2019 15th International Conference on Engineering of Modern Electric Systems (EMES)

Segmenting Users of an Online Store Using Data Mining Techniques

Alexandrina-Mirela Pater
Dept. of Computers and Information Technology
University of Oradea
Oradea, Romania
mpater@uoradea.ro

Ștefan Vári-Kakas
Dept. of Computers and Information Technology
University of Oradea
Oradea, Romania
vari@uoradea.ro

Otto Poszet
Dept. of Computers and Information Technology
University of Oradea
Oradea, Romania
poszet@uoradea.ro

Ionel Gabriel Pintea
S.C. Rewine S.R.L
Oradea, Romania
pintea_gaby@yahoo.com

***Abstract*—An efficient marketing strategy is based on the proper analysis of the data collected from the history of the transactions made. The large amount of data is explored with data mining techniques, which permit the segmentation of customers depending on certain criteria. In this paper we have adopted one of the most widespread analysis process and applied it as a case study to a small business consisted of an online store.**

Keywords—data mining, data segmentation, cluster, RFM model, CRM

I. INTRODUCTION

In the last decades the online commerce has experienced an unprecedented spread due to the accessibility of internet services for household users. The seller, besides the data associated with the sold products (type, price, amount), can obtain a lot of information about the buyer, such as his address and the payment modality. The technology provides for the seller information about each customer that makes possible his personalization. This way makes possible the establishment and follow of a relationship based on the knowledge of the loyal customers, known as CRM (Customer Relationship Management). Based on the analysis of this kind of information, the selling company will be able to submit personalized offers for each client, which is to the benefit of both parties: the sales volume will increase, and the customers will be warned about the products they are interested in.

In order to get the desired data about the clients, the administrators of online shops have to use data extraction techniques. Data mining has become so very important in CRM [1, 2]. Based on the extracted data, the customers are divided by the segmentation process in groups that have common features. Such a segmentation method is the RFM model that makes the clusterization based on recency, frequency and monetary criteria [3, 4]. There are many segmentation algorithms used in various cases, such as k-means, decision trees, genetic algorithms [5, 6, 7].

Our paper focuses on the use of customers segmentation based on RFM model and k-type clusters. Section II deals with modeling and segmentation concepts adopted for our practical study. Section III presents a case study which applies the model for a very small online store. The last section contains the conclusions on the addressed topic.

II. THEORETICAL FUNDAMENTALS

In order to differentiate the customers with distinct behavioral features it is necessary to divide them on more homogenous groups based on the selected attributes. Such attributes may be related to demographic, geographic, social, financial, loyalty, life-style data. The criterion is chosen depending on the specific objectives pursued in the business.

Customer segmentation using RFM (Recency, Frequency, Monetary) is an important model in marketing analysis. It permits to identify various categories of customers for personalized services to be used in marketing campaigns. Without the use of this category identification the promotion campaigns are exposed to the waste of financial resources and the desired results will not appear [8].

In the case of a small company, such as the subject of our study, data can be collected relatively easily based on the clients' activity on the webpage of the online store. These behavioral data are usually available in the database associated to the site. It is necessary to know the buyers' identity, the number of transactions made during the analyzed period, the financial amount of transactions and the date of the most recent shopping. It is clear, that the most valuable clients are those who made in the recent past the highest number of great value shopping. Based on the results of the analysis, the future behavior of the client can be anticipated in order to make him the most suitable offer. Thus, the seller can create the best management strategy.

The RFM analysis uses for each costumer as input data the number of days/month since the last purchase (R), the number of transactions made (F) and the total amount of spent money (M) in the period under study. Based on these data, customers can be categorized in different segments beginning with the one that contains the best clients (who bought most recently, most often and spent the most money) and ending with the clients with the weakest activity. The simplest method for segmentation is to divide the customers into a number of equal groups for each of the three input parameters and assigning a score for each group. In this way, in the case of five groups theoretically $5\times5\times5=125$ segments can be created, each corresponding to a different behavior. Of course, these segments should be reduced to a manageable number, as in [9]. Analyzing the characteristics of each segment, the seller can decide upon different marketing actions in order to increase the benefit of its business.

978-1-7281-0773-8/19/$31.00 ©2019 European Union

The most popular segmentation algorithm is k-means, where data is divided in k clusters. The algorithm aims in more turns to find for each cluster a centroid against which the objects are closest [10].

The k-means clustering algorithm can be described by the following pseudocode:

```
procedure k-means-Clustering-Algorithm ();
begin
 for i := 1 to k do Clusters[i] := {0};
 for i:= 1 to k do
        Focus[i ] := rnd_object_from_DBO();
 nr_repeat := 0;
repeat
 for j:=1 to N do
 begin
      for i:= 1 to k do
          distance[i] :=
                Compute_Distance (DBO{j}, Focus[i]);
        //sqrt(sqr(j.attr1-i.attr1)+sqr(j.attr2-i.attr2)+…);
         mindist := index_of_Minimum(distance[]);
         Cluster[mindist] := Cluster[mindist]+{DBO{j}};
        //add the object to the cluster with closest focus
          element_number[mindist]:=
                element_number[mindist]+1;
  end;
 //compute now the centroid element of each cluster
 for i := 1 to k do
 begin
   for p:= 1 to element_number[i] do
    begin
      sum[p]:=0;
      for q:= 1 to element_number[i] do
          sum[p] := sum[p] +
              + Distance (Cluster[i]{p}, Cluster[i]{q});
    end;
    centroid_index := index_of_Minimum(sum[]);
    Centroid[i] := Cluster[i]{centroid_index};
  end;
nr_repeat := nr_repeat +1;
exit_flag := true;
for i := 1 to k do
  exit_flag := exit_flag and equal(Focus[i], Centroid[i]);
for i := 1 to k do  Focus[i] := Centroid[i];
until (exit_flag or nr_repeat > max_nr_repeat);
end procedure k-means-Clustering-Algorithm;
```

We denoted with DBO the database containing all of the objects, *N* is the number of the objects in the database, *k* is the number of clusters, *Focus* is the vector of the k randomly chosen initial objects, *Centroid* is the vector of the computed centroid objects, and *Clusters* denotes the array of the computed *k* clusters set.

The algorithm finishes, when it converges (the centroids are equal with the current focuses), or *nr_repeat> max_nr_repeat*.

The integration of RFM model with the k-means segmentation technique may offer a simple and often used solution for the analysis of the customers' behavior. This can make the marketing activity more efficient and the business more profitable.

III. Case Study

In this chapter is presented a case study that wants to demonstrate the efficiency of online store customer segmentation, using data mining techniques. Our analysis follows the method steps presented in [11] which process customer's data base. Customer's segmentation is realized by k-means grouping algorithm based on the RFM model.

The online retailer considered in this study is one that has its headquarters and target users in Romania. For many years, the company has relied solely on the physical store but 5 years ago the store entered the online environment and since then has had a steady increase in sales and customers. This article examines all transactions between August 2013 and May 2018. During this period 1420 associated orders with 1030 unique users were made.

This analysis, made with SAS Enterprise Miner, uses a set of variables such as: client (ID), recency (in months), first purchase data, buying frequency, monetary (total amount spent by client), monetary minimum, monetary maximum and medium value of monetary.

The Merchant Data Set has 25 variables, of which a preprocessing set of target data for analysis was generated. A series of two PHP scripts were used to convert the original data set (given in SQL code) to be used by the software tool and to calculate the Frequency, Recency, and Monetary variable values for each user code, respectively.

Total amount of online shop clients represents target data set. These data were filtered to cutoff irrelevant data from our analysis, the threshold being set at 1% of the total number of cases examined.

Next step of analysis were the clients segmentation based on values of tree parameters: recency, frequency and monetary. K-means algorithm used for data segmentation groups the objects in clusters. The resulted k clusters were generated based on the distances between objects. For each cluster is calculated the centroid object which is the element closest to each other. To get the most relevant results, the algorithm can run multiple times. The algorithm recommends 5 clusters (segments) for a relevant analysis of the target data set by a marketing specialist.

The total customer number was assigned by the above mentioned algorithm in 5 segments, as shown in Table I, their percentage distribution being shown in Fig. 1.

TABLE I. NUMBER OF CUSTOMERS IN EVERY CLUSTER

Segment	Number of customers
1	11
2	369
3	49
4	63
5	459

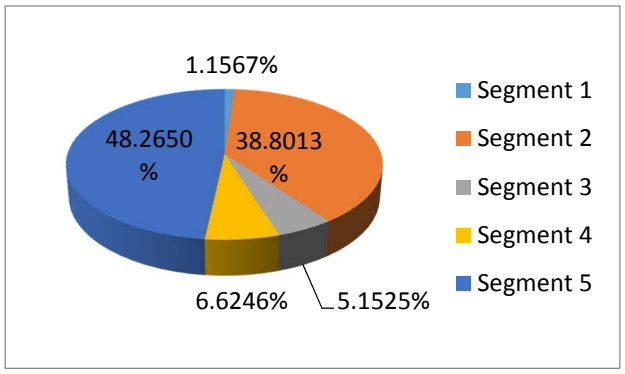

Fig. 1. Percentage of customers assigned to clusters

The importance of the variables for each segment is presented in Table II, through their rank. In segments 2 and 5, the most important variable is recency. In segments 1 and 4, the most important variable is the monetary, and in segment 3 is frequency.

The parameters that resulted to be more important for the analysis are represented in Fig. 2. The blue lines are the values calculated for the different user groups in the cluster created by the algorithm until it has ended. The red lines are the finite centroids that are closest to all the values in that cluster group.

Segment 1 refers to 11 customers and represents a modest value of only 1.1 percent of the total number of users. This group is the least interesting because it is not highlighted by either a recent or frequent high shopping. This group represents the group with the lowest payroll per capita rate and is the most unprofitable for the business.

There are 369 clients in Group 2, representing 38.8% of the total of users in the test data set. Compared to the other groups, it has a fairly low average frequency and also the amount spent per user is somewhere at the average. What highlights this group is the high recency, so those users have high recency, low frequency, and value per average user.

Unlike group 1, group 3 has a larger number of components being 49 and accounting for 5.15% of the total number of users. This group is highlighted by a rather high money rate, well above average.

Group 4 contains 63 members representing 6.62% of all users. This group stands out over monetary value above average, and this group contains those users who had the orders with the highest value per order.

Group 5 contains the biggest number of customers, 449 representing 48.26% of all shop clients. Clients of this group buy fairly frequently, spending a little but acceptable amount of money. Over time, these consumers can become very profitable or less profitable for the store.

TABLE II. EVERY CLUSTER STATISTICS

	Value	Rank
Segment 1		
recency	0.001241	3
frequency	0.003153	2
monetary	0.013823	1
Segment 2		
recency	0.41662	1
frequency	0.02059	3
monetary	0.05112	2
Segment 3		
recency	0.004363	3
frequency	0.084181	1
monetary	0.011123	2
Segment 4		
recency	0.004051	2
frequency	0.001654	3
monetary	0.079152	1
Segment 5		
recency	0.34276	1
frequency	0.02976	3
monetary	0.06972	2

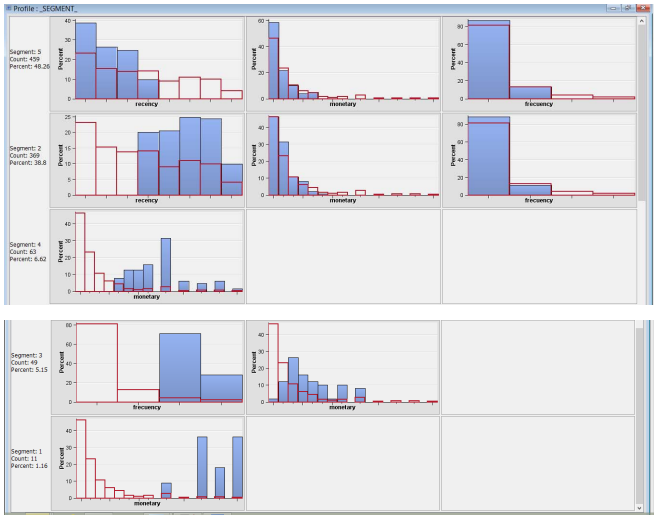

Fig. 2. Results profile segment

Our analysis can be synthesized as follows:

- 48% of the customer bought at a significant frequency, spending over average amount of money
- 38% of customers have made great profit for the store
- 7% of customers have brought average profit
- 7% of customers have brought small profit

It was found that a small number of customers contributed more than 50% of total sales.

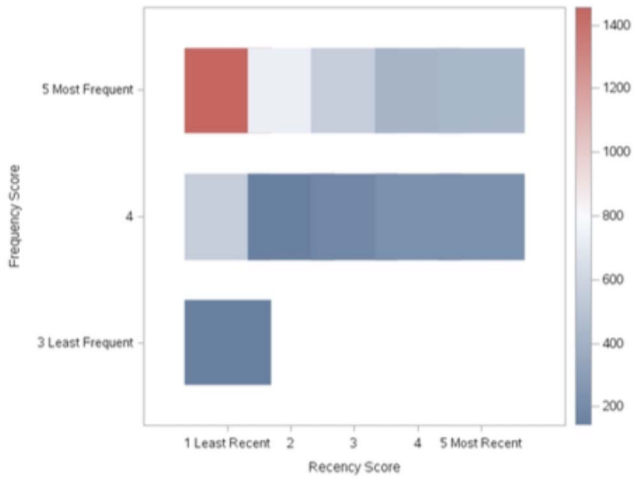

Fig. 3. Monetary distribution

Fig. 3. presents the distribution of the money spent by customers based on recent and frequency scores. Red color zone represent the number of users who bought the most frequently and spent the most money.

IV. Conclusions

The paper shows how CRM data extraction techniques help provide customized service tailored to the needs of individual clients. By analyzing purchase patterns and web links, even small businesses can make ads and promotions tailored to customer profiles. These actions can lead to increased sales and substantial savings for companies.

In the case study, it was shown that there are some stages in the data mining process, which are very important and that consume the greatest time: data acquisition, data segmentation and marketing interpretation of the results.

After applying this method, it was found that the success rate of the algorithm depends on the variables considered and the random selection of the objects.

Finally, it was found that this method can be successfully used even for a small business with a relatively small number of clients, offering relevant results for targeted marketing for client clusters created by the algorithm.

References

[1] K. Tsiptsis, A. Chorianopoulos, "Data Mining Techniques in CRM: Inside Customer Segmentation", John Wiley and Sons, 2009.

[2] J. A. Berry, S. G. Linoff, „Data Mining Techniques for Marketing, Sales, and Customer Relationship Management", Wiley, 2004.

[3] M. Mohammadian, I. Makhani, "RFM-Based customer segmentation as an elaborative analytical tool for enriching the creation of sales and trade marketing strategies", International Academic Journal of Accounting and Financial Management 3(6), 2016, pp. 21-35.

[4] K. Stormi, A. Lindholm, T. Laine, T. Korhonen. "RFM customer analysis for product□oriented services and service business development: an interventionist case study of two machinery manufacturers", Journal of Management and Governance (online), January 2019.

[5] K. Kim, H. Ahn, "A recommender system using GA K-means clustering in an online shopping market", Expert Systems with Applications, 34(2), 2008, pp. 1200–1209.

[6] L. Rokach, O. Maimon, "Data Mining with Decision Trees: Theory and Applications", World Scientific Publishing, Series in Machine Perception and Artificial Intelligence, Vol. 69, 2008.

[7] Er. Rupampreet Kaur, Er.Kiranbir Kaur, "Data Mining on Customer Segmentation: A Review", International Journal of Advanced Research in Computer Science , Volume 8, No. 5, May-June 2017.

[8] O. Dogan, E. Aycin, Z. A. Bulut, "Customer Segmentation by Using RFM Model and Clustering Methods: A Case Study in Retail Industry", International Journal of Contemporary Economics and Administrative Sciences, Volume 8, Issue 1, 2018, pp. 1-19.

[9] A. Hebbali, "Customer Level Data", https://rfm.rsquaredacademy.com/articles/rfm-customer-level-data.html (accessed March 20, 2019).

[10] Jules J. Berman, "Principles of Big Data", Morgan Kaufmann, 2013.

[11] D. Chen, S.L.Sain, K.Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining", Macmillan Publishers Ltd., 1741-2439, Database Marketing & Customer Strategy Management, Vol.19, 3, 2012, pp. 197-208.

[12] A. Kristin, "Customer Relationship Management", McGraw-Hill, 2002.

[13] C. Hung, C.-F. Tsai, "Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand", Expert Systems with Applications, 34 (1), 2008, pp. 780–787.

[14] Ph. Kotler, K. Lane Keller, "Marketing Management", Prentice Hall, 2005.

[15] C. H. Mooney, J. F. Roddick, "Sequential Pattern Mining: Approaches and Algorithms", ACM Computing Surveys 45(2), June 2013.

[16] Naveen Venkat, "The Curse of Dimensionality: Inside Out", 2018.

[17] J. Han, M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann, 2006.

[18] P. N. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining", Addison-Wesley, Boston, 2005.

[19] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinart, C. Shearer, R. Wirth, "CRISP–DM Step-by-Step Data Mining Guide", 2000.

[20] P. Muley, A. Joshi, "Application of Data Mining Techniques for Customer Segmentation in Real Time Business Intelligence", International Journal of Innovative Research in Advanced Engineering, Issue 4, Volume 2, April 2015.

[21] S. Tripathi, A. Bhardwaj, E. Poovammal. "Approaches to clustering in customer segmentation", International Journal of Engineering & Technology, 7 (3.12), 2018, pp. 802-807.

[22] Y. P. Raykov, A. Boukouvalas, F. Baig, M. A. Little. "What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm", PLoS ONE, 11(9), September 2016.

[23] T. Upadhyay, A. Vidhani, V. Dadhich. "Customer Profiling and Segmentation using Data Mining Techniques", International Journal of Computer Science & Communication, Vol. 7., No. 2., 2016, pp. 65-67.

[24] M. Khajvand, K. Zolfaghar, S. Ashoori, S. Alizadeh, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study", Procedia Computer Science, Vol. 3, 2011, pp. 57-63.