Applying Data Mining Techniques and Analytic Hierarchy Process to the Food Industry: Estimating Customer Lifetime Value

Fátima Carneiro
Kaizen Institute Western Europe, Faculty of Engineering of the University of Porto
Rua Manuel Alves Moreira 207, Vila Nova de Gaia, Portugal
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
fcarneiro@kaizen.com

Vera Miguéis
Faculty of Engineering of the University of Porto, INESC TEC
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
vera.migueis@fe.up.pt

Abstract
Customer segmentation is increasingly needed in a context where customer interests are vital for companies to survive. This study proposes the use of the weighted RFM (Recency, Frequency, Monetary) supported by data mining techniques and the Analytic Hierarchy Process (AHP), to classify the customers according to their lifetime value (CLV). The customer segments obtained can be used to boost marketing strategies, as these segments enable to differentiate the customers. Each segment of customers is described by a set of rules based on the customers' purchasing patterns. The methodology developed is validated by using a real case study, i.e. a food industry company, whose core business is the production of biscuits.

Keywords
AHP, Clustering, Customer lifetime value (CLV), RFM model and WRFM model.

1. Introduction
Customers should be seen as one of the most critical business assets for companies’ success (Van den Poel and Lariviere 2004), but not all of them can be managed equally. Targeting customers properly means categorising them to provide valuable relationships, that may exceed their expectations. Given the impossibility of any market offer to satisfy all markets and their respective individuals, segmentation appears as an important marketing tool (Assael and Roscoe Jr 1976, Constantinides and Zinck Stagno 2011). Wedel and Kamakura (2002) reinforce this idea by stating that segmentation has become a need for marketers since it is the only possible way to save money and efforts that directly affect companies’ profit. Segmentation assumes that customers with similar purchasing needs and patterns tend to demonstrate a similar response to marketing actions (Tsai and Chiu 2004). Ngai, Xiu et al. (2009) also mention that segmentation enables to obtain a deeper understanding of a specific market and an enhanced capacity to identify and explore opportunities for commercial profit. The result of a good segmentation allows organisations to define their objectives (Lamb, Hair et al. 2008).

Companies need to know the customers, what they want and how much they are willing to pay (Kotler 2000). However, for many companies, it is difficult to understand customers' needs, especially if there is no direct interaction. However, the vast amount of data organizations is collecting, if used properly, enable to discover important patterns and trends. For this purpose, data mining techniques have gained relevance in Customer Relationship management (CRM) context (Berson, Smith et al. 2000).
According to Khajvand, Zolfaghar et al. (2011), one of the simplest and most powerful models to support CRM is the RFM (Recency, Frequency, Monetary) model. This model is based on three purchasing variables that are usually used in the segmentation phase to aggregate customers and finally to describe those groups.

The company under study categorises the market using geographic factors, due to the differences in the markets belonging to different countries. Segmenting customers by geographic area is a traditional segmentation approach, not being used to the same extent as before, as many services have become electronic, reducing the inherent distance difficulties (Weinstein 2013). Moreover, the location may not represent similar purchasing behaviour. Therefore, this paper proposes to support customer relationship management through customers segmentation using the RFM approach. Nevertheless, this paper also proposes the integration of the geographical dimension in customers segmentation. Classification techniques are used to describe the segments obtained after clustering. In the end, the results of the segmentation will enable to choose the most appropriate groups of customers to target considering the company’s goals.

The structure of the paper is as follows. Section 2 presents a literature review, in order to obtain a general view of the case studies on this research topic, as well as to understand the existing gaps in the literature. Section 3 describes the methodology proposed. In section 4, the analytical results are described. In the last section, the main conclusions and perspectives of the future work are presented.

2. Introduction

The Recency, Frequency and Monetary model, proposed by Hughes (1996), allows to identify customers who are more likely to respond to new offers, being simple and easy to apply (Bauer 1988, Bult and Wansbeek 1995), given the small number of variables it considers (Kaymak 2001). Using the RFM model, customers are grouped based on three variables associated with the consumer’s behaviour: the time interval between the moment of analysis and the time of the most recent purchase (Recency), the number of purchases that the customer made within a considered period (Frequency) and the average value of the transactions (Monetary Value). This model has proven to be effective when applied to transactional data (Bauer 1988, Fader, Hardie et al. 2005), facilitating the decision-making process through a clear interpretation of the results (Marcus 1998). However, RFM is limited. In fact, in addition to these three attributes, there are usually other relevant attributes to support the implementation of a successful marketing strategy (Fitzpatrick 2001).

Thus, over time, several authors have been adding new variables to the RFM model in order to complete the analysis of customer behaviour. Wei, Lin et al. (2012) refined the model by looking at the length of the relationship between the company and the customer (L) - LRFM. L represents the period between the first transaction and the most recent transaction of a consumer, distinguishing long and short-term consumers.

In recent research, some authors propose to apply differential weights for each of the three components, depending on the importance of each of them in the organisational context. This is called the Weighted RFM (Stone and Jacobs 1988, Liu and Shih 2005, Shen and Chuang 2009) model. In practical terms, the Weighted RFM implies defining the relative importance (weights) of the RFM variables, which depend on the industry under analysis. For this purpose, Stone and Jacobs (1988) determined the weights subjectively without the application of a systematic approach.

Table 1 lists some studies using RFM, WRFM and LRFM models, and highlights the application context.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Application context</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Online Sports Store</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grocery Store</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Commercial Store</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Online Retail Business</td>
<td></td>
</tr>
<tr>
<td>WRFM</td>
<td>Hardware Retailing Company</td>
<td>Shih and Liu (2003)</td>
</tr>
</tbody>
</table>
The methodologies previously identified are mostly applied in retailing. It is noteworthy that studies in industry are only a few, and none in the sector under analysis in this case study, i.e. the food industry. It is also important to highlight that, although the models presented are the most used segmentation approaches, by meeting transaction information, they ignore other information, such as individual differences (for example, values, motivations, lifestyles), that may be crucial in segmenting customers and targeting.

Segmentation are also used with the purpose of evaluating the value of the customers for the company. Kotler (1974) introduced one of the key CRM concepts, i.e. the Customer Lifetime Value (CLV), used to assess customer value in order to establish differentiated relationship between the segments according to their value (Kumar 2008). Usually, CLV represents the current value of all future profits from a customer during its life cycle (Gupta and Lehmann 2003, Rust, Lemon et al. 2004). Nevertheless, there are companies that rank customers based on the CLV inferred from the past transactions, disregarding the future value. Thus, the RFM variables may constitute an estimative of the CLV value. The analytical hierarchy process (AHP) may be used to determine the relative importance of RFM variables, supporting the Weighted RFM model (WRFM) (Liu and Shih 2005, Parvaneh, Abbasimehr et al. 2012). This AHP is a multi-criteria decision making method that helps the decision-maker facing a complex problem with multiple conflicting and subjective criteria (Saaty 2008, Ishizaka and Labib 2009).

In parallel with the momentum gained by the segmentation approaches in practical settings, data mining techniques also gained relevance in the last decades. One of the most popular data mining techniques refers to the clustering techniques. In CRM context, clustering techniques allow to group customers with similar buying patterns into the same group and separate customers who are different (Kesavaraj and Sukumaran 2013). On the other hand, some classification algorithms may be used to create rules which describe the clusters obtained in the clustering phase.

3. Methodology

This paper proposes to compare two different segmentation approaches, which may enable to group customers based on their purchasing behaviour. Furthermore, this paper proposes to assess the value of the segments obtained, in order to establish a ranking of the segments that may help managers taking more efficient decisions, given the importance of their customers. The proposed methodology, illustrated in Figure 1, is based on the CRISP-DM approach.
The first phase involves the definition of the requirements to determine the objectives to be achieved. The data understanding consists of analysing the available data, including its collection, classification and quality verification. The third step prepares the data for the clustering phase and involved the following procedures:

1. To select the attributes related to the RFM model and to eliminate unnecessary information
2. To link the customer ID to their geographic location and their purchasing data
3. To exclude the records (or customers) whose purchase value is equal to zero
4. To withdraw customer returns, subtracting them from the value of the transactions in which the product was purchased
5. To merge the rows of the same transactions into one
6. To estimate the following attributes for each customer: Recency, Frequency and Monetary. The recency is presented in months, counting the time since the last transaction was made. In turn, frequency counts the number of transactions made during the analysis period. The monetary value shows the average amount spent per transaction, in euros.
7. To detect and analyse the presence of outliers

In the clustering stage, two different approaches are proposed. In the first approach, the RFM segmentation model is used, whereas in the second approach, a geographic segmentation is first performed and then the RFM segmentation model is applied. The K-Means clustering method is one of the best-known unsupervised algorithms for the clustering process. To identify the optimal number of clusters using the K-Means algorithm, several metrics can be used to evaluate the quality of the result achieved (Tibshirani, Walther et al. 2001). In this paper we used the Elbow criterion, a method that analyses the percentage of variance explained as a function of the clusters number. The first clusters will add a lot of information, but at some point, the marginal gain drastically decreases and forms an elbow in the graph. When it is not possible to get a clear elbow, there are other alternatives, such as the average silhouette method, to infer the number of clusters to consider (Rousseeuw and Kaufman 1990). The silhouette value is a measure of how an object is similar to its cluster (cohesion) compared to other clusters (separation), indicating a high value that the object is quite compatible with its cluster, unlike neighbouring clusters.

In the evaluation stage, the quality of the segments was measured based on the heterogeneity within and between segments (Wedel and Kamakura 2012) and the feasibility of a practical implementation.

After comparing the two segmentation approaches and having selected the most appropriate, the WRFM is applied in order to evaluate the value of each segment for the company. For this purpose, it was necessary to define the weights of each of the RFM variables. This was performed using the Analytic Hierarchy Process. To define the relative importance of the criteria, the main decision makers of the company used as case study were involved to compare each possible pair of criteria. After the comparison exercise, it was evaluated the consistency of the judgements made. In some cases, the comparison of the alternatives may not be consistent, meaning that the AHP process will not be effective. A perfect consistency implies an index of zero inconsistency. However, it is typically not achieved due to the human being biased and inconsistent when making judgements. According to Shih and Liu (2003), an inconsistency index of less than 10% is acceptable.

Considering the weights inferred from the AHP method, the average value of the CLV can be evaluated regarding the normalised variables R, F and M, through the Equation 1 (Hosseini, Maleki et al. 2010). $C^j$ corresponds to the value of the customer $j$.

$$ C^j = w_R C_R^j + w_F C_F^j + w_M C_M^j $$  \hspace{1cm} (1)

Where:

- $w_R, w_F, w_M$ represent the relative weights of the RFM variables
- $C_R^j, C_F^j, C_M^j$ are the normalised variables under study
Note that for the calculation of $C^I_j$, the complementary to one of the normalised recency values was taken. This is due to the fact that $C^I$ aims at measuring the value of the customer, which is higher when the recency is lower.

4. Case Study

This paper proposes to investigate the segmentation model that better represents the customers of the company used as case study. This section shows the results of the implementation of the methodology described in section 3. Two approaches are presented, distinguished by the inclusion or not of the customers’ geographic information.

4.1 Business Understanding

In competitive environment, customer attraction and satisfaction depend on the ability of companies to understand the nature and characteristics of their customers. The company under study is considered a sustainable and competitive company, investing on quality and innovation to build unique biscuits, almonds and sweets. Over the years, it has always been concerned about customer needs. For this reason, investing in the customers is a long-term strategy, essential to increase market share and stand out at national and international levels.

4.2 Data Understanding

This study includes 28,259 transactions of 296 customers, comprising the time period between November 1, 2017 and October 31, 2018. A transaction involves the time of purchasing, the item purchased, the quantity purchased, and the monetary value of the transaction. The data source provided by the company was used to validate the proposed methodology. It is noteworthy that it was necessary to invest a lot of time during this stage of the process, since the information collected was not organized.

It should be noted that the customers of the company used as case study do not correspond to the end customer, but to wholesalers and retailers, as this is a B2B company.

4.3 Data Preparation

The data transformation step culminated in a dataset of 296 rows, each one representing one customer. This dataset included 4 variables, i.e. the RFM variables and the geographic location of the customer. After completing this step, a descriptive analysis was performed to describe and summarize the data. Subsequently, the presence of outliers and their treatment was analysed.

By analysing Table 2, it is possible to observe the difference between the average and maximum values of the RFM variables. The average recency and frequency are below the 3rd quartile, that is the value from which they are 25% of the highest values. However, with the monetary variable, the average value is higher, indicating that less than 25% of the data contains the highest values. It is important to mention the amplitude of each of the variables, highlighting the recency, where it is possible to infer that there is at least one customer who did not make any purchases during a year.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Min.</th>
<th>1st Q</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Q</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recency</td>
<td>0,0</td>
<td>0,0</td>
<td>1,0</td>
<td>2,7</td>
<td>5,0</td>
<td>12,0</td>
</tr>
<tr>
<td>Frequency</td>
<td>1,0</td>
<td>6,0</td>
<td>18,0</td>
<td>78,9</td>
<td>47,3</td>
<td>3035,0</td>
</tr>
<tr>
<td>Monetary</td>
<td>3,0</td>
<td>69,0</td>
<td>184,5</td>
<td>1101,2</td>
<td>725,2</td>
<td>22926,0</td>
</tr>
</tbody>
</table>

To visualize the distribution and outliers of the data, the boxplots of the variables R, F and M are shown in Figure 2. Some instances have higher values of M and F compared to most instances in the dataset, leaving those values outside the interquartile range.
According to Hautamäki, Cherednichenko et al. (2005), the K-Means clustering, one of the best-known unsupervised algorithms for the clustering process algorithm, is sensitive to outliers and variables with unmatched magnitudes (MacQueen 1967). In this sense, it is important to analyse if the data present an atypical behavior and if it can be considered outliers, after evaluation. The Frequency has 29 records with unusual behaviour and the Monetary variable 50, corresponding to a total of 78, with a record in common between the variables F and M. Although the data are formally considered outliers these instances correspond to actual behaviour of the customers and consequently these customers were not excluded from this study.

To calculate the linear dependence between the variables of the RFM model, the Pearson correlation coefficient was used (Benesty, Chen et al. 2009). According to the correlation matrix, Figure 3, it is possible to verify that the variables present a negligible correlation, given the ρ to be, positive or negative, between 0 and 0.3 (range indicating a non-significant correlation) (Lee Rodgers and Nicewander 1988).

The three variables are not on the same scale and the intervals differ considerably: R [0,12]; F [1,3035] e M [3, 22926]. Since the distance-based methods are affected by the difference in scale between the attribute values, before the clustering step, the variables were normalised. The method used was the min-max normalisation, which allows the linear transformation of the original data to prevent attributes with large ranges out-weight attributes with small ranges.
4.4 Modelling

Based upon the data mining objective, after preparing the data, the next step is to build the two models design with subsequent quality and validity testing.

4.4.1 1st Approach

Figure 4 presents the sum of the squares of each cluster, considering a maximum number of 12 clusters. In addition to this measure, the relationship between SS and its total is also considered to help classifying the properties of internal cohesion and external separation. Consequently, its value of 93.9% reflects the total variation in the dataset that is explained by this 12 clusters.

![Figure 4 - Elbow Curve](image)

Given that Figure 4 does not show a clear elbow, it was applied the average silhouette method. By analysing Figure 5, the number of clusters to be constructed was set to 5, representing the number of customer groups that have similar RFM behaviour.

![Figure 5 - Average Silhouette Method](image)

As shown in Table 3, the selected clusters cover 83.6% of the total variance present in the dataset. This table also shows the Sum of Square Error (SSE) for each cluster. In this case, the cluster 1 has the greater dispersion value, while cluster 3 has the highest cohesion.

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of different customers</strong></td>
<td>63</td>
<td>150</td>
<td>7</td>
<td>67</td>
<td>9</td>
</tr>
<tr>
<td><strong>Within-cluster sum of square</strong></td>
<td>1.97</td>
<td>0.71</td>
<td>0.42</td>
<td>0.94</td>
<td>0.44</td>
</tr>
<tr>
<td><strong>Between SS/Total SS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83.6%</td>
</tr>
</tbody>
</table>

© IEOM Society International 2021
Figure 6 allows comparing the variables R, F and M, for each cluster, considering the dataset normalised. Each line represents one variable and the corresponding values of the centroids of the clusters. For example, cluster 3, presents the highest value of frequency, with a minimum recency and monetary value.

In Figure 7, for a better understanding of each cluster, a decision tree presents the rules that enable to distinguish the segments. Both cluster 1 and 4 differ at recency level, with the first cluster having a value equal to or greater than 6.5 months and the fourth clustering between 2.5 and 6.5 months. On the other hand, clusters 2, 3 and 5 present recency values less than 2.5 months. Cluster 2 is distinguished from the 5 at the frequency level, with a value of less than 829 purchases. Cluster 3 is distinguished from 2 because it presents a monetary value greater than 5 538 €.

4.4.1 2\textsuperscript{nd} Approach

In the 2\textsuperscript{nd} approach, the segmentation was carried out in two moments: in the first considering the geographic region and in the second one applying the RFM model to each region. Table 4 identifies the number of clusters obtained in each moment, the distribution of the customers per cluster and the SSE for each cluster.

<table>
<thead>
<tr>
<th></th>
<th>C1 - America</th>
<th>C2 - Africa</th>
<th>C3 - Asia</th>
<th>C4 - Oceania</th>
<th>C5 - Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>1\textsuperscript{st} Moment</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Number of different customers</td>
<td>26</td>
<td>8</td>
<td>17</td>
<td>1</td>
<td>244</td>
</tr>
<tr>
<td>Between SS/Total SS</td>
<td>81,1%</td>
<td>100%</td>
<td>86,9%</td>
<td>100%</td>
<td>44,5%</td>
</tr>
</tbody>
</table>

© IEOM Society International
For illustration purpose we present, in Table 5, each of the sub-clusters of cluster 5 (Europe). The SS relationship has a value of 44.5%. Figure 8 shows the respective parallel coordinate graph, which enables to characterize the clusters.

### Table 5 - Sum of Square by Cluster (cluster: Europa)

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of different customers</td>
<td>125</td>
<td>119</td>
</tr>
<tr>
<td>Within-cluster sum of square</td>
<td>17.12</td>
<td>14.58</td>
</tr>
<tr>
<td>Between SS / Total SS</td>
<td>44.5%</td>
<td></td>
</tr>
</tbody>
</table>

![Cluster Plot (Cluster Europa)](image)

**Figure 8 - Cluster Plot (Cluster Europa)**

### 4.5 Evaluation and CLV Calculation

The purpose of this section is to evaluate the two models presented and calculate the CLV for the selected one.

#### 4.5.1 Evaluation

The next step is to evaluate the quality of the clusters formed in both models, in order to understand how the business objectives are met by the proposed models and if there are commercial reasons to prove their applicability.

The SSE of each one of the clusters obtained in the two approaches were calculated. In model 1, a value of 83.6% was obtained. In order to be compared with model 2, a weighted average of the SSE was estimated, based on the respective number of clients. It was obtained a value of 51.8%. These values revealed that the customer segmentation based on the second model minimises the sum of the cluster distance between the elements, as well as the cluster centre, ensuring further reduction of clustering errors and a greater precision of the formed clusters.

At this stage, the company under study was involved to evaluate the models and to analyse the potential benefits of the study. The company concluded that that there is no capacity to provide a differentiated relationship with the 21 clusters identified by the second approach. Furthermore, by analysing the characteristics of the segments of the 1st approach, inferred from the decision trees, the company chose the first model, with 5 clusters, as the most appropriate.

#### 4.5.2 Estimating CLV for Clusters

The pairwise comparison of all criteria (recency, frequency and monetary) was performed by the company's decision makers and are shown in Table 6.
Table 6 - Relative degree of importance for pairwise comparisons

<table>
<thead>
<tr>
<th>Which criterion with respect to AHP priorities is more important?</th>
<th>Equal?</th>
<th>How much more?*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recency</td>
<td>Frequency</td>
<td>1</td>
</tr>
<tr>
<td>Recency</td>
<td>Monetary</td>
<td>1</td>
</tr>
<tr>
<td>Frequency</td>
<td>Monetary</td>
<td>1</td>
</tr>
</tbody>
</table>

*AHP Scale: 1- Equal Importance, 3- Moderate importance, 5- Strong importance, 7- Very strong importance, 9- Extreme importance (2, 4, 6, 8 values in-between)

According to the assessments obtained through the AHP, a consistency ratio (CR) of 7.7% was achieved, which validates the model. The resulting relative weights of the variables were: $w_r = 0.073$, $w_f = 0.166$ and $w_m = 0.761$. Having these values, for each customer, it was estimated a weighted RFM. Thus, their individual normalised R, F and M were combined using the weights obtained from AHP. Note that in the case of Recency, the complementary to one was taken, as the higher its value, the less value should be given to its customer. This parameter differs from frequency and monetary, where their values have a direct relation to the weight to be given.

By comparing the average weighted of the parameters of RFM, it is possible to rank customers according to their CLV - Table 7. For example, segment number 5 has the highest CLV value. In this case, most of its customers have recently purchased, with the highest buy-per-transaction average. These customers show the characteristics of a very valuable segment whose segmentation strategies should consider these specifications.

Table 7 - CLV Rank

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Recency</th>
<th>Frequency</th>
<th>Monetary</th>
<th>Weighted RFM</th>
<th>CLV Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.34</td>
<td>0.00</td>
<td>0.04</td>
<td>0.06</td>
<td>5</td>
</tr>
<tr>
<td>C2</td>
<td>0.97</td>
<td>0.02</td>
<td>0.03</td>
<td>0.09</td>
<td>3</td>
</tr>
<tr>
<td>C3</td>
<td>1.00</td>
<td>0.51</td>
<td>0.01</td>
<td>0.17</td>
<td>2</td>
</tr>
<tr>
<td>C4</td>
<td>0.71</td>
<td>0.01</td>
<td>0.03</td>
<td>0.08</td>
<td>4</td>
</tr>
<tr>
<td>C5</td>
<td>0.87</td>
<td>0.00</td>
<td>0.56</td>
<td>0.49</td>
<td>1</td>
</tr>
</tbody>
</table>

5. Conclusions and Future Work

This paper aims at demonstrating how a customer-centric analysis can be created through the application of data mining techniques and AHP. Two segmentation models were presented, being the first one selected, considering the feasibility of implementation. Subsequently, the respective CLV for each segment of customers was estimated.

The objective of the customers segmentation is not only to identify segments, but to discover how they can articulate and integrate into the 4P strategy (Price, Place, Product and Promotion). As a next step, it is necessary to explore the development of the strategy to be launched, resulting from the segmentation process. With the information resulting from this study, the company could define a different strategy depending on the type of clusters. For the segment referring to the lowest value, marketing campaigns and discounts could be launched to stimulate buying, make current products known (for example, by providing samples), promote experimentation, dispose of products or blur seasonality of consumption. For clusters that currently have a higher value, premium treatment and a closer relationship with the customers, is recommended to maintain customer value. Given the company's focus on innovation and the consequent launch of new products, it could be interesting to involve the most valuable customers in the ideation and/or to guarantee exclusive sales, in order to reinforce their special nature.

Throughout the development of this project, most of the time was spent on processing the information provided by the company, which corroborates with several data mining studies that emphasize that the preparation stage is one of the most relevant and often time-consuming aspects of these type of projects (Khajvand, Zolfaghar et al. 2011, Bunnak, Thammaboosadee et al. 2015). This reinforces the importance of keeping information organized and up to date, providing a central point of consultation and record of information.
Despite the relevance of the study, some limitations regarding the results should be noted. Some other factors, mainly important variables in the food industry, could be considered in the segmentation models in order to achieve more significant profiles (e.g.: production changeover times). Moreover, the variable related to the duration of the customer relationship (LRFM) could be included in the RFM model. In the current version of the study, due to the manual required effort to obtain the contractual duration of the customers, this variable was not considered.

As future work, the study of the first approach should be expanded to accommodate the last three years of data (2016 - 2018).

References


MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA.


**Biographies**

Fátima Carneiro is a student of the Doctoral Program in Industrial Engineering and Management at Faculty of Engineering of the University of Porto (FEUP). She holds a master’s degree in Industrial Engineering and Management from FEUP. At the same time, she is a Project Leader at Kaizen Institute Western Europe with experience in the Field of Operational Excellence and Continuous Improvement Programs using Kaizen, Lean and 6 Sigma Methodologies. Her completed projects cover several areas, including Public Sector, Retail, Automobile, Metallurgy, Wood Transformation, Paper, Food, Wine, Pharmaceutical and Textile.

Vera L. Miguéis is an Assistant Professor in the Department of Industrial Engineering and Management at the Faculty of Engineering of the University of Porto, Portugal. She received her Ph.D. in Industrial Engineering and Management from the same school. Invited Researcher at Faculty of Economics and Business Administration of Ghent University and Tepper School of Business of Carnegie Mellon University. She is also a researcher at INESC TEC, and her research interests include data analytics and quantitative methods to support the decision making process. She has been mainly working on analytical customer relationship management and data mining. She has published papers in several international journals in Computer Science. She is the external relations manager of the Industrial Engineering and Management Master of the School of Engineering of the University of Porto.