Supermarket customer segmentation: a case study in a large Brazilian retail chain

Wellerson V. Oliveira
Supermercados Nordestão / UFRN
Natal, RN, Brazil
wellerson@nordestao.com.br

Daniel S. A. Araújo
IMD, UFRN
Natal, RN, Brazil
daniel@imd.ufrn.br

Leonardo C. T. Bezerra
IMD, UFRN
Natal, RN, Brazil
leobezerra@imd.ufrn.br

***Abstract*—In order to obtain commercial advantages over competitors, companies in all segments are improving their *customer relationship management* (CRM). The supermarket segment is no different, and investments in CRM are increasing over the last years. The first step towards a successful CRM strategy is to know customers better, for which *customer segmentation* plays an important role. In this work, we segment customers from Nordestão, the third largest supermarket chain in the Northeast of Brazil. To do so, we adapt the recency-frequency-monetary model, enrich it with new features, and use Gaussian mixture models to cluster the data. Furthermore, we employ a well-established *a priori* segmentation from the Brazilian supermarket literature. Our segmentation considers stores individually, and for each store we further refine its *a priori* segments into customer groups, with each group representing a different customer profile. Among the most interesting are *prime* and *opportunity* customers, who respectively focus on high-end and on-sale products. Importantly, most of the behaviours are consistent across the different stores, varying only as to store-specific parameters. We conclude our work with a further algorithmic validation and interpretability analysis of our findings.**

***Index Terms*—customer relationship management, customer segmentation, unsupervised learning, retail supermarket**

I. Introduction

In the last years, the world has been experiencing a transformation in the relationship between companies and their customers, largely motivated by the need to get commercial advantages over competitors. In part, this transformation is explained by the recent escalation of *customer relationship management* (CRM [1]). Within CRM, an important subject is *customer segmentation* [2], i.e. splitting customers into segments that “can be targeted in the same manner, because they have similar needs or preferences”. In this context, customer similarity is defined as a function of the available customer characteristics a company collects. Splitting customers into segments helps companies execute marketing strategies, and is an effective method to satisfy customer needs [3]. Furthermore, a good market segmentation enables customer behavior prediction, among other advantages [4].

Beyond the importance for industry, customer segmentation is also an academically-explored subject. Indeed, customer segmentation is one of the main applications for machine learning algorithms, and it is common to see publications applying these techniques as case studies for companies [5]. Our work fits this industry-academia collaboration context. Specifically, in this work we segment customers from the Nordestão supermarket chain in the Northeast region of Brazil. Nordestão has nine retail and two wholesale stores spanning three different municipalities within the metropolitan area of Natal, the capital city of the state of Rio Grande do Norte, and has recently started its expansion to nearby states. As a chain, Nordestão is the third largest in sales in the Northeast of Brazil, and the 27th in the country [6].

To perform customer segmentation, we adapt the recency-frequency-monetary (RFM) model [5] to our context. In more detail, we conduct our analysis on 2019 data, given the effects of the COVID-19 pandemics on more recent data. For this reason, recency is not included in our analysis as a feature for clustering. Frequency and monetary values are averaged from the data available for a 6-month period. In detail, monetary values are represented by average ticket item price, total value, and total item quantity. Furthermore, we enrich this adapted RFM model with (i) the ratio of items on sale, and (ii) a measure of item diversity.

Besides the enriched RFM model, we also adopt an *a priori* segmentation well-established in the Brazilian supermarket retail scenario [7], by splitting customers into *full* and *complementary* segments. In particular, full customers get their groceries primarily from Nordestão, whereas complementary customers use Nordestão as a secondary supplier. To cluster customers within each of the *a priori* segments, we adopt Gaussian mixture models (GMM [8]), which can isolate outliers and provide a model significant to this analysis. We cluster customers individually per retail store given their particularities, but observe a similar pattern across them. In fact, profiles found for most stores differ only as to the inherent parameters of each store, such as average ticket item price.

For complementary customers, five major profiles were found. Among the most insightful are (i) *specific*, who have a high average ticket item price, and; (ii) *opportunity*, i.e., customers with a high ratio of items on sale in their tickets. All profiles observed for the complementary customer segment were also identified for the full customer segment, with the addition of a new one, named *prime*, i.e. customers with a high average item price and average item quantity. Importantly, we also discuss profile variations from the perspective of store characteristics. For instance, we observe that the number of

This research was partially funded by Supermercados Nordestão

specific customers is higher in stores placed in residential areas than in stores placed in commercial areas.

Finally, in order to validate our results from an algorithmic perspective, we compare our findings with the outcomes of two alternative algorithms and observe that our conclusions are quite stable even if we replace GMM for the k-means algorithm. Furthermore, we conduct a multi-dimensional visualization assessment of the clusters produced by GMM using principal component analysis (PCA [9]). Besides confirming the separation between clusters, this assessment helped us better understand feature importance. For instance, complementary customer clusters differ primarily by frequency, and secondarily by on-sale item ratio.

The remainder of this work is structured as follows. Section II briefly reviews background concepts, including the literature on customer segmentation. The sections that follow address the major steps of CRISP-DM methodology [10], which we adopt in this work. Specifically, Section III details the Nordestão chain and its available customer data. In Section IV, we describe our proposed approach, delimiting data preparation and modeling. Evaluation from business and algorithmic perspectives are respectively given in Sections V and VI. Finally, we discuss deployment and future work possibilities in Section VII.

II. Background and related work

In this section, we briefly review the literature on customer segmentation. In detail, we initially discuss clustering applications to customer segmentation. Later, we review related works in retail, highlighting the Brazilian supermarket context.

A. Clustering and customer segmentation

Clustering algorithms are unsupervised machine learning (ML) approaches that receive unlabeled data as input and try to group these examples into homogeneous groups. Among the best-established families of clustering algorithms are *centroid*-, *density*-, and *distribution*-based. One of the main applications of clustering algorithms is customer segmentation [11]. Though clustering is a relatively recent approach, customer segmentation has been addressed for several decades now [12], defined as an approach to view a heterogeneous market as a combination of smaller, homogeneous markets.

The most widely used model in the context of customer segmentation is known as *recency-frequency-monetary* (RFM), along with its variations. The original RFM model consists in calculating these three features for each customer and then using a clustering algorithm to segment customers [5]. As surveyed by [8], most of the work in customer segmentation is performed using heuristic procedures, the most common being k-means. In addition, some pitfalls identified by the authors included a (i) small sample size of the data available, and (ii) lack of discussion of techniques used to pre-process data. Also, the authors point that clustering algorithms are usually selected for their simplicity, without taking into account more elaborate methods.

Though widely adopted in the literature, RFM and k-means are not the only approaches observed. Below, we briefly summarize alternatives that can be identified:

Features. RFM models have been replaced and/or enriched by other business features in different works [7], [11], [13]. For instance, [7], [13] use product categories that are bought by customers. Another example is [11], where authors use age and income for bank customer segmentation.

Algorithms. Though it is possible to use customized algorithms [13], most works in the customer segmentation literature adopt traditional approaches [8], [11], [14]. For instance, [11] compare k-means and DBSCAN in the context of bank customer segmentation. Other works adopt GMMs [8], [14]. In particular, [8] argues that GMMs provide, beside segments, a statistical model for the analysis of segment composition.

Though GMMs have a wide background literature and several applications in customer segmentation [8], [14], we have not identified case studies applying GMMs to customer segmentation for supermarkets.

B. Customer segmentation in retail

The literature on customer segmentation in retail is rich, including examples of supermarket chains around the world. For example, [15] proposed an integrated method for customer segmentation in a Chinese supermarket chain. The proposed approach combines both demographic and transactional data from customers. In detail, information such as gender, age, and marital status of the customers are used to generate a first segmentation. A second segmentation is performed using transactional data and a combination of PCA and k-means. Finally, the resultant groups are combined and another k-means clustering comprises the final segmentation, resulting in a set of 27 clusters.

Adopting a multi-stage segmentation is also the approach used by [16] to segment customers from an e-commerce store. Specifically, authors have used k-means as clustering algorithm to segment customers based on RFM features. The study discusses how to define each of the RFM dimensions, with the authors performing a two-stage segmentation: one for recency and another for the combination of frequency and monetary. Recency segmentation generated three clusters based on the number of days since the last purchase: active, lapsing, and lapsed customers. For each of these groups, k-means clustering was performed using the frequency and monetary dimensions. As a result, authors were able to find ten customer segments. Furthermore, authors also proposed and evaluated a strategy for each cluster which included an SMS campaign using different discount levels.

Regarding Brazil, the literature on customer segmentation for supermarkets is limited, but relevant. The only study identified targeted the largest Brazilian retail chain, Grupo Pão de Açúcar (GPA) [7], and discussed how to improve sales investing in CRM. One of the strategies adopted was to segment customers, though not much information is disclosed on the techniques considered. In line with the works

TABLE I: Retail stores considered and their characteristics.

Store	Neighborhood	Area	Mix	Value (%)	Ticket (%)
A	Residential	1.00	9795	5.38	6.23
B	Commercial	1.17	9618	5.86	6.54
C	Commercial	2.76	16705	19.93	18.81
D	Commercial	2.25	16282	12.31	12.03
E	Residential	2.03	10721	10.27	10.81
F	Residential	1.21	9302	4.91	5.47
G	Residential	2.57	16317	17.82	16.32
H	Residential	2.34	16281	13.97	14.81
I	Commercial	2.14	16320	9.55	8.99

TABLE II: Customer transaction features adopted in this work. Each feature is averaged over the 6-month period considered.

Label	Description
frequency	Number of tickets
value	Total value of the ticket
quantity	Total number of items in a ticket
item_price	Ratio between value and quantity
sale_ratio	Ratio between ticket items on and off sale
diversity	Number of different product categories in a ticket

previously discussed, GPA also employed a two-stage segmentation approach, addressing the dimensions of the RFM model individually. Given the lack of work in this area in the Brazilian supermarket field, this work aims to fill this gap, providing an algorithm-based analysis for customer segmentation. First, three broad customer categories were defined, based on increasing ticket item diversity, namely (i) *sporadic*, (ii) *complementary*, and (iii) *full* customers. Each category was further refined, though only the constitution of the latter two are described. Concretely, complementary customers were further segmented based on product categories, whereas full customers were further segmented based on frequency. The reported segmentation is used as support to match marketing actions to the group that should receive it. For instance, promotions on wine and organic products targeted party and healthy customers, respectively.

As discussed in this section, different techniques have been employed for customer segmentation, but less so in the context of Brazilian supermarkets. In the next sections, we follow the CRISP-DM methodology [10] to segment customers from the Nordestão supermarket chain.

III. Business and data understanding

As previously discussed, our work proposes a customer segmentation approach for Nordestão. In this section, we initially detail the chain and its data. Later, we conduct an exploratory analysis to identify features useful for modeling.

A. Business and data description

The Nordestão supermarket chain holds the third-largest sales value among all supermarkets chains in the Northeast of Brazil, and the 27th in Brazil [6]. The chain comprises nine retail and two wholesale stores located in the metropolitan area of Natal, Rio Grande do Norte. Despite the presence of the two greatest Brazilian supermarkets chains, Nordestão holds a market share of around 65%. Below, we detail the retail stores and available customer data we consider.

Stores. Table I describes stores in order of their inauguration year, labeled from A to I for anonymity. For each store, we provide (i) a description of its neighborhood; (ii) the number of distinct products available (labeled *mix*), and; (iii) its built-up area, relative to the area of the smallest store. The remaining characteristics given on Table I will be discussed later in this section. Overall, the stores are nearly equally distributed over commercial and residential areas. Furthermore, the product mix of a store is defined by based on features such as area, customer profile, and supplier negotiation. Therefore, it is natural that product mix and area show some correlation.

Data. The data used in our work comes from the fidelity program Nordestão hosts, where identified tickets accumulate points for a monthly lottery. In total, about half of the 30,000 daily tickets from Nordestão stores are identified. We restrict our analysis to data collected between 02/01/2019 and 30/06/2019, for two major reasons. First, the company changed its internal information technology solutions a few years ago, and hence some of the differences between the systems would affect our assessment. Second, data more recent than 2019 would be hindered by the COVID-19 pandemic effects.

B. Data engineering

The data retrieved from the customer identified purchase database comprise product, price, and quantities. From this data, we produced the features given in Table II, as a result of filtering, aggregating, and enriching ticket data:

Filtering. We removed customers with less than three tickets in the period assessed, since their purchase history is too small to model their behavior. In addition, we removed tickets with at least 200 units of the same product. This is a company rule used to identify small companies such as groceries and hotels who shop at the chain sporadically and do not represent the customer behaviour the company would like to model.

Aggregation. We aggregate the data in two steps. First, we aggregate product-level data into a ticket-level granularity. Our rationale is that a product-level segmentation would increase profile model complexity. Second, we aggregate ticket-level data to obtain customer-level features. The *frequency* feature for a given customer is computed as the six-month average of his/her monthly number of tickets. By contrast, a monetary feature is computed as a 6-month average of per-customer monthly averages. Specifically, we adopt (i) *ticket value* and (ii) *total item quantity*. Due to the COVID-19 effects on sales, we leave the recency dimension for future work.

Enrichment. We engineered three additional features to enrich our RFM model. The first is *average item price*, a feature that can help distinguish customers who usually buy first-need products from customers that buy expensive items. The second is the *ratio of items on sale*, since on-sale products are expected to have some influence on customer behaviour. Finally, we measure *item diversity* as the number of different product categories in a ticket, following [7].

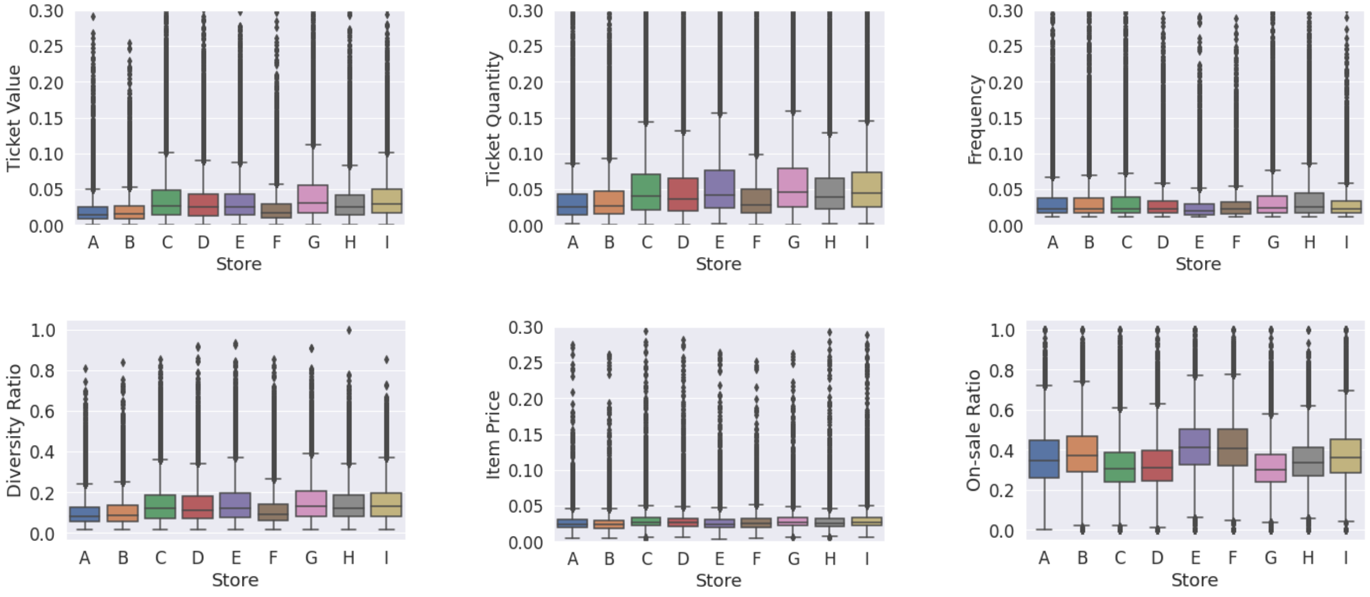


Fig. 1: Univariate feature analysis: boxplots depicting individual retail stores. Ranges were chosen for clarity.

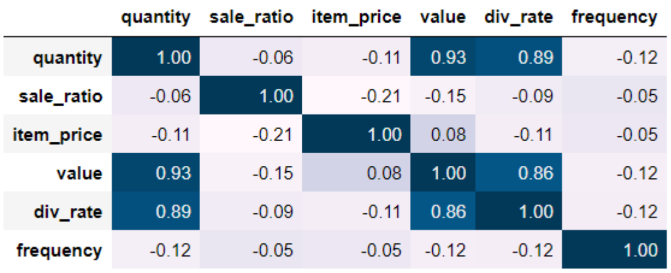
C. Exploratory data analysis

To gain a better understanding of the features we have engineered, we conduct exploratory analyses from both an uni- and multivariate perspectives. To aid this discussion, we initially detail the remaining store characteristics from Table I.

Value and ticket participation are the ratios between the given variable in each store and the total amount of the chain. For example, store B represents 5.86% of the total amount of sales and 6.54% of the total amount of tickets of the chain in the period assessed (after filtering). Importantly, we observe that these features are highly correlated between themselves and also with store size.

Univariate analysis. Boxplots for the features engineered are given in Figure 1, in which y-axis ranges have been scaled for confidentiality and clipped for clarity. Except for on-sale ratio, distributions are strongly asymmetric, with no clear distinction between stores for frequency and average item price. Interestingly, on-sale ratio values are generally higher for stores with smaller participation ratios. Finally, the between-store variation pattern observed for ticket value, quantity, and diversity partially matches store characteristics previously discussed, as the smaller stores present smaller feature values. The distinction between medium-sized and large stores is not as clear, though.

Multivariate analysis. Figure 2 presents Pearson correlation between features in the given row and column. The value, quantity, and diversity features are highly correlated between themselves. Yet, the correlation between value and diversity is slightly smaller than the correlation between value and quantity, since it is possible to increase these latter features while keeping diversity unaltered. Among the remaining feature pairs, correlation is either non-existing or non-significant.



	quantity	sale_ratio	item_price	value	div_rate	frequency
quantity	1.00	-0.06	-0.11	0.93	0.89	-0.12
sale_ratio	-0.06	1.00	-0.21	-0.15	-0.09	-0.05
item_price	-0.11	-0.21	1.00	0.08	-0.11	-0.05
value	0.93	-0.15	0.08	1.00	0.86	-0.12
div_rate	0.89	-0.09	-0.11	0.86	1.00	-0.12
frequency	-0.12	-0.05	-0.05	-0.12	-0.12	1.00

Fig. 2: Pearson correlation coefficient for all pairs of features.

As discussed in this section, Nordestão collects customer data that enables segmentation. Yet, further data preparation is required to improve modeling, which we discuss next.

IV. Data preparation and modeling

The customer segmentation approach we propose for the Nordestão supermarket chain builds on the insights from the literature review and the exploratory analysis. Initially, we perform an *a priori* segmentation [7], and select and prepare uncorrelated features for modeling. Later, data is clustered using GMM, as detailed below.

***A priori* segmentation.** Given its business relevance, we adopt the *a priori* segmentation approach proposed by GPA [7]. Sporadic customers were already filtered from the database during data engineering. We split the remaining customers into complementary and full based on the monthly number of distinct product categories a customer buys, averaged over the 6-month period considered. Concretely, customers below the third quartile for this feature were labeled complementary, and customers above it were labeled full. The choice of the third quartile as a threshold was empirical.

TABLE III: Number of clusters per store and *a priori* segment.

	A	B	C	D	E	F	G	H	I
Complementary	5	4	5	5	6	5	5	5	5
Full	3	7	6	6	5	3	7	5	6

Feature selection. Our modified RFM model comprises frequency along with three monetary features, namely quantity, item_price and sale_ratio. We choose quantity over value and diversity for two reasons. First, quantity is expected to vary less among different stores than value, and also promotes confidentiality. Second, the *a priori* segmentation already used a metric similar to diversity, reducing its per-segment variance.
Outlier analysis. Three of the remaining features are likely to present some outliers, namely frequency, quantity, and item_price. The first two have been addressed to some extent during data filtering. Regarding item_price, we can see that extremely high values appear only in customers with a low frequency and usually in tickets with only one product. Since those tickets represent circa 1% of all tickets, we decided to remove them using the inter-quartile (IQR) distance approach. In detail, for each store and *a priori* segment, all tickets with an item_price value greater than the given third quartile plus 1.5 times the given IQR were removed from the given dataset.
Data transformation. The transformations we perform are conducted on a per-store and *a priori* segment basis, split into two steps. First, a logarithmic transformation is applied to each feature, to render the tail and log-normal distributions more similar to a normal distribution. Next, we scale the resulting data to the $[0, 1]$ range, making every feature range identical.
Clustering. We fit a GMM model to each store/*a priori* segment combination. The number of clusters $k \in \{2, \ldots, 20\}$ is configured for each combination using BIC analysis, and is summarized in Table III. In particular, we choose BIC to optimize the number of clusters because it provides a balance between error variance and number of clusters in the model, producing a more unbiased result.Nearly all stores present $k = 5$ complementary customer clusters, stores B and E being the exception. We observe a higher variance in the number of full customer clusters, with no clear pattern regarding store characteristics. For instance, the smallest stores include both the smallest (stores A and F) and the largest (store B) k values.

In this section, we have proposed a customer segmentation pipeline for Nordestão. The next sections respectively evaluate this approach from a business and an algorithmic perspectives.

V. Evaluation from a business perspective

To evaluate customer segmentation from a business perspective, we initially identify customer profiles that repeat across the different stores considered. Later, we discuss the between-store variations we observe, trying to understand whether the features considered had an impact on the differences between store profile outcomes. Finally, we discuss customer belonging, one of the major benefits of employing GMM as clustering algorithm.

TABLE IV: Average feature values for complementary (top) and full (bottom) customers identified in store C. Q: quantity; SR: sale_ratio; IP: item_price; F: frequency. Cluster cardinality (N) is also given, both absolute and relative to the number of customers in the given *a priori* segment.

Complementary customers

Q	SR	IP	F	N	N (%)	Profile
6.26	0.24	1.00	1.62	3546	18.40%	Specific
7.22	0.33	0.86	5.66	2300	11.90%	Frequent
11.22	0.54	0.88	1.86	2809	14.60%	Opportunity
16.97	0.30	0.93	2.90	4545	23.60%	Regular
25.94	0.30	0.92	1.44	6091	31.60%	Large

Full customers

Q	SR	IP	F	N	N (%)	Profile
15.77	0.34	0.67	13.23	620	9.70%	Frequent
23.01	0.26	0.88	7.27	979	15.30%	Specific
32.95	0.42	0.65	3.95	760	11.90%	Opportunity
40.37	0.28	0.72	4.11	2145	33.60%	Regular
43.61	0.27	1.00	2.47	667	10.50%	Prime
70.36	0.30	0.66	1.60	1210	19.00%	Large

A. Overall profile analysis

Though customer segmentation was performed for each store and *a priori* segment, we observe common profiles across stores. In addition, some of the profiles we observe are present in both *a priori* segments. We then initially discuss the profiles we observe more often among complementary customers, and later comment on the differences regarding full customers. For brevity, we illustrate the main profiles we observe among complementary customers with results from store C, and later comment on how they differ from full customer profiles.

Table IV (top) provides complementary cluster mean feature values for store C, which are also given as radar charts in Figure 3. Each profile was labeled to help its understanding. We make a few remarks about the data prior to discussing results. First, the values given in Table IV are untransformed values, for interpretability. In detail, we provide cluster means prior to logarithmic transformation or scaling. The only exception is item_price, for which values have been scaled for confidentiality (though not subject to a logarithmic transformation). Second, profiles are sorted as to quantity values. Finally, cardinality (N) is included to understand how representative a given profile is, both absolute and relative to the total number of customers in the given store/segment combination. Next, we discuss each profile identified and its potential CRM impact.

Specific customers only shop selected items. This is indicated by the low on-sale item ratio in contrast to a high average item price and low item quantity. One possible explanation is that these customers shop for items that are not available in other supermarket chains, which are less often on sale. Marketing campaigns targeting this customer profile should then focus on their preferred items rather than on items on sale.
Frequent customers shop much more often than the remaining profiles. Across all stores, customers from this profile usually shop on a weekly frequency. Being frequent, their tickets are usually small in value and moderate as to the

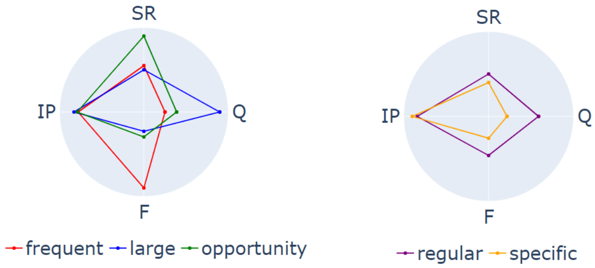


Fig. 3: Radar chart depicting store C complementary customer profiles. Ranges in both charts are the same for comparability.

average item price. In this context, marketing campaigns targeting these customers should likely focus on increasing their ticket quantities rather than bringing them to the stores.

Opportunity customers stand out as to their ticket on-sale item ratio. Indeed, the mean value for this feature in this profile is 54%, but the highest values reach circa 65%. As a result of the high on-sale item ratio, the average item price for their tickets are usually low. In principle, this profile should be the most commonly targeted by discount campaigns, as they demonstrate significant interest in offers.

Regular customers present moderate values for all features considered. In a sense, they represent the regular behavior expected for complementary customers. Indeed, this is one of the largest profiles observed among complementary customers, and could be used if marketing campaigns were to target complementary customers in general.

Large shoppers generally present the highest ticket quantities among all complementary profiles. As such, the frequency values for this profile are usually the lowest. Marketing campaigns targeting these customers would benefit from predictive analysis of when these customers come to the store.

The complementary customer profiles discussed above do not differ considerably from the full customer profiles given in Table IV (bottom) and in Figure 4, except for the absolute ticket quantities observed. Below, we comment on the profiles for which we observe interesting insights not previously discussed for complementary customers.

Frequent full customers shop with a very high frequency, namely two or three times a week. Given that frequent full customers already buy a significant number of product categories, marketing campaigns targeting this profile should focus on increasing their average ticket item price.

Prime full customers present a high average item price. Differently from specific customers, prime customers also present a high item quantity. Though this customer set is not among the largest within full customers (and was not directly observed among complementary customers), prime full customers are likely the most profitable segment we identified among Nordestão customers. Marketing campaigns for prime customers should then focus on high-end products.

Large shopper full customers fit the opposite description of frequent full customers. Besides having an average item quantity which is generally the largest among all segments, customers from this profile shop sometimes less often than every other week. As such, predicting when these customers will come to the store is even more critical than what was previously discussed for large shopper complementary customers.

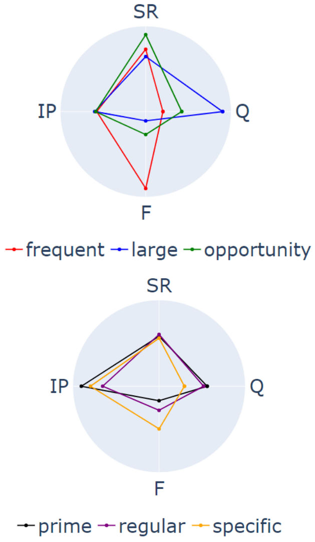


Fig. 4: Radar chart depicting store C full customer profiles. Ranges in both charts are the same for comparability.

B. Store-level analysis

The profiles discussed above are observed across the different stores. However, not all profiles are present at every store, or profiles may present slight differences in mean values. Those differences may be explained by store characteristics, such as those presented in Section III. We then further investigate customer profiles from a per-store perspective. For brevity, the full description of store profiles is compiled as a business report that is not disclosed for confidentiality. Here, we discuss only the most relevant insights observed, grouped by the feature and store characteristic interaction they concern.

Frequency and store size. Figure 5 (left) presents boxplots of the per-store average frequency values observed for the frequent complementary and full customer profiles, grouped by store size. The frequency with which these complementary customers attend stores increases with store size. Though further investigation would be required to understand the reason behind this interaction, we conjecture that a larger store promotes a better customer experience, with wider aisles and better parking facilities, thus increasing the frequency with which customers visit the store. When we assess store size effects on frequent full customers, we see that large and medium-sized stores also follow the pattern of increased frequency for increased store size. The exception concerns small stores, for which variance is high.

Average item price and store neighborhood. Figure 5 (right) illustrates the per-store proportion of specific complementary and full customers w.r.t. total complementary and full customers for the given store, respectively. The impact of

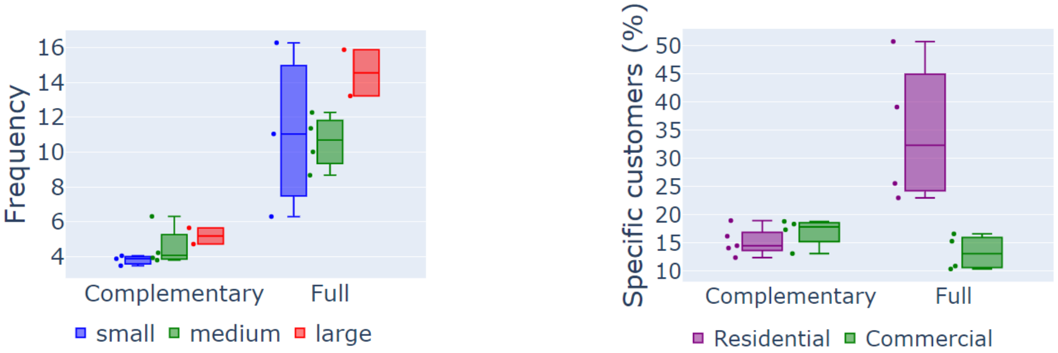


Fig. 5: Per-store profile analysis and interactions with store characteristics. Left: average frequency of the frequent customers. Right: proportion of specific customers.

store neighborhood in these proportions vary as a function of *a priori* segment. In detail, commercial-neighborhood stores present a higher median proportion of specific complementary customers than residential stores, though their difference in distribution is not strong. This effect is reversed for specific full customers, with the difference in proportions being much more striking. In this context, marketing campaigns targeting specific customers should reflect these interactions.

C. *Customer belonging*

As previously discussed, GMM clusters may overlap, since models comprise Gaussian components. As such, the GMM model provides probability scores $p_{ij} \in [0, 1]$, where a higher score indicates a higher probability of customer i belonging to cluster j. In addition, $\sum_j p_i = 1$, so a high probability of belonging to a given cluster also implies a low probability of belonging to other clusters. Below, we first perform an overall belonging analysis both for complementary and full customers.

Complementary customer profiles are given in Figure 6 as histograms of the per-customer probability scores from store C. In general, each of the distributions given is shifted to the right. The extreme situations are observed for the opportunity and frequent profiles, which present a tail distribution and a peak near one, representing customers with an almost perfect profile match. Furthermore, the regular profile presents a distribution very similar to the distribution of the large shopper profile.

Full customer profiles from store C are given in Figure 7. Note how (i) the distribution for prime customers resembles the distribution for the opportunity customer; (ii) the distribution for frequent full costumers resembles the distribution for frequent complementary customers, and; (iii) only for large shoppers the distribution is right-shifted. Understandably, the regular profile presents a left-shifted distribution, as this is the baseline full customer profile.

The insights above reveal that many customers belong to more than one profile, especially full. To further assess customer belonging, we next define *strong belonging*, which renders results more reliable from a business perspective.

Univariate rule: customers whose highest cluster probability score exceed 0.7. Our rationale is that the second-highest score for such a customer would be at least 50% smaller than the first. In addition, we remove customers whose highest probability score are lower than 0.4, to ensure that customers with a set of low scores do not affect the following rule.

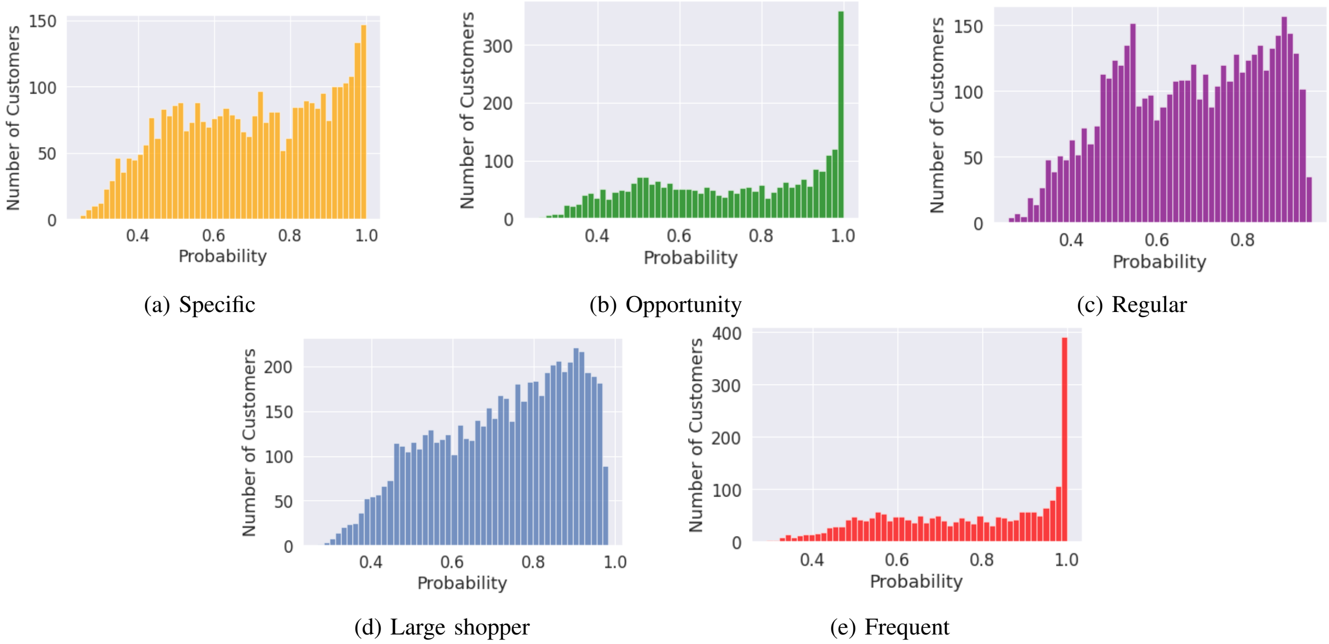
Multivariate rule: customers whose highest probability score is at least two times their second highest score. Our rationale is that even customers for which the highest score is between 40% and 70% may be proportionally strongly belonging, if their remaining probability scores are low.

Figure 8 presents the proportion of strongly belonging customers we identify for each complementary (top) and full (bottom) customer profiles for all stores. In the complementary group, we observe that median values are always above 65%, which indicate that cluster belonging is strong for the majority of the customers assessed. Interestingly, for the frequent and large clusters, median values are at least 75%, and no profile from any store ever presents less than 50% of strongly belonging customers. A similar pattern is observed for full customers, though values change slightly. In detail, (i) the median values for all profiles are always above circa 60%, and; (ii) the median values for frequent and large shopper profiles are always above 70%.

In this section, we have evaluated our proposed customer segmentation from a business perspective. Besides insightful profiles, we have discussed their interaction with store characteristics and how customers may be strongly belonging or not. In the next section, we evaluate our approach from an algorithmic perspective.

VI. Evaluation from an Algorithmic Perspective

Given that customer segmentation builds on clustering, validating results from an algorithmic perspective is essential. In this section, we first conduct a robustness assessment to investigate the impact of changing clustering algorithm. Later, we use principal component analysis (PCA) to assess cluster separation and investigate feature importance.



(a) Specific

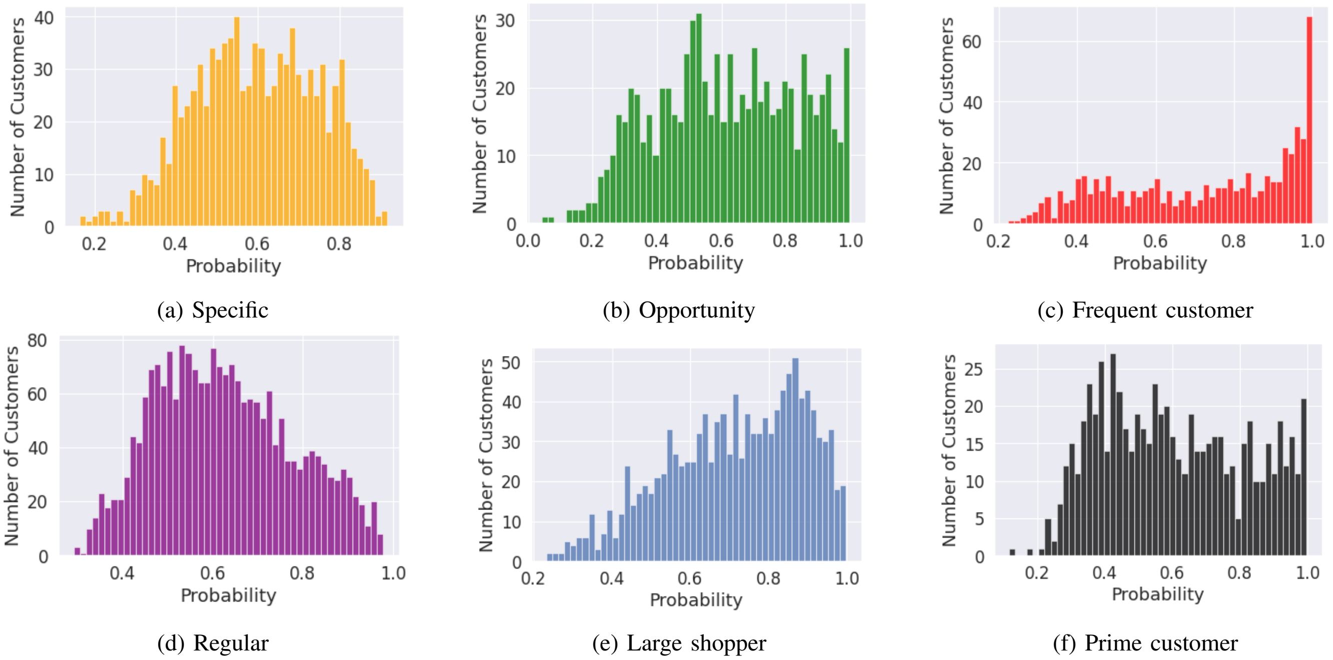
(b) Opportunity

(c) Regular

(d) Large shopper

(e) Frequent

Fig. 6: Probability distribution analysis for complementary customer profiles from store C.



(a) Specific

(b) Opportunity

(c) Frequent customer

(d) Regular

(e) Large shopper

(f) Prime customer

Fig. 7: Probability distribution analysis for full customer profiles from store C.

Alternative algorithms. For most stores and *a priori* segment combinations, preliminary experiments show that DBSCAN produces a single cluster and regards the remaining few points as outliers. For this reason, we restrict our comparison to GMM and k-means, given as boxplots of different validation metrics in Figure 9. We can see that for both silhouette and Calinsk-Harabasz (maximization), the median values obtained by k-means are slightly better than the ones obtained by GMM. However, significant difference is only observed for Davies-Bouldin (minimization). Given the performance similarity and the added benefit of customer belonging analysis, we consider GMM an appropriate choice for Nordestão.

Cluster separation and feature importance. To assess cluster separation, Figure 10 gives scatter plots of the complementary (left) and full (right) customer clusters using the first two principal components (PC1 and PC2) obtained for each analysis. For complementary customers, we can see from Figure 10 (left) that clusters are overall well-separated. Regarding

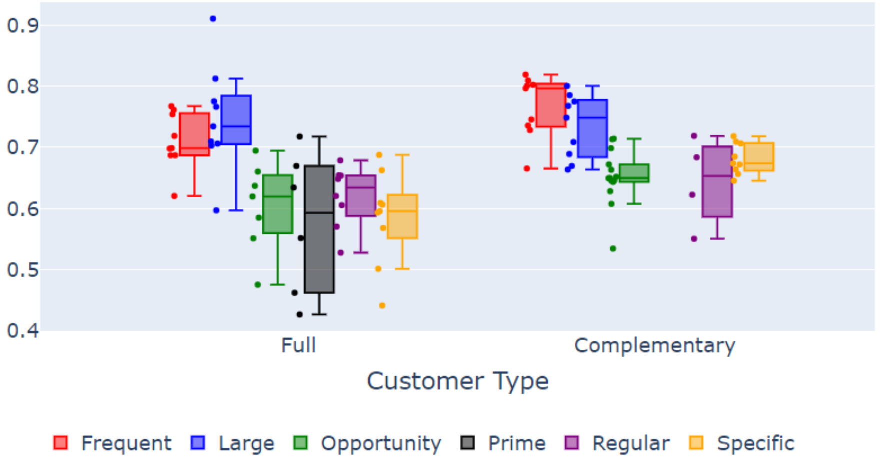
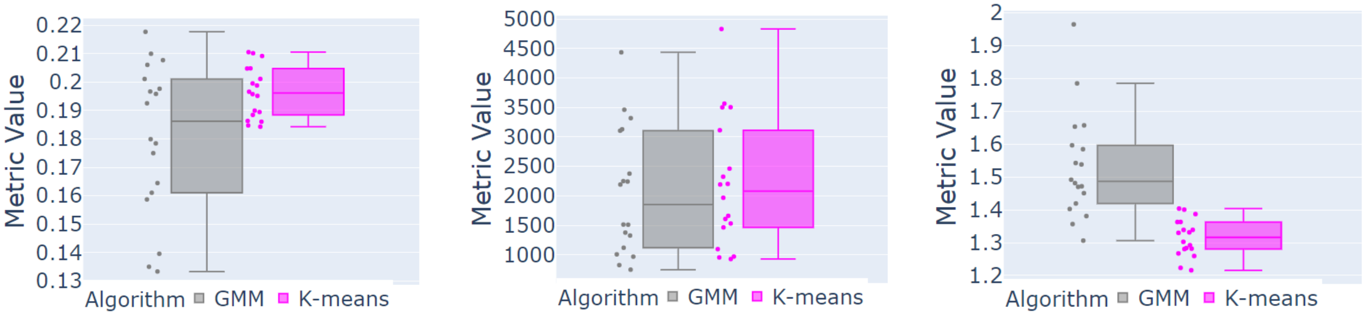


Fig. 8: Proportion of strongly belonging customers per profile, grouped by *a priori* segment.



(a) Silhouette

(b) Calinsk-Harabasz

(c) Davies-Bouldin

Fig. 9: Clustering metric boxplots for k-means and GMM, using results from all store and *a priori* segment combinations.

full customers, the separation given in Figure 10 (right) is not as clear, specially around the central region of the plot.

To provide interpretability into cluster separation and also understand feature importance, Figure 11 gives an assessment of the correlation between PCs and the original features for complementary (left) and full (right) customer clusters. For complementary customers, Figure 11 (left) shows that PC1 is strongly positively correlated with frequency, and to a lesser extent negatively with item_quantity. In turn, PC2 is highly positively correlated with sale_ratio, and to a lesser extent negatively with item_price. These results are coherent with the separation given in Figure 10 (left). In detail, along PC1 we observe the separation between large, regular, and frequent customers, which differ precisely as to the combination of frequency and item_quantity. In complement, the separation along PC2 splits specific, regular, and opportunity customers, which differ w.r.t. sale_ratio and item_price.

Finally, Figure 11 (right) shows results for full customer profiles, where correlation insights also match the scatter plot separation given in Figure 10 (right). In detail, PC1 is once again strongly positively correlated with frequency, but this time to an also high extent negatively with item_quantity. In the scatter plot given in Figure 10 (right), along PC1 we observe the separation between large and frequent customers on the opposite ends of the axis. Furthermore, closer to the center of the axis we see the separation between prime and specific customers, which differ precisely as to the combination of frequency and item_quantity. In turn, PC2 is highly negatively correlated with item_price, and to a lesser extent positively with sale_ratio. In the plot, the separation along PC2 contrasts prime and specific customers from opportunity customers, for which the combination of item_price and sale_ratio is critical.

VII. Conclusions

Customer segmentation is a central strategy within customer relationship management (CRM [1]). This is evidenced by the CRM literature, where a number of proposals build on knowledge of customer behavior [5]. In this work, we have proposed a customer segmentation approach for the Nordestão supermarket chain. Initially, we performed an *a priori* segmentation following the literature on customer segmentation for Brazilian supermarkets, splitting customers into *full* and *complementary* segments. For each *a priori* segment from a given store, we have used an adapted recent-frequency-monetary (RFM [5]) model and the GMM algorithm to characterize and cluster customers. Results showed that six customer profiles could by identified, namely *frequent*, *specific*, *regular*, *opportunity*, *prime*, and *large shopper*. Finally, we have provided likely

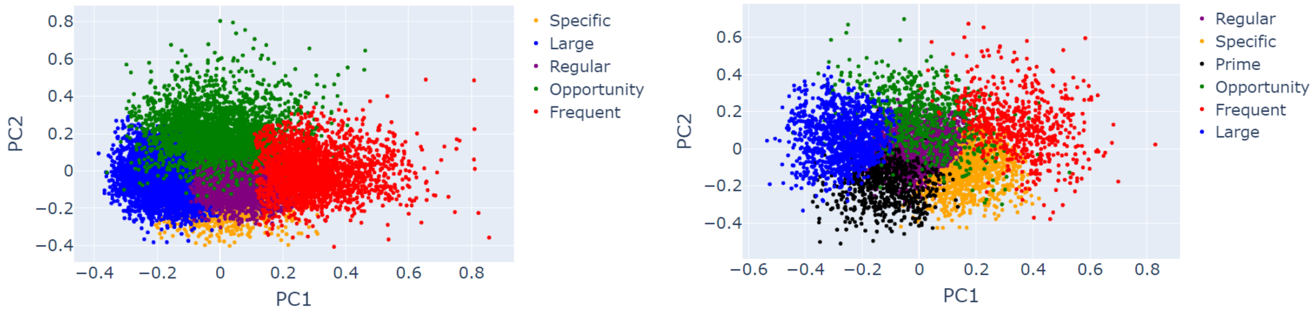
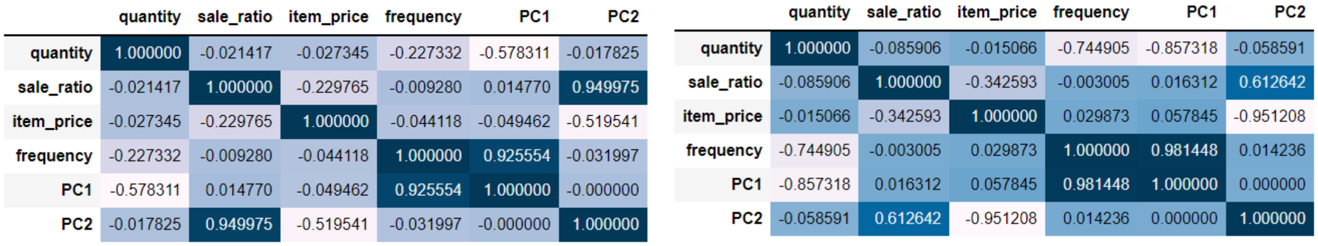


Fig. 10: Store C: PCA analysis to visualize complementary (top) and full (bottom) profile separation.



	quantity	sale_ratio	item_price	frequency	PC1	PC2
quantity	1.000000	-0.021417	-0.027345	-0.227332	-0.578311	-0.017825
sale_ratio	-0.021417	1.000000	-0.229765	-0.009280	0.014770	0.949975
item_price	-0.027345	-0.229765	1.000000	-0.044118	-0.049462	-0.519541
frequency	-0.227332	-0.009280	-0.044118	1.000000	0.925554	-0.031997
PC1	-0.578311	0.014770	-0.049462	0.925554	1.000000	-0.000000
PC2	-0.017825	0.949975	-0.519541	-0.031997	-0.000000	1.000000

	quantity	sale_ratio	item_price	frequency	PC1	PC2
quantity	1.000000	-0.085906	-0.015066	-0.744905	-0.857318	-0.058591
sale_ratio	-0.085906	1.000000	-0.342593	-0.003005	0.016312	0.612642
item_price	-0.015066	-0.342593	1.000000	0.029873	0.057845	-0.951208
frequency	-0.744905	-0.003005	0.029873	1.000000	0.981448	0.014236
PC1	-0.857318	0.016312	0.057845	0.981448	1.000000	0.000000
PC2	-0.058591	0.612642	-0.951208	0.014236	0.000000	1.000000

Fig. 11: Store C: correlation between principal components and features for complementary (left) and full (right) profiles.

explanations for the variations seen on results from a few stores, discussed cluster belonging, and assessed alternative algorithms and cluster separation.

The customer segmentation approach proposed in this work is seminal for a number of future investigations and strategical planning opportunities for Nordestão. For instance, the profiles identified in this work will be used by the marketing department to direct promotions for each customer group and store. In addition, the continuous monitoring of customer data will help the company assess whether these marketing campaigns are able to migrate customers between profiles. Also, the segmentation analysis proposed here should help company to better understand its business and which segments are more profitable. This knowledge may drive to a better business model, which can lead to a better relationship with customers and a increase of sales.

It is also important to remark that the segmentation approach we propose can be further improved. More information on product categories and store characteristics could help understand variations in profiles that this study could not address. Furthermore, demographic information such as age and gender could help further target marketing campaigns.

References

[1] A. Payne, *The handbook of CRM: Achieving excellence in customer management*. Elsevier Butterworth-Heinemann, 2006.

[2] M. Wedel and W. Kamakura, "Introduction to the special issue on market segmentation," *Int. J. Res. Mark.*, vol. 19, no. 9, pp. 181–183, 2002.

[3] P. P. Pramono, I. Surjandari, and E. Laoh, "Estimating customer segmentation based on customer lifetime value using two-stage clustering method," in *ICSSSM*, pp. 1–5, IEEE, 2019.

[4] M. N. Tuma, R. Decker, and S. W. Scholz, "A survey of the challenges and pifalls of cluster analysis application in market segmentation," *Int. J. Mark. Res.*, vol. 53, no. 3, pp. 391–414, 2011.

[5] D. Chen, S. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *J. Database Mark. & Cust. Strategy Manag.*, vol. 19, no. 3, pp. 197–208, 2012.

[6] "Ranking ABRAS 2020," https://www.abras.com.br/edicoes-anteriores/Main.php?MagNo=259, May 2020, accessed: 2021-05-06.

[7] V. Riegel and M. L. A. Pereira, "Pão de Açúcar Mais: O desafio do relacionamento," https://docplayer.com.br/69475-Central-de-cases-pao-de-acucar-mais-o-desafio-do-relacionamento-www-espm-br-centraldecases.html, November 2010, accessed: 2021-05-06.

[8] M. Tuma and R. Decker, "Finite mixture models in market segmentation: A review and suggestions for best practices," *Electron. J. Bus. Res. Methods*, vol. 11, no. 1, pp. 2–15, 2013.

[9] I. Jolliffe, "Principal component analysis: A beginner's guide," *Weather*, vol. 45, no. 10, pp. 375–382, 1990.

[10] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1. Springer-Verlag London, UK, 2000.

[11] D. Zakrzewska and J. Murlewski, "Clustering algorithms for bank customer segmentation," in *ISDA*, pp. 197–202, IEEE, 2005.

[12] W. R. Smith, "Product differentiation and market segmentation as alternative marketing strategies," *J. Mark.*, vol. 21, no. 1, pp. 3–8, 1956.

[13] X. Chen, Y. Fang, M. Yang, F. Nie, Z. Zhao, and J. Huang, "Purtreeclust: A clustering algorithm for customer segmentation from massive customer transaction data," *IEEE Trans. on Knowl. and Data Eng.*, vol. 30, no. 3, pp. 559–572, 2017.

[14] Z.-K. Huang and K.-W. Chau, "A new image thresholding method based on Gaussian mixture model," *Appl. Math. and Comput.*, vol. 205, no. 2, pp. 899–907, 2008.

[15] X. Bai, X. Xia, H. Wang, W. Yin, and J. Dong, "An integrated customer segmentation method for China's supermarkets," in *SOLI*, pp. 440–445, IEEE, 2010.

[16] M. Tavakoli, M. Molavi Hajiagha, V. Masoumi, M. Mobini, S. Etemad, and R. Rahmani, "Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: A case study," in *ICEBE*, pp. 119–126, IEEE, 2018.