Customer Clustering using RFM analysis

VASILIS AGGELIS
WINBANK
PIRAEUS BANK
Athens
GREECE

DIMITRIS CHRISTODOULAKIS
Computer Engineering and Informatics Department
University of Patras
Patras
GREECE

Abstract: RFM (Recency, Frequency, Monetary) analysis is a method to identify high-response customers in marketing promotions, and to improve overall response rates, which is well known and is widely applied today. Less widely understood is the value of applying RFM scoring to a customer database and measuring customer profitability. RFM analysis is considered significant also for the banks and their specific units like e-banking. A customer who has visited an e-banking site Recently (R) and Frequently (F) and created a lot of Monetary Value (M) through payment and standing orders is very likely to visit and make payments again. After evaluation of the customer’s behaviour using specific RFM criteria the RFM score is correlated to the bank interest, with a high RFM score being more beneficial to the bank currently as well as in the future. Data mining methods can be considered as tools enhancing the bank RFM analysis of the customers in total as well as specific groups like the users of e-banking.

Key-words: Data Mining, e-banking, RFM analysis, Clustering

1 Introduction
RFM analysis [5] is a three-dimensional way of classifying, or ranking, customers to determine the top 20%, or best, customers. It is based on the 80/20 principle that 20% of customers bring in 80% of revenue.

In order to group customers and perform analysis, a customer segmentation model known as the pyramid model [4] is used. The pyramid model groups customers by the revenue they generate, into the categories shown in Figure 1. These categories or value segments are then used in a variety of analytics. The advantage of this approach is that it focuses the analytics on categories and terminology that are immediately meaningful to the business.

The pyramid model has been proven extremely useful to companies, financial organisations and banks. Indicatively some issues that can be improved by the use of the model follow:
- Decision making.
- Future revenue forecast.
- Customer profitability.
- Predictions concerning the alteration of customers’ position in the pyramid.

Essentially RFM analysis suggests that the customer exhibiting high RFM score should normally conduct more transactions and result in higher profit for the bank.

RFM analysis [3, 7, 8, 9] nowadays can be conducted by the use of Data Mining methods like
clustering. These methods contribute to the more efficient determination and exploitation of RFM analysis results.

In the present paper, the RFM scoring of active e-banking users is studied along with the ranking of these users according to the pyramid model. The software used is SPSS Clementine 7.0. Description of various clustering techniques and algorithms follow in section 2 while in section 3 the calculation of the RFM scoring of active e-banking users is described. Section 4 contains experimental results derived from the data set of section 3 and conclusions and future work is stated in section 5.

2 Clustering Basics
Clustering techniques [2, 6] fall into a group of undirected data mining tools. The goal of undirected data mining is to discover structure in the data as a whole. There is no target variable to be predicted, thus no distinction is being made between independent and dependent variables.

Clustering techniques are used for combining observed examples into clusters (groups) that satisfy two main criteria:
- each group or cluster is homogeneous; examples that belong to the same group are similar to each other.
- each group or cluster should be different from other clusters, that is, examples that belong to one cluster should be different from the examples of other clusters.

Depending on the clustering technique, clusters can be expressed in different ways:
- identified clusters may be exclusive, so that any example belongs to only one cluster.
- they may be overlapping; an example may belong to several clusters.
- they may be probabilistic, whereby an example belongs to each cluster with a certain probability.
- clusters might have hierarchical structure, having crude division of examples at highest level of hierarchy, which is then refined to sub-clusters at lower levels.

2.1 K-means Algorithm
K-means [1, 2, 6, 11] is the simplest clustering algorithm. This algorithm uses as input a predefined number of clusters that is the \(k\) from its name. Mean stands for an average, an average location of all the members of a particular cluster. When dealing with clustering techniques, a notion of a high dimensional space must be adopted, or space in which orthogonal dimensions are all attributes from the table of analysed data. The value of each attribute of an example represents a distance of the example from the origin along the attribute axes. Of course, in order to use this geometry efficiently, the values in the data set must all be numeric and should be normalized in order to allow fair computation of the overall distances in a multi-attribute space.

K-means algorithm is a simple, iterative procedure, in which a crucial concept is the one of \textit{centroid}. \textit{Centroid} is an artificial point in the space of records that represents an average location of the particular cluster. The coordinates of this point are averages of attribute values of all examples that belong to the cluster. The steps of the K-means algorithm are given in Figure 2.

1. Select randomly \(k\) points (it can be also examples) to be the seeds for the centroids of \(k\) clusters.
2. Assign each example to the centroid closest to the example, forming in this way \(k\) exclusive clusters of examples.
3. Calculate new \textit{centroids} of the clusters. For that purpose average all attribute values of the examples belonging to the same cluster (\textit{centroid}).
4. Check if the cluster \textit{centroids} have changed their "coordinates". If yes, start again form the step 2). If not, cluster detection is finished and all examples have their cluster memberships defined.

![Fig.2 – K-means algorithm](image)

Usually this iterative procedure of redefining \textit{centroids} and reassigning the examples to clusters needs only a few iterations to converge.

2.2 Two Step Cluster
The Two Step cluster analysis [10] can be used to cluster the data set into distinct groups in case these groups are initially unknown. Similar to K-Means algorithm, Two Step Cluster models do not use a target field. Instead of trying to predict an outcome, Two Step Cluster tries to uncover patterns in the set of input fields. Records are grouped so that records within a group or cluster tend to be similar to each other, being dissimilar to records in other groups.
Two Step Cluster is a two-step clustering method. The first step makes a single pass through the data, during which it compresses the raw input data into a manageable set of subclusters. The second step uses a hierarchical clustering method to progressively merge the subclusters into larger and larger clusters, without requiring another pass through the data. Hierarchical clustering has the advantage of not requiring the number of clusters to be selected ahead of time. Many hierarchical clustering methods start with individual records as starting clusters, and merge them recursively to produce ever larger clusters. Though such approaches often break down with large amounts of data, Two Step’s initial pre-clustering makes hierarchical clustering fast even for large data sets.

3 RFM scoring of active e-banking users
The data sample used concern the period between January 1st and December 12th of the year 2002.

The term «active e-banking user» describes the user who has conducted at least one financial transaction during this period. In order RFM scoring to express customer profitability, all values concerning financial transactions are taken into consideration [12].

The following variables are calculated for this specific time period.

Recovery (R)
R is the date of the user’s last transaction. Since the R value contributes to the RFM scoring determination, a numeric value is necessary. Therefore, a new variable, \( R_{\text{new}} \) is defined as the number of days between the first date concerned (1/1/2002) and the date of the last active user’s transaction. For example a user who has conducted his last transaction on 29/11/2002 is characterized by \( R_{\text{new}}=332 \), while one who has conducted his last transaction on 4/4/2002 will have \( R_{\text{new}}=93 \).

Frequency (F)
R is defined as the count of financial transactions the user conducted within the period of interest (1/1/2002 to 12/12/2002).

Monetary (M)
M is the total value of financial transactions the user made within the above stated period.

RFM Score (RFM Factor) is calculated using the formula:
\[
\text{RFM Factor} = R_{\text{new}} + F + M.
\]

A sample of the data set on which data mining methods are applied lies in Table 1.

<table>
<thead>
<tr>
<th>User</th>
<th>( R_{\text{new}} )</th>
<th>F</th>
<th>M</th>
<th>RFM Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>User522</td>
<td>330</td>
<td>20</td>
<td>€20.8</td>
<td>21.206,3</td>
</tr>
<tr>
<td>User523</td>
<td>216</td>
<td>6</td>
<td>€16.9</td>
<td>17.154,1</td>
</tr>
<tr>
<td>User524</td>
<td>304</td>
<td>8</td>
<td>€12.8</td>
<td>13.178,2</td>
</tr>
<tr>
<td>User525</td>
<td>92</td>
<td>1</td>
<td>€27.2</td>
<td>27.338,4</td>
</tr>
</tbody>
</table>

Table 1 – Sample Data

The sample includes 1904 active users in total. Customer classification is performed using the K-means and Two Step Clustering methods.

4 Experimental Results
As seen in the histogram of Figure 3, RFM distribution is high over values less than 1,000,000. This is a natural trend since, as concluded in paragraph 1, 80% of the customer exhibits low RFM Factor.

Application of the K-means algorithm results in the 4 clusters of Figure 4. Next to each cluster one can see the number of appearances as well as the average value of each variable.
The above clustering results in the distribution of Figure 5. The similarity of this distribution to the pyramid model is apparent:

- **Cluster 1**: 81.93% Small 80%
- **Cluster 3**: 15.65% Medium 15%
- **Cluster 4**: 2.26% Big 4%
- **Cluster 2**: 0.16% Top 1%

Additionally as another way of certifying the existence of different customer clusters, the Two Step Cluster method was used. This method yielded the four clusters of Figure 6. The number of appearances is also supplied in this case accompanied with the average value and standard deviation of the variables of each class.

The distribution derived from the above clustering procedure is seen in Figure 7. The similarity to the pyramid model is even greater. Specifically:

- **Cluster 1**: 81.3% Small 80%
- **Cluster 2**: 14.55% Medium 15%
- **Cluster 3**: 3.52% Big 4%
- **Cluster 4**: 0.63% Top 1%

Finally, cross-tabulation of the fields of the two different clustering procedure results in the Matrix of Figure 8.
The lines are interpreted like this:
From the total of 1560 records of Cluster 1 as derived from the K-means algorithm 1260 (80.76%) are included in Cluster 1, 243 (15.57%) records in Cluster 2, 48 (3.07%) in Cluster 3 and 9 (0.6%) in Cluster 4 derived from the Two Step method. Next lines are interpreted the same way.

The most important observations concerning the matrix are the following:
- 1260 users belong to the category of «small» customers, 13 in the «medium» category, 17 in the «big» and 3 in the top category as a result of the application of the two models combined.
- 9 users of the «top» class as derived from the Two Step method belong to the «small» category of K-means method.
- 5 users that belong to the «small» class according to the Two Step method belong to the «big» category according to the K-means.
- Despite some differences in the classification of customers in the various categories of the pyramid, the two models are certified from the majority of the records.

5 Conclusions and Future Work
In the present paper it is shown that the knowledge of RFM scoring of active e-banking users can rank them according to the pyramid model. This result was highlighted by the use of 2 clustering methods. Therefore, the e-banking unit of a bank may easily identify the most important users-customers. The model continuously trained reveals also the way customers are transposed between different pyramid levels so that the bank administration has the opportunity to diminish customer leakage.

At the same time customer approach and new services and products promotion is improved since it is the bank’s knowledge that it is more likely a customer to respond to a promotion campaign if this customer belongs to the 20% of more beneficial ones.

Correct recognition and analysis of the clustering results offers an advantage to the e-banking unit of a bank over the competition. Users-customers clustering could be subjected to further exploitation and research.

Crucial hints of future work are the payment types preferred in each category, e-banking use, user profiles and other general characteristics of each customer category.

The use of other clustering algorithms as well as other data mining methods is a promising and challenging issue for future work. The application of RFM analysis can also be used in larger data sets, in order to produce completed results that will be updated continuously by training of the models.

References:
[7]. Im K. and Park S., (1999) “A Study on Analyzing Characteristics of Target Customers from Refined Sales Data”, APIEMS.