in two categories. They are explained below.

A. Internal Cluster Validation:

The goal of a cluster model is to have lower intra-cluster distance (between objects of same cluster) and higher inter-cluster distance (between different clusters). The performance of cluster models is evaluated using the below measures:

- *(i) Silhouette Width*: It measures the closeness of each point in one cluster in comparison to other clusters. It ranges from 0 to 1 with 1 indicating the observations are well clustered. Silhouette width for the three cluster models are listed in the table below:
- *(ii) Dunn Index:* It is the ratio of the smallest inter-cluster distance to the largest intra cluster distance. Ideally, a higher value of Dunn Index is desired.

	K-means	Hierarchical	Fuzzy c-means
Silhouette Width	0.39418	0.381595	0.35838
Dunn Index	1.46043	1.580904	1.34221

As indicated in the table, K-Means has marginally higher Silhouette width in comparison to Hierarchical model. It should be noted that K-Means & Fuzzy means was built with 5 clusters, whereas Hierarchical was built only with one cluster. Based on the above information, we can confirm Hierarchical model provided better results, but the business value of the results is not so deep as it shows only two clusters.

B. External Validation:

External validation is done for comparing cluster models and study the similarity between different cluster models. It is evaluated using below parameters:

- *(i) Adjusted Rand Index*: It is a measure of the agreement between different cluster models. It varies from 0 to 1 with 1 indicating maximum agreement.

Adjusted Rand Index	K-means	Hierarchical	Fuzzy c-means
K-means	N/A	0.32519	0.51907
Hierarchical	0.32519	N/A	0.317132
Fuzzy c-means	0.51907	0.317132	N/A

The above table indicates that K-Means and Fuzzy C means produce relatively similar results and more than 50% of objects are clustered similarly.

(i) Variation of Information Index: It is a measure of variation between two cluster models. In general, a lower Variation Index is desired.

VI Index	K-means	Hierarchical	Fuzzy c-means
K-means	N/A	1.184	1.2072
Hierarchical	1.184	N/A	1.89365
Fuzzy c- means	1.2072	1.89365	N/A

The above table highlights the variations shown by Hierarchical & Fuzzy C means and helps in studying the difference in results by cluster models.

VI. CONCLUSION AND FUTURE WORK

In this research, various cluster models were developed to segment customers based on Recency, Frequency and Monetary (RFM) model and these models were evaluated. The Dunn Index indicates that the hierarchical model performed better compared to k-means and fuzzy c-means model in terms of producing a good cluster. Likewise, the Adjusted Rand Index score indicates that k-means and hierarchical models provided more varying results. These results indicate that hierarchical model provided better clusters that satisfy the primary goal of lower intra cluster distance and high inter cluster distance. However, performance of these models can be evaluated with more parameters like entropy, partition co-efficient, and so on, thereby comparing results. Furthermore, hierarchical models with different distance and partitioning methods such as Manhattan, bit-vector, and hamming can be implemented and compared with k-means.

REFERENCES

[1] A. Ammar, Z. Elouedi, and P. Lingras, "Meta-clustering of possibilistically segmented retail datasets," Fuzzy Sets and Systems, vol. 286, pp. 173–196, Mar. 2016.

[2] S. C. Oner and B. Oztaysi, "An interval type 2 hesitant fuzzy MCDM approach and a fuzzy c means clustering for retailer clustering," Soft Comput, vol. 22, no. 15, pp. 4971–4987, Aug. 2018

[3] M.-H. Masson and T. Doneux, "ECM: An evidential version of the fuzzy c-means algorithm," Pattern Recognition, vol. 41, no. 4, pp. 1384–1397, Apr. 2008.

[4] S. Ghosh and S. K. Dubey, "Comparative analysis of k-means and fuzzy c-means algorithms" IJACSA vol. 4, no. 4, pp. 35-39, 2013

[5] G. Singh, N. Kaur, "Implementation of hybrid clustering algorithm with enhanced k-means and hierarchal clustering" IJARCSSE vol. 3, no. 8, pp. 608-618, Aug. 2013.

[6] B. Romania, "Clustering the Grocery Retail Market" in Business Excellence Challenges During Economic Crisis in 7th Int. Conf. on Business Excellence, Ohio, USA, pp. 98-312, Oct. 12-13, 2013.

[7] O. Doğan, E. Ayçin, and Z. A. Bulut, "Customer segmentation by using RFM model and clustering methods: A case study in retail industry," IJCEAS, vol. 8, no. 1, pp. 1-19, 2018.

[8] M. Khalaji, S. J. Mirabedini, "Recommender System Based on Association of Complementary and Similarity in Electronic Market" *Middle East Journal of Scientific Research* vol. 13, no. 6, pp. 823-828, 2013. DOI: 10.5829/idosi.mejsr.2013.13.6.2497

[9] S. Škerlič and R. Muha, "Identifying warehouse location using hierarchical clustering," Transport Problems, vol. 11, no. 3, pp. 121–129, 2017.

[10] A. Ammar, Z. Elouedi, and P. Lingras, "Meta-clustering of possibilistically segmented retail datasets," Fuzzy Sets and Systems, vol. 286, pp. 173–196, Mar. 2016.

[11] Griva, A., Bardaki, C., Pramatari, K. and Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data.

[12] M. Abirami and V. Pattabiraman, "Data Mining Approach for Intelligent Customer Behavior Analysis for a Retail Store," in Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC – 16'), 2016, pp. 283–291.

[13] A. Ansari and A. Riasi, "Taxonomy of Marketing Strategies Using Bank Customers' Clustering," IJBM, vol. 11, no. 7, p. 106, Jun. 2016

[14] A. Riasi and S. Pourmiri, "Effects of online marketing on Iranian ecotourism industry: Economic, sociological, and cultural aspects," Management Science Letters, vol. 5, no. 10, pp. 915–926, 2015.

[15] C. H. Park, Y. H. Park, and D. A. Schweidel, "A multi-category customer base analysis," International Journal of Research in Marketing, vol. 31, no. 3, pp. 266–279, Sep. 2014.

[16] Z. Ji, Z. Xia, Q. Sun, and G. Cao, "Interval-valued possibilistic fuzzy C-means clustering algorithm," Fuzzy sets and systems, vol. 253, pp. 138-156, 2014.

[17] K.B. Munroe, Pricing: Making Profitable Decisions, McGraw-Hill, New York, 1990.

[18] M. Khajvand and M. J. Tarokh, "Estimating customer future value of different customer segments based on adapted RFM model in retail banking context," Procedia Computer Science, vol. 3, pp. 1327–1332, Jan. 2011.

[19] B. Lantz, Machine learning with R, 2nd ed, Birmingham: Packt Publishing Ltd., 2015.

[20] F. Yoseph and M. Heikkila, "Segmenting Retail Customers with an Enhanced RFM and a Hybrid Regression/Clustering Method," 2018 International Conference on Machine Learning and Data Engineering (ICMLDE), Sydney, Australia, pp. 108-116, 2018.