Marketing Analytics

Project Work

Group 17

A.Y. 2022/2023
Professor: Lamberti Lucio
Assistant professor: Di Dalmazi Michele

Chiara Piccioni	*10866770*
Matteo Mascheroni	*10638118*
Giovanni Galeotti	*10619759*
Edoardo Gronda	*10826399*
Federico Cantarelli	*10596312*

POLITECNICO
MILANO 1863

Recency State Measure

In the ordinary RFM model, the *recency* measure is defined as the number of days that have passed from the last purchase of a given customer. This measure was developed and adopted originally for direct mailing campaigns because representing a good proxy for the customer's responding probability.

For our purposes, we redefined such measure in order to suit it to our case, namely a national general retailer. In order to do so, we firstly identified the weaknesses the "base" *recency* measure which are listed below:

- The *recency* cannot be weighted on the *frequency* with which a customer purchases. For instance, having a recency of 4 days should be evaluated differently for a customer base that buys every week or every other day.
- The measure does not consider the customer's regularity, which should be actually taken into consideration when evaluating its recency state.

To solve these issues, a new way of measuring the recency has been devised. How? Firstly, we wanted to compute a single value for it, but soon we realised that the *recency* and *regularity* were independent. Therefore, summarizing the two with a unique value would have reduced the measure's interpretability and caused an important loss of information. So, we opted for a matrix representation with the "*new recency*" as a dimension and the *regularity* as the other. After that we defined a "*new recency*" measure:

Recency_mod = recency(customer)/ average interpurchase time (customer)

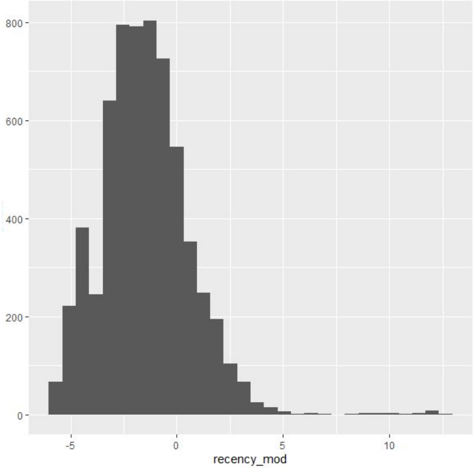
Regularity = standard deviation of interpurchase time(customer)/average interpurchase time (customer)

Recency State Measure

After computing these two measures, it has been necessary to find a way to cluster them into significative groups. For the regularity, we divided it into tertiles: From the lowest to the higher value we assigned these labels respectively: **Creature of habit, Regular, Unstable**.

For the new recency measure we were interested in dividing customers that have a recency smaller than their average interpurchase time and the one with recency higher than their average interpurchase time. To do so, we considered the **logarithm** of the new recency. This way, it has been possible to stretch the values lower than 1 and condense those larger than that. What we obtained consisted of an almost normal distribution as we can see from the graph (slightly screwed on the left with a longer tail on the right). Finally we created 4 groups by taking the median considering only the negative values and defined the median of that group as one limit (*llimit* in the code) and did the same thing for the values higher than 0 (*mlimit* in the code). We then assigned the labels:

```
recency_mod ≤ llimit ~ "Super Hot",
recency_mod > llimit & recency_mod ≤  0 ~ "Hot",
recency_mod > 0 & recency_mod ≤  mlimit ~ "Cold",
recency_mod > mlimit ~ "Super Cold"
```



Histogram of the recency_mod

This procedure allows us to represent the customer base on a matrix, we opted for 3 categories for the regularity and 4 for the recency because it was the best compromise in terms of quality of clusters and interpretability of the obtained matrix.

Exploratory Data Analysis - Descriptive statistics

The data sets supplied (MA_Assignment_transactions and Product DB) stem from the database of an **Italian retailer** and consists of a multitude of data representing partially its activities registered from the October 2020 to October 2022. More specifically,

1. **MA_Assignment_transactions**, composed of several variables which describe the transactions occurred. These are: Store ID, Date, Ticket number, Cash Desk number, Customer ID, Product code, quantity and price.
2. **Product DB**, where the product code is associated to its specific description.

Total number of customers	9998
Average frequency	16 times in 2 years
Average total price per basket	11.5 € per basket
Average basket size	5 products per basket

Database descriptive statistics

The **size** of the basket, both, in terms of **price** and **products bought**, let us think that these supermarkets are not primary retailers for the consumers.

Objective

Redefine the loyalty company’s programme (1 point for € spent) by focusing on the implementation of a **tier program** that provides different long terms offers to the most valuable customers in terms of profitability. To reach this goal, a RegFM analysis, taking into account the inter- and infra-clusters migrations, and a Market Basket Analysis have been carried out.

Data preprocessing

Before conducting the due analyses, we have gone through a required modification of the data set in such a way to allow R and the consequently exploited deployed functions to work properly, i.e. without returning any error related to the data type. In fact, due to the import function that has been used, namely *read.csv* the data were all read by the programme either as numeric or integers. This had to be modified in some peculiar cases.

1. **Date**: In order to allow to compute easility all the operations on data, such as differences between two or lags, this column had to be turned into a *date-time* object. This has been achieved by first converting the column type into character and then by exploiting the function *parse_date_time* contained in the library *lubridate*, specifying the exact type of output we wanted, i.e. a *date-time* object that showcased from the year to the second in which the customer had purchased a given ticket.

2. **Missing values**: No missing values were present in any of the 25617408 entries given, thus nothing had to be done with this respect.

3. **Variables as factor**: Some variables need to be treated as categorical. More specifically, *customer_id*, *store_id* and *cashdesk_no* have been turned into factors.

4. Unique key (**basket**): Due to the fact that tickets numbers are re-initialized after a given period, we had to come up with a method to allow to identify the specific Basket purchased by a given customer in a given point in time. To do so we created a new column containing a unique Key, lately renamed as *Basket,* made of several information, namely: *customer_id, ticket_no, store_id, date_ticket.* Still, before pasting these information within one single column, we made sure that all the values in each of the four columns respected a precise pattern, i.e. we performed a *gsub()* to guarantee for instance that no extra white spaces were existing. This was required because otherwise we could have obtained two different results, i.e. unique Keys, for the same Basket only because in the *date_ticket* of one of the contained an extra space. Note: we decided to avoid to use the *cash_desk* column, since we have found just one single entry in the entire data set that had a different *cash_desk* number but everything else equal. We considered this as a system's error, since we believed it was physically impossible that the same person has been able to buy at the same shop, at the same time (seconds) and has received the same *ticke_no* but at two different *cash_desks.* Thus, in the following analyses this case has been treated as unique, i.e. single entry.

5. **Negative tickets elimination**: Being -877.03 in total for 136 customer, thus ca. 6.5 € each in 2 years, we believed that there was no point in keeping these values. To be more precise, in 42 cases (tickets) we have found that both prices and quantity were negative, probably displaying some returned products. Despite this could have been an interesting aspect to analyze, we also believed that considering the extremely low number of cases with respect to the total of 63956 tickets, could not justify the efforts, moreover representing only -359.56 €.

RegFM analysis

In order to assess an actual buying behavior, we have chosen to perform the Regularity, Frequency and Monetary (RegFM) analysis and to use the *Recency_mod* measure as a personal variable for each customer. The historical value of the client is assumed to be an indicator of the client's future contributions. We have chosen the **quarter** as timeframe (**1st July- 30rd September 2022**), because of the sector in which the company operates and also due to the fact that it does not seem to be a primary retailer for the customers (low average frequency, low average monetary value of the ticket and low number of products for basket). Given the reasons mentioned above, it seems right to consider a customer as disappeared after 3 months.

For what concerns the variables selection we have formulated the *Monetary* as each customer’s total expenditure, the *Frequency* as the amount of repeated purchases (unique tickets) and the *Regularity* as the coefficient of variation, which is obtained as the ratio between the inter-purchase time ‘s standard deviation (representing the regularity state of a customer) with the mean (in order to scale it).

Variables selection and operationalisation	
Monetary	*Number of transactions * Money spent in a single transaction*
Frequency	*Number of tickets*
Regularity	*Coefficient of variation = sd(Inter-purchase time)/mean(Inter-purchase time),* *where* *Inter-purchase time = number of days between consecutives purchases*
Recency_mod	*Regularity * log(Recency/mean(Inter-purchase time)),* *where* *Recency = difference between last purchase and first day of the next month*

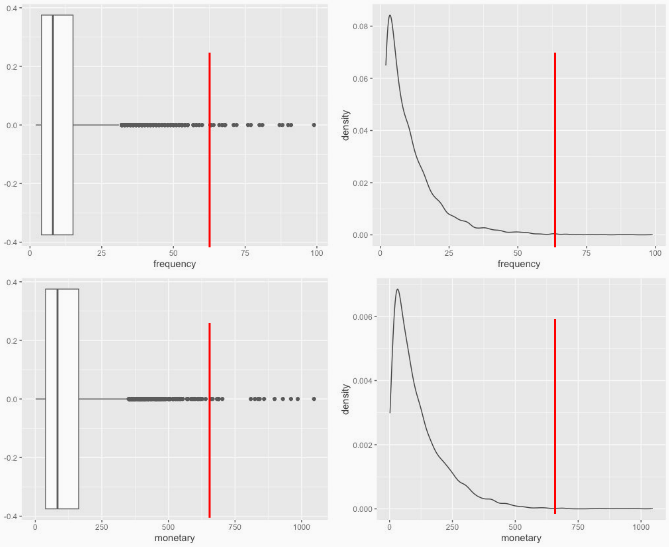
Anomaly detection, outliers and threshold definition

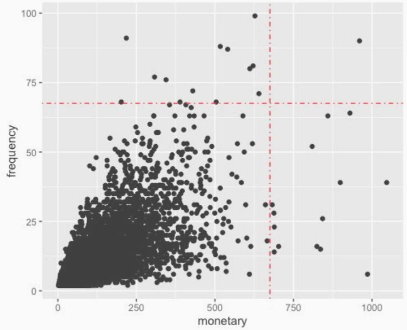
To identify the possible outliers, we have decided to follow a mixed approach. Starting from the interquartile range rule, the threshold have been defined where there was a flattening of the density function of the *Frequency* and *Monetary* variables. Indeed, customers will be considered as **outliers** if they have a *Frequency* value higher than 67 times in 90 days disposable (ca. 3/4) and a *Monetary* value higher than 675 € in total during the period. This process resulted in the elimination of 28 observations from the dataset. In addition, we have removed from the dataset the **one visit customers** (i.e. customers that had *Frequency* = 1) being them not the target of our analysis. We have decided to keep those customers whose frequency is greater than 1, despite having visited the supermarket during the same day, although they should have been considered as one visit customers as well (3 observations). Finally the threshold for the scoring attribution (1,2,3,4) have been defined algorithmically on the 1st quartile, the median and the 3rd quartile.

	Frequency	Monetary	Regularity
Minimum	2	1.67	0.3012
1st Quantile	**4**	**38.63**	**0.6459**
Median	**8**	**83.41**	**0.7617**
Mean	11.14	118.23	0.8517
3rd Quantile	**15**	**162.97**	**0.9619**
Maximum	99	1046.66	2.4033

Temperature measure

Based on the *Recency_mod* measure, we classified the intrinsic state of a customer. With this respect, though a proper counterintuitive logic had to be followed to arrive to the correct conclusions. Knowing that this newly-devised variable distribution is not symmetric, suggesting us that negative values are "hotter" than positive ones thus that values near to 0 are "preferred" than larger values in absolute terms; we came up with the decision to create 4 clusters using 2 medians: one for the "hot" side and one for the "cold" one. Although devised it, we came up with the conclusion that for our purposes, namely devising a **tier program**, which focuses on a medium-long- term objective, such measure would not be of any use. Despite this, we have noted a positive correlation between the tiers that we have created (see Slide 8) and temperature suggesting an higher retention rate for the top tiers.





Outliers identification

Target definition

As a result of the RegFM analysis, we obtained 64 different clusters based on the customers' estimated value. To achieve our objective **(tier program)** we identified 3 different classes focusing on their **profitability**. Our target will be the *primary* and *secondary* shoppers with the aim of increasing company's revenues and retention rates on a segment basis. Intuitively, the key issue in implementing this program is to develop actions that push lower tier customers into the upper ones.

Primary shopper

It is made of the customers that come often and spend much. In fact, they have been defined as the customer with the highest *Monetary* (4) and *Frequency* (4) values, regardless the *Regularity.* These customers accounts for the 57% of the company's revenues. The objective is to **retain** these ones and to increase the share of them. It should also be highlighted that in this group 740 *Super Hot* and 186 *Hot* customers are present, thus showing high level of retention.

Secondary shopper

It is composed of customers that have still high level of *Monetary* (3,4) but any any level of *Frequency* on the contrary of the *Primary*.. The *Secondary* and the *Primary* accounts for the 82% of the company's revenues and the focus here will be to increase their *Frequency* and *Monetary* levels, nudging them towards the first group.

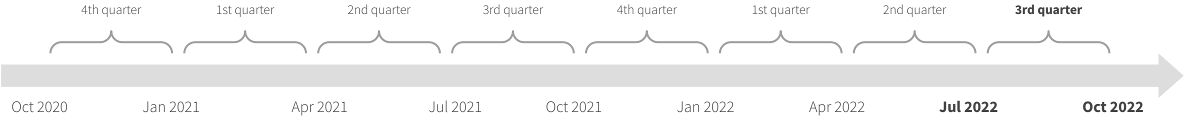
Non shopper

It is composed by customers that have very *Monetary* low value (1,2). This group is on average less regular, and has a low level of *Frequency*. The objective is to increase the *Monetary* and to push them to become *Secondary shopper*. In this cluster it is possible to identity two different sub-groups: the ***Potentially good customers***, that have high value of *Frequency* (3,4) composed of 511 subjects and the ***Not worth it*** that shows very low *Frequency* values (1) composed of 1641 customers.

	Av. monetary score	**Av. monetary**	**Av.regularity score**	**Av. frequency score**	**Av. frequency**	**Number of customers**
Primary	4	279,1 €	3,47	4	27,8 times	949
Secondary	3,28	152,3 €	2,83	2,76	11,5 times	2169
Non shopper	1,5	40,7 €	1,97	1,66	5,5 times	3118

Descriptive statistics of Primary, Secondary and Non Shopper

Longitudinal RegFM

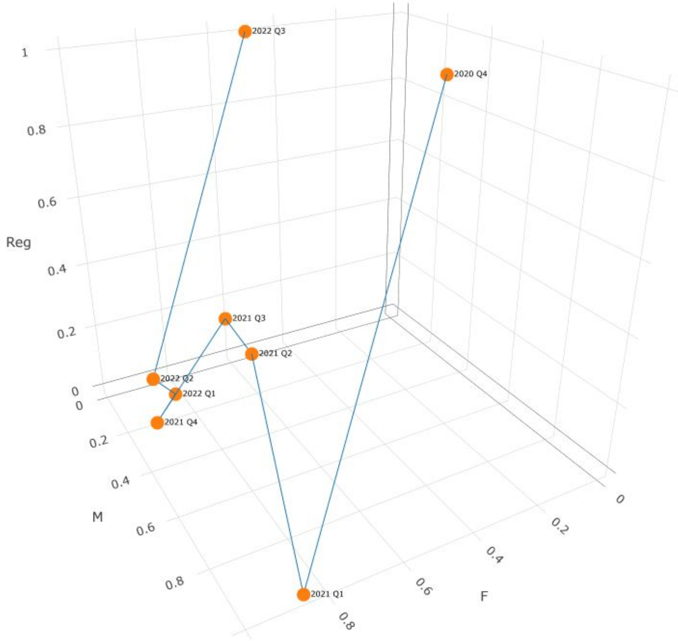


4th quarter, 1st quarter, 2nd quarter, 3rd quarter, 4th quarter, 1st quarter, 2nd quarter, **3rd quarter**

Oct 2020, Jan 2021, Apr 2021, Jul 2021, Oct 2021, Jan 2022, Apr 2022, **Jul 2022**, **Oct 2022**

In order to catch and analyse customers' dynamic patterns, we have performed a longitudinal RegFM on the 8 **full quarters** (disregarding the last month in the data set, namely October 2022) solely on the customers (outliers have been removed) present in the 3rd quarter 2022, thus remaining consistent with our long-term objective. To define the threshold of the longitudinal RegFM on each quarter, we have decided to use a **mixed approach**. Starting from the 1st quartile, median and 3rd quartile of the last quarter and by considering that the latest distribution in time is the most valuable one, we have adjusted the thresholds with a coefficient ("g") which represents the thresholds' average shift.
Once we got the clusters, we have been looking at:

Infra-cluster migration analysis: focusing on the evolution in time of the average of the RegFM values, it emerged that: concerning the *Primary shopper* the monetary value decreases steadily, while frequency and *Regularity* face a drastic worsening in the last quarter. For what concerns the *Secondary shopper* we noticed that, while the monetary and frequency values decrease steadily, the regularity achieves a more variable behavior with respect to the *Primary ones*. Finally, the *Non shopper Monetary* value shows a similar trend with respect to the previous clusters except being more stable (smaller variability). On the other hand, *Regularity* and *Frequency* seem highly unstable.
Moreover, we calculated a proxy of the cluster (i.e. *Monetary, Frequency* and *Regularity*) stability as well as the distance between the centroid of the same cluster in different quarters (e.g.: see picture on the right). Unfortunately, this value has not provided any particular insights apart from being correlated with the sample size.



Evolution in time of the Primary shopper's RegFM centroids

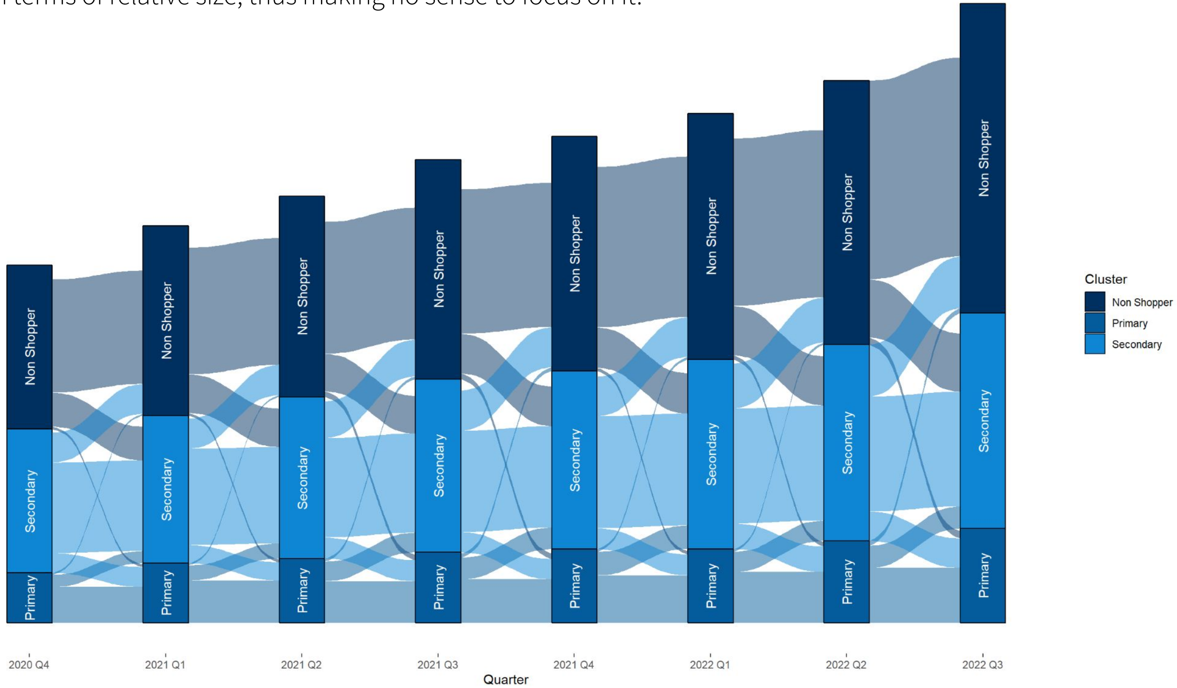
Longitudinal RegFM

Inter-cluster migration analysis: looking at the alluvial graph, it is possible to assess the 3 different classes of customers' flows during the 8 periods. In general, the in- & out-flows seem to be well-balanced with a **great share of stable** customers. An interesting pattern regards the one from *Non-shopper* to *Primary*. However, it does not seem to be very significant in terms of relative size, thus making no sense to focus on it.

Ther reverse is to some extent true as well, although it might be interesting to understand the reasons because of which a customer switched from being a *Primary* one to a *Non shopper*.

Instead, the most interesting and relevant (for our purposes and in terms of size) patterns are those concerning the switches from *Secondary* to *Primary* and conversely, and from *Primary* to *Primary*, thus regarding those customers that decided to remain within the same cluster.

Furthermore, the migration from *Non shopper* to *Secondary* highlights that in the lowest valuable group there might be a portion of *Potentially good customers* that might deserve more attention, i.e. customized marketing actions in order to harvest possible unleashed potentials.



Migration patterns

Marketing implications

The clusters analysis suggests which CRM actions might be possible to adopt: the main idea consists of creating a **tier program** with two separated tiers, dividing *Primary* customers (**gold tier**) and *Secondary* customers (**silver tier**). To better understand which policies could have been beneficial we conducted a **Market Basket Analysis** on these clusters (See. MBA). The more practical concepts have been summarized below.

For the **gold** tier the aim would be to increase as much as possible the **retention**. This could be achieved in various ways. The underlying idea would be to **increase** the tier's **switching costs**. Some possible solutions could be:

- Ad-hoc loyalty programs connected to specific benefits (say, "rewards' goods") which will be obtainable though by collecting a higher number of points;
- Unique discounts and exclusive points reward for that tier;
- Possibility to pay in different ways or pay-back policies and;
- In case of the existence of a home delivery program, possible exclusive services like free delivery or fast delivery.

For the **silver** tier we noticed a general stability of the cluster but with a still relevant flow to the next, more valuable, tier. So, the CMR actions to adopt should be aimed at **increasing the *Frequency*** and/or the ***Monetary*** purchases' value.

The best choice would be to have an improved loyalty program (for example using the budget allocated for the actual loyalty program to non-shoppers) and trying to promote more frequent purchases of complementary goods. Moreover, a further interesting action that could be performed in order to increment our target's frequency along the aforementioned one, concerns of policies such as targeted discounts (more generally direct marketing actions), that nudges our tier to visit the shops specifically in those days in which we have found they tend not to (or rarely) do groceries (see. MBA).

Market Basket Analysis

To have a broader view of our customer base, we approached the market basket analysis in two different ways: we used an **“apriori”** algorithm on virtual variables (i.e. dummy items) such as *week-day*, *hour* and *total expense* to determine if frequent purchasing pattern emerged regarding specific cluster of customers. Secondly, we applied the **“eclat”** algorithm to perform a **WARM (Weighted Association Rules Mining)**: the frequency of an itemset may not be a sufficient indicator of interestingness because it does not reveal the utility of an itemset, which can be measured in terms of cost, profit, or other expressions of user preference.

If frequencies of items vary, two problems encountered are:

1. At high *minsup* value, then rules or patterns of rare items will not be found, as rare items fail to satisfy the *minsup* value.
2. To find rules that involve both frequent and rare items, *minsup* has to be set very low. This may cause combinatorial explosion, that is too many rules or frequent pattern are generated.

Before proceeding with the analysis we need to make a clear assumption, useful for computing item-sets’ weights .

Assumption: Equal profit margin for every product

With equal profit margin, we can push more items with higher price that would not emerge if we considered only frequency. *Eclat* algorithm allows us to compute the support of each item A in the following way:

Weighted support (A) = $\sum w_i T / \sum w_i$

Where:

Weight (A) = Unitary price (A), scaled on a range from 0 to 1
T = All those transactions containing A

Association Rules Mining (ARM)

Best practices suggest to choose an aggregation level such that the support for single items is almost the same for all the items in the transactions set. From the plot on the right, we can see that apart from the first five items, this rule is somehow respected.

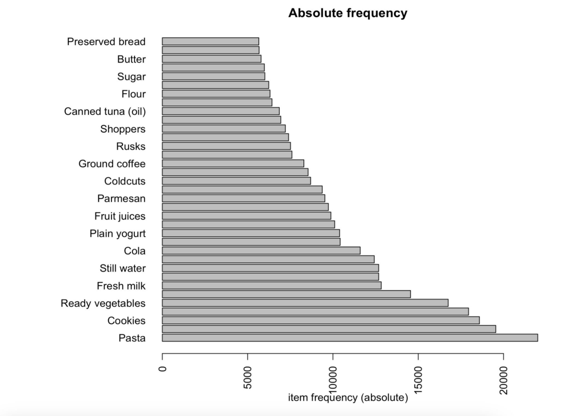
Both explorations on purchase pattern and consumer's basket have been conducted on the clusters deriving from the RegFM analysis, namely "Primary" and "Secondary" customers.

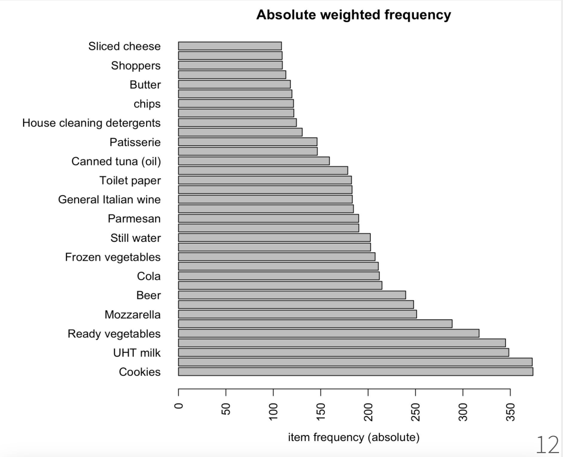
Since the supermarket chain is already set up, it would be interesting to collect customers' data in order to perform further analyses with more demographic data and with added information regarding promotions on particular products.

	Basic MBA	MBA with Virtual variables
Confidence	0.001	0.001
Support	0.3	0.3
Max Length	4	8

We started the rules research by setting up lower-bound parameters of support and confidence, and upper-bound for the length of the rule.

Note that we did not set any rule's length that considered more than 4 items since facing the risk of becoming too specific, thus loosing in terms of generality, i.e. actionability. .





ARM and WARM on *Primary* customers

Since the goal for this cluster consists mainly in **retention**, we are interested in understanding customer purchasing patterns and preferences in other to “accommodate” them.

Through a purposive MBA, we imposed RHS to be equal to “Medium-High” and “High” ticket value, because we wanted to understand what purchasing patterns on average triggers higher ticket values.

The results suggest **high ticket value in correspondence to the weekend, from Friday afternoon, until Saturday evening,** while regarding items, we highlighted particular rules that showcase sparse preferences that vary from recipes to cat-related products.

Finally it is important to underline that ticket values are derived from the quartile distribution of the tickets and for that reason depends strongly on the cluster chosen for the analysis: for example 80% of the “Primary” tickets are below 15 €, while if we consider “Secondary” ticket, it raise up to 20€. This is due to a **tradeoff between frequency and monetary**, where best customers make purchases more frequently with slightly lower average price.

Rules (apriori algorithm)	**Support**	**Confidence**	**Coverage**	**Lift**
{Saturday,Evening} => {High ticket value}	0.00693	0.236	0.0293	1.18
{Saturday, Afternoon} => {High ticket value}	0.00906	0.25	0.0362	1.25
{Friday, Afternoon} => {High ticket value}	0.00762	0.236	0.0324	1.18

Rules (eclat algorithm)	**Support**	**Confidence**	**Coverage**	**Lift**
{Bakery ingredients, Sugar} => {Flour}	0.00122	0.343	-	8.75
{Canned tuna (oil), Tomato sauce} => {Pasta}	0.00126	0.437	-	3.2
{Cat accessories} => {Cat food}	0.00126	0.426	-	9.87

ARM and WARM on *Secondary* customers

Our goal for the *Secondary* cluster is to nudge customers migrate towards the *Primary*, which in concrete terms means increasing monetary spending, frequency of purchase or both.

That is why we want to incentivize the purchase of complementary products through offers of frequently demanded items, but also to understand which are the days when customers on average are not used to make any purchases.

In this case, for the apriori algorithm we set LHS equal to every day of the week and looked for **weaker rules** in terms of support, but still relevant in terms of confidence.

For this particular cluster, it emerged that Monday, Tuesday and Thursday are days with lower frequencies if compared to Friday, Wednesday and the weekend. Thus, it might be of interest to push in-store promotions selecting these precise days.

If instead we look at products most frequently purchased together, we can see that there's a strong interest relating to **"bakery products"**, being both profitable for the retailer and demanded by the customer. This is still valid for **"house cleaning products"**

Rules (apriori algorithm)	**Support**	**Confidence**	**Coverage**	**Lift**
{Thursday, Early morning} => {low ticket value}	0.0106	0.24	0.044	1.2
{Monday, Early morning} => {low ticket value}	0.0094	0.241	0.039	1.2
{Wednesday, Afternoon} => {low ticket value}	0.00785	0.238	0.033	1.19

Rules (eclat algorithm)	**Support**	**Confidence**	**Coverage**	**Lift**
{Bakery ingredients, Pasta, UHT milk} => {Flour}	0.00119	0.42	-	7.97
{Dishwashing soap, Laundry soap, Paper and disposable tools} => {House cleaning detergents}	0.0011	0.352	-	7.44
{Canned tomato pulp, Sugar} => {Pasta}	0.00119	0.561	-	3.55

Testing association rules

Through WARM we highlighted shared rules between clusters:

Primary: {Bakery ingredients, Sugar} => {Flour}
Secondary: {Bakery ingredients, Pasta, UHT milk} => {Flour}

Common ground: {Bakery ingredients} => {Flour}

Considering the **anti-monotone** property of support, all subsets of a frequent itemset must also be frequent, which means that the common ground rule is itself frequent.

That’s why we want to propose a marketing intervention regarding bakery-related products. On one side we can “accomodate” the *Primary* cluster, by promoting products of their interest, and on the other side we can incentivize the *Secondary* cluster to purchase these products in low frequency days.

“Sweet Monday” Campaign

Campaign content:
On Monday bakery ingredients can be found at 30% discount showcased on a dedicated store isle. Purchasing these products in that particular time-frame allows customer to collect more points in their loyalty card.

Campaign distribution:
The campaign is promoted only on owned media (Instagram, newsletter) to minimize costs.

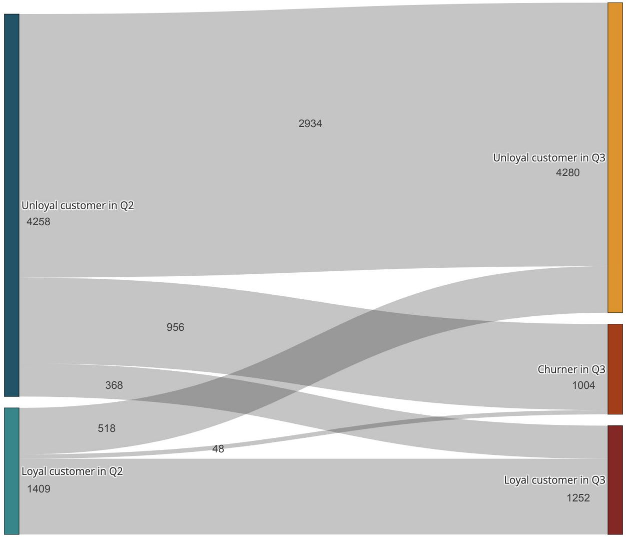
Campaign KPIs:
Bakery product sales on Monday/Bakery product sales on rest of the week

Timeline:
Q4 2022

Detect customer partial defection

The main problem with trying to detect partial defections in the retail sector is that you cannot know whether a customer has become a churner until she has really become one. So we used a different perspective, described by W. Buckinnx and D. Van den Poel in their article. In doing so, we had to detach slightly from the work done so far, in order to remain consistent with the suggestions provided by the two aforementioned researchers. Indeed, in this particular case, it has been reasonable to avoid to exploit the third dimension represented by *Monetary* since considered as a double-edged sword: there might the possibility to exclude (i.e. considered as churned) some customers just because purchasing low sums, despite having a quite regular and frequent behaviour, thus representing potentially good customers. Anyhow, how do we know if a customer is behaving like a loyal customer? Two conditions help us identify them: (1) the frequency of purchases is greater than the quarter average and (2) regularity (expressed by the traditional formula) is below the quarter average. If both conditions are met then we can say that a customer is behaving as a loyal customer in the relevant quarter. Before we continue, we need to make a further premise of the use of historical data in classification algorithms: the classification model we want to find is not a typical classification algorithm, but a it is one that uses a stream of data (albeit of low frequency). This means that since we have identified the quarter as the timeframe of interest, we will use **Q2 of 2022** and **Q3 of 2022** to perform our analysis concerning whether a customer behavior is changing or not.

Using the data from the first quarter we will find the customers who behaved like loyal customers in Q2 of 2022, and the data from Q3 of 2022 to see whether the customers have changed their behavior or not. Next, we will focus only on customers who behaved as loyal customers in the first quarter, since it is now well known that regular (or even potentially regular since we have excluded the monetary component to define a loyal customer) customers are those who guarantee the major companies and have changed their behavior in the second quarter. We will not consider loyal customers who have churned because with our model we are interested in **predicting a churn before it happens** and therefore we want to see what the reasons that drive a customer to change purchase behavior are. From the graph on the right, we can see how the customer behavior has evolved in different quarters. Moreover, we can also notice that the rule used to detect loyal customers is working fine, since just 48 loyal customers have churned. We selected three algorithms to classify customers who changed behavior and we chose them because they are suitable to support decision-making process and their results are easily interpretable: binary tree classification, logistic regression, and random forest. For the complete dataset used, you can refer to the annexes where we have provided another basic idea that came up to our minds but which we did not implement it.



Customer Behaviour in Q2 and Q3

Results of classification

Disclaimer: it is important to notice that we have just a part of the complete dataset, so we need to keep in mind the fact that if we had more data, the classifications algorithm could have performed even better.

Initially we had also thought of including in the dataset the cluster resulting from our previous analysis and the cluster from the previous quarter. The problem that arose once we began the model selection was that all the factors with greatest importance were related to those two categorical variables. We therefore decided to remove them from the model. To assess model performances we decided to use F1 score and AUC metrics. In the table on the right you can find model performances. During the training of the model, we balanced the target variable, to avoid misleading results of classification models. You can refer to the Jupyter notebook for all the script. In the annexes you can find all detailed information about the different models we found. From the model found we gained some important insights, which we decided to summarize in three main points:

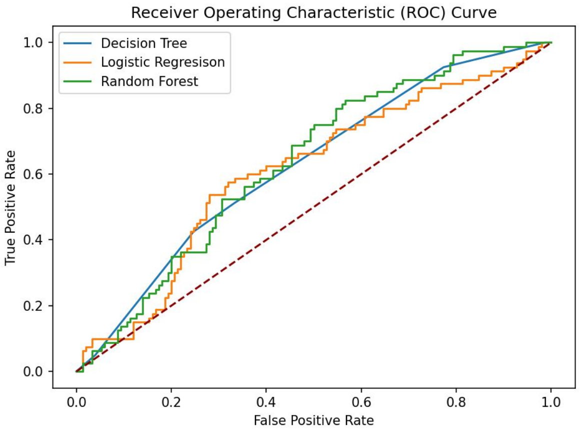
- The time component, both in terms of recency and inter-purchase time is definitely dominant over the importance of the monetary component, so we should try through marketing campaigns to increase the number of visits and as regularly as possible

- We suggest conducting further analysis of why people spent more than 38.56€ on food during the quarter. This could be because the food has a low quality and therefore customers who spend so much on food the next quarter tend to change their behavior. Moreover, we suggest also to focus on the fruits and vegetables, since results obtained with the logistic regression showed that spending on fruits and vegetables tends to "squezee" the probability toward 0. We want to emphasize the importance of this analysis since in the decision tree it is one of the highest-level and therefore most impactful nodes. The importance is also confirmed by the feature importance of the random forest model we found.

- The number of categories purchased does not seem to influence the change in customer behavior, this means that the retailer should not necessarily continue to sell so many categories and indeed, should conduct a profitability analysis and possibly, carry out marketing campaigns aimed at increasing sales in the less purchased categories.

In conclusion, considering that all models have equivalent performance, we suggest using the decision tree classifier to determine partial defections relying on the time rules obtained from the tree. In the annexes you can see the complete treeplot

Model	F1	AUC
Decision tree classifier	0.48	0.63
Logistic regression	0.522	0.61
Random forest classifier	0.51	0.64



Conclusion

Customers churn because of a misalignment between their perceptions of the relationship with the supplier and the service received. With our analysis it’s not possible to identify the specific reasons behind this behavior, but it’s clear that despite there are not many loyal customers that churn, almost ⅓ of them subside to the un-faithful categories in the subsequent period. This is a red flag for the company because the unloyal customers are the most at risk of churning. To avoid this phenomenon it is prominent to implement specific marketing actions such as Loyalty programs and price tactics (such as discounted products for the loyal group, see MBA) to retain them. A clear definition of partial defection increases the effectiveness of a churn prevention policy and allow the company to not lose the good customers, that in our case, can generate the 80% of the revenues. It is preferable to increase retention rate because the costs of acquiring new customers are increasing more and more.

References

- «Cluster evolution analysis: Identification and detection of similar clusters and migration patterns», R. Ramon-Gonen & R. Gelbard, Expert Systems With Applications 83 (2017) 363–378

- “Database Marketing Analyzing and Managing Customers”, Robert C. Blattberg, Byung-Do Kim and Scott A. Neslin, Springer

- Wouter Buckinx, Dirk Van den Poel, Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting, European Journal of Operational Research, Volume 164, Issue 1, 2005, Pages 252-268, ISSN 0377-2217

- “Customer Relationship Management, Concept, Strategy, and Tools”, V. Kumar, Werner Reinartz ISBN 978-3-662-55380-0

- “Marketing Analytics, A practical guide to improving consumer insights using data techniques”, Mike Grigsby,ISBN 978 0 7494 8216 9

- “A new framework for predicting customer behavior in terms of RFM by considering the temporal aspect based on time series techniques”, Hossein Abbasimehr,Mostafa Shabani

- “A review of the application of RFM model”, Jo-Ting Wei, Shih-Yen Lin and Hsin-Hung Wu

- “Cluster evolution analysis: Identification and detection of similar clusters and migration patterns”, Roni Ramon-Gonen, Roy Gelbard

- “ A modern approach to RFM segmentation”, Roy Wollen

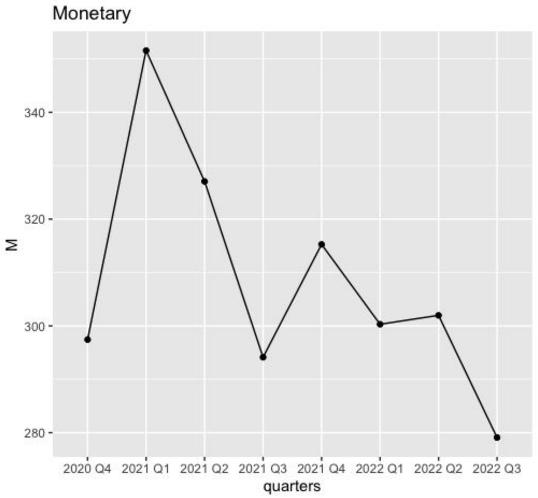
- Overcoming the “recency trap” in customer relationship management, Scott A. Neslin,Gail Ayala Taylor, Kimberly D. Grantham, Kimberly R. McNeil

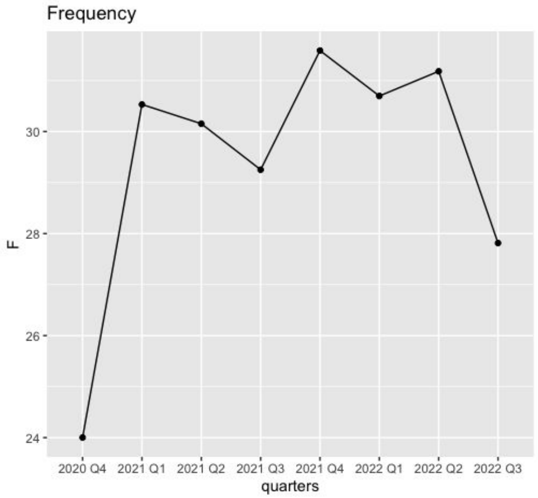
- Association Rule – Extracting Knowledge Using Market Basket Analysis, Raorane A.A., Kulkarni R.V., and Jitkar B.D., 2012

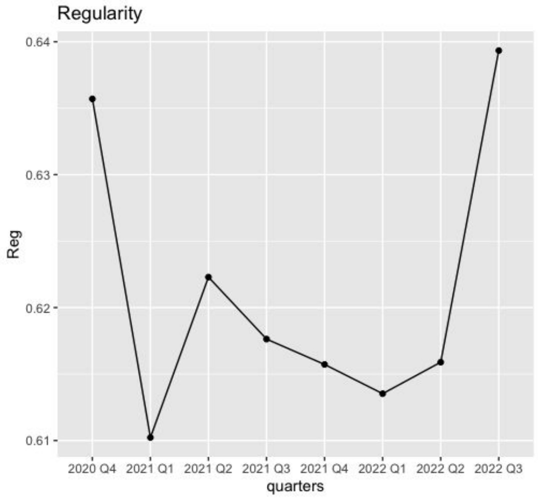
- Mining utility-oriented association rules: An efficient approach based on profit and quantity, Parvinder S. Sandhu, Dalvinder S. Dhaliwal and S. N. Panda, 2012

Annexes

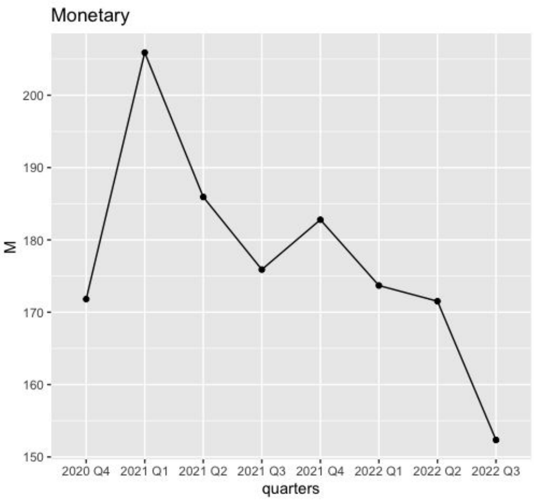
Primary shoppers analysis

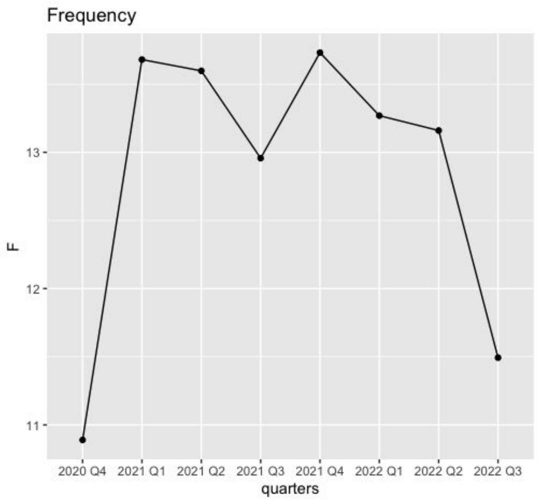


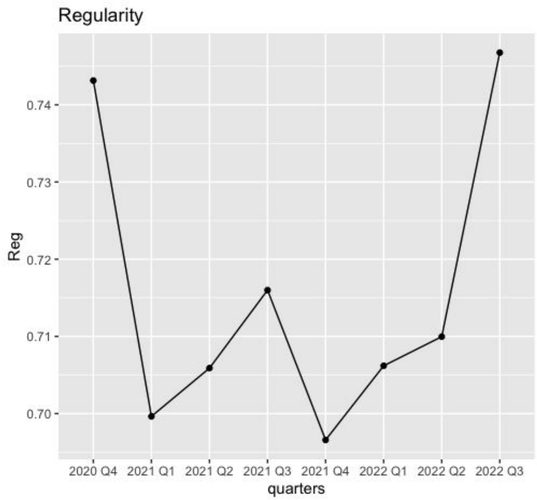




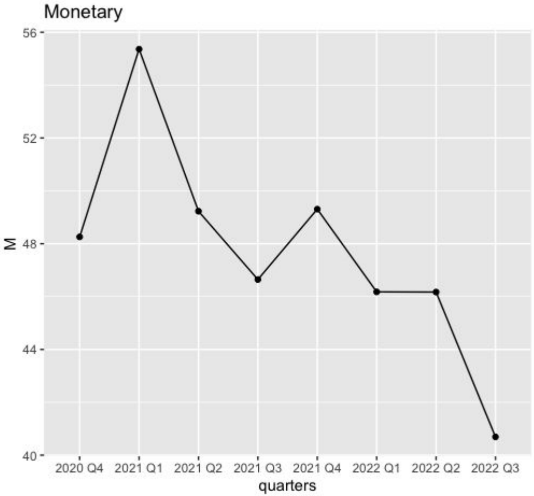
Secondary shoppers analysis

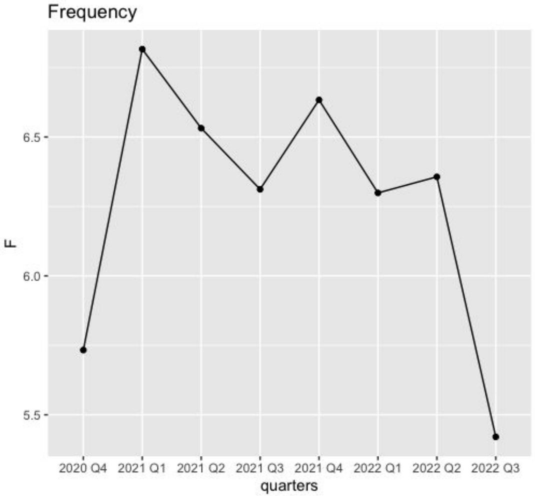


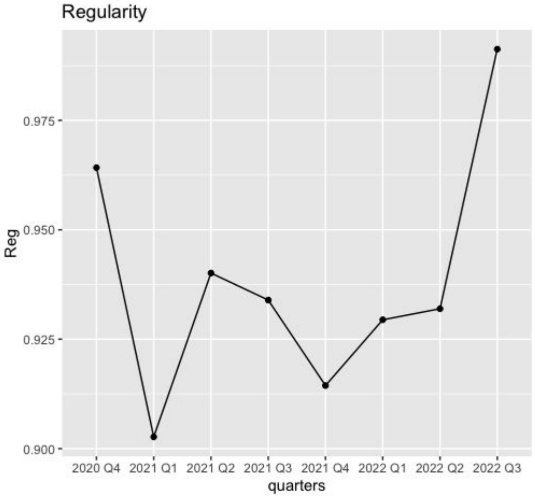




Non Shopper analysis





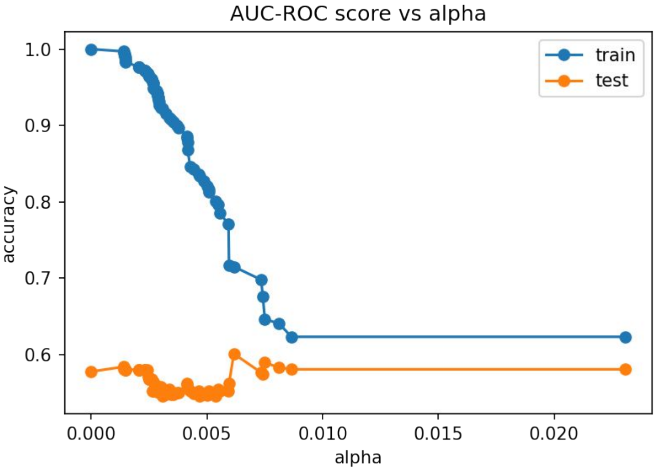


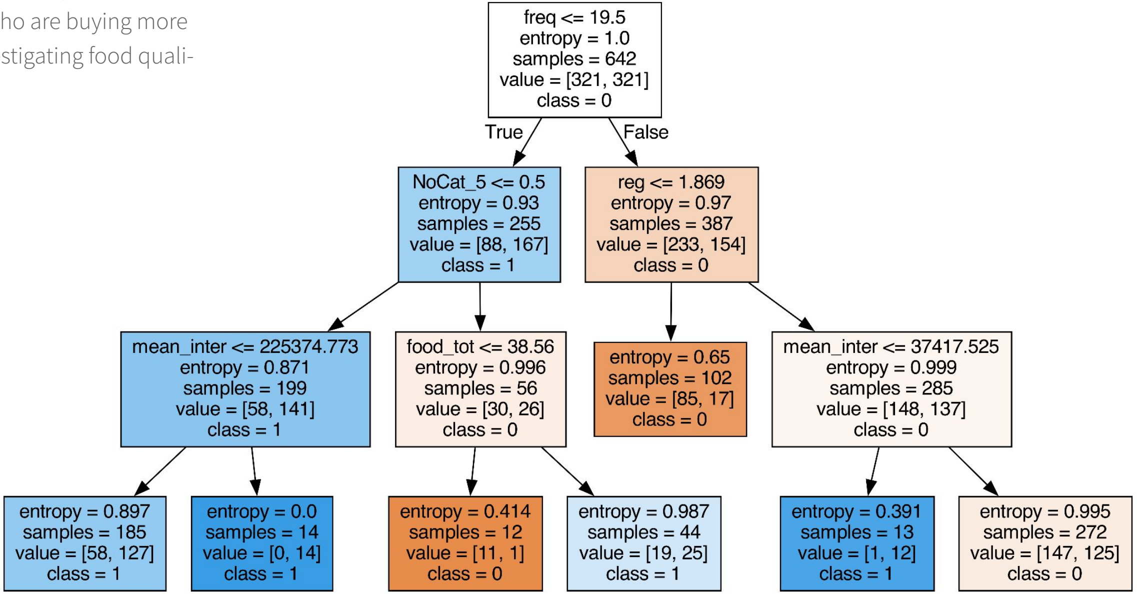
Dataset for partial defection detection

Attribute	Description
customer_id	Unique ID for customers
Reg	Standard deviation of interpurchase time divided by the mean of the interpurchase time
Freq	Frequency of customer visits in the quarter
Recency	Recency of customer
std_inter	Standard deviation of interpurchase time
mean_inter	Mean of interpurchase time
mean_timing	Mean of visit timing: this is the mean(hour(date_ticket)) for the quarter
std_timing	Standard deviation of visit timing: this is the mean(hour(date_ticket)) for the quarter
monetary	Monetary of the customer in the selected quarter
NoCat	Number of different macro-categories purchased in the quarter. We divided products in 14 different macro-categories: Food, Drinks, Dairy, Meat, Home, Personal care, Pet, Fruits and vegetables, Fish, IT, Clothes, Hobby, Motors, Other
pct_expense_above_average	Percentage of visits spending more than average spending
*_tot	Monetary expenses in each macro-category. You can replace * with category name
Target	Target column: 1 if the customer changed h* behavior or 0 if s/he remained loyal

Decision tree plot

The decision tree has such a small depth because we pruned the tree to avoid overfitting to the test set data, using the cost complexity pruning technique. In our case, we set the parameter `ccp_alpha=0.00865771`. From the plot below you can see the minimum point of AUC metrics versus alpha. From the tree plot, as we could have expected, we can see that loyal customers, whenever switching their behavior, visit the store less frequently and with an higher coefficient of variation of interpurchase time. By using this thresholds we can detect if a customer is about to change h* behaviour. Another important insight could be the one related to the total expense in food: it seems that people who are buying more food are more likely to change their behavior, thus, we suggest in investigating food quali--ty because could be a drive of change in behavior.





freq <= 19.5
entropy = 1.0
samples = 642
value = [321, 321]
class = 0

True

False

NoCat_5 <= 0.5
entropy = 0.93
samples = 255
value = [88, 167]
class = 1

reg <= 1.869
entropy = 0.97
samples = 387
value = [233, 154]
class = 0

mean_inter <= 225374.773
entropy = 0.871
samples = 199
value = [58, 141]
class = 1

food_tot <= 38.56
entropy = 0.996
samples = 56
value = [30, 26]
class = 0

entropy = 0.65
samples = 102
value = [85, 17]
class = 0

mean_inter <= 37417.525
entropy = 0.999
samples = 285
value = [148, 137]
class = 0

entropy = 0.897
samples = 185
value = [58, 127]
class = 1

entropy = 0.0
samples = 14
value = [0, 14]
class = 1

entropy = 0.414
samples = 12
value = [11, 1]
class = 0

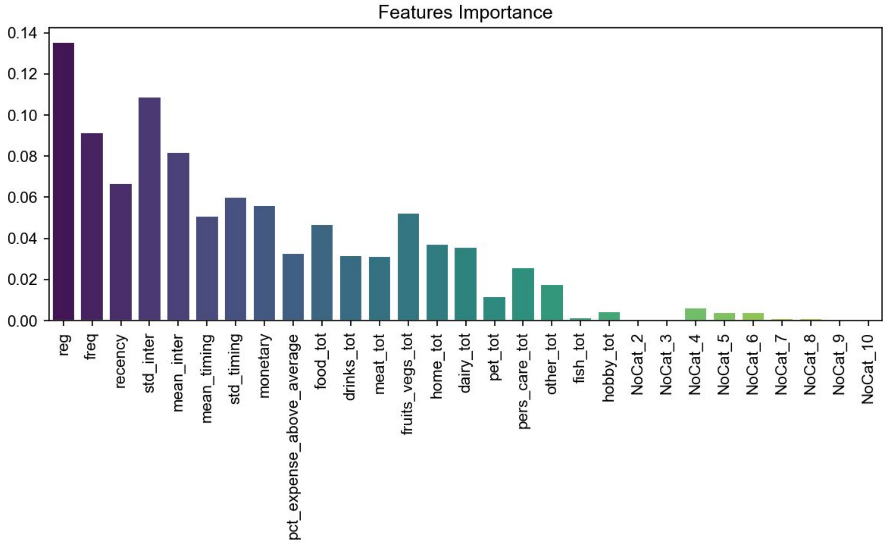
entropy = 0.987
samples = 44
value = [19, 25]
class = 1

entropy = 0.391
samples = 13
value = [1, 12]
class = 1

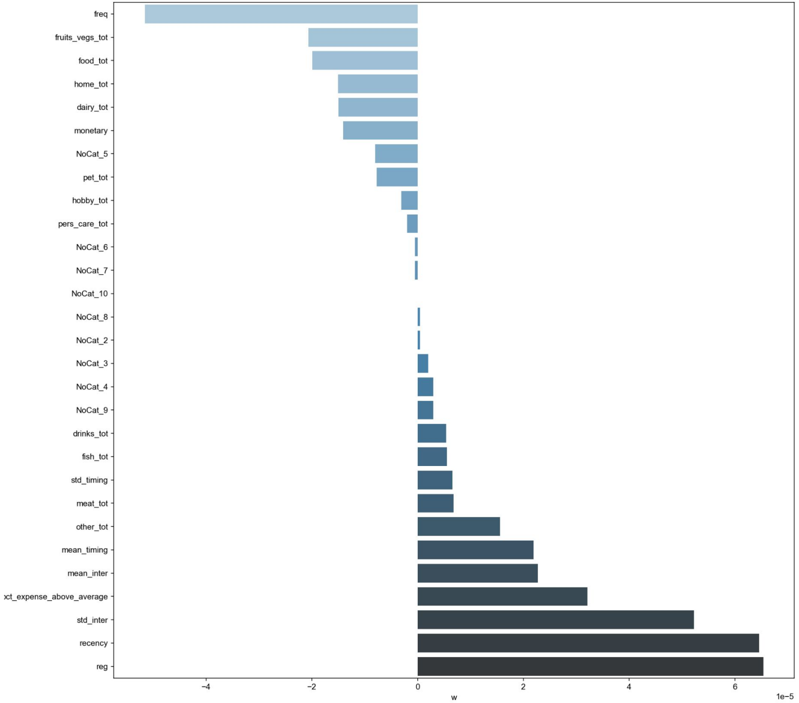
entropy = 0.995
samples = 272
value = [147, 125]
class = 0

Random forest features importance

From the image on the left, we can see the *feature importance* for all the variables included in the model. We can see that the numbers of categories purchased in the quarter have little importance. Thus, it could make sense to drop them from the dataset and retrain the model for detecting partial defecting customer, i.e. decision tree classifier. On the other hand, we can see once again the importance of regularity, frequency and recency, united with the relevance of the interpurchase time's standard deviation and the mean. On the other hand, monetary seems less important with respect to the time. So, we have to keep in mind that monetary component is less important then the temporal one, or in other words, it is not necessarily true that a customer who spends a lot will never change his behavior: the time component is more important in making sure that a customer does not change h* behavior. In a certain way, the best way to make a customer not change behavior is through changing his shopping habits.



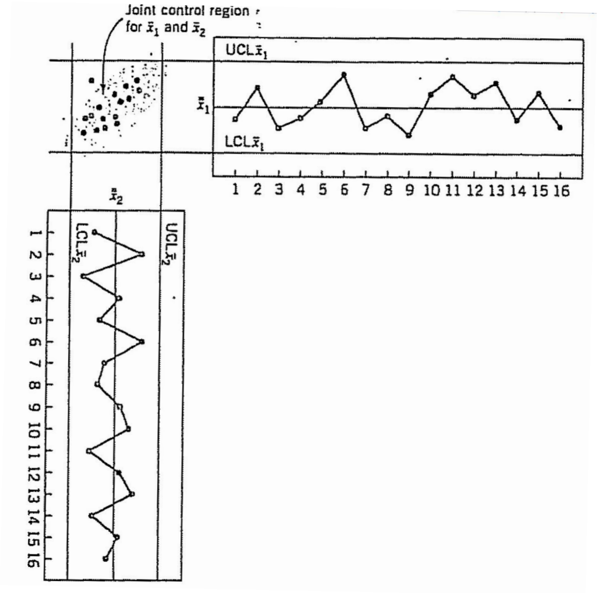
Logistic regression weights



Note that we used standardized numerical variable
to fit the logistic regression model

A control chart approach for detecting pattern changes in custumers behavior

We want to propose also another method to detect abnormal patterns in the behavior of the costumer. This approach, unlike the one shown earlier in the slideshow, does not focus only on loyal customers, but consists of a "tailor made" tool for each customer. This way, we could avoid the biases that might be caused by aggregating available data. The idea is to use a **multivariate process control chart**. Control charts are statistical tools that allow us to define whether a process is under control. Our idea is to develop a control chart to simultaneously monitor monetary and interpurchase time. Being single observations we will have to use an I-MR control chart if we are interested in monitoring variability as well or an exponential weighted moving average control paper if we are interested in detecting even small shifts from process mean. In this way we could know at any instant whether the client's behavior is deviating from normal, either positively or negatively. For example, if the control chart constructed on monetary shows a pattern of above-average spending subsequent to a marketing campaign, it means that the marketing campaign was effective for that customer. In addition, by monitoring also interpurchase time, we can easily detect anomalies patterns of store visits. In fact, it is very difficult for a customer to churn all at once, especially in the retail sector. By doing so, we will be able to target the individual customer with offers aimed at keeping him or her active again. Finally, since control charts are statistical tools, we can calculate the expected value of losses in case the pattern of anomalies detected is a false alarm. By doing so, we could more effectively organize marketing campaigns by **ensuring that the ROI of the campaign is positive**.


Joint control region for $\bar{\bar{x}}_1$ and $\bar{\bar{x}}_2$
UCL$\bar{\bar{x}}_1$
$\bar{\bar{x}}_1$
LCL$\bar{\bar{x}}_1$
$\bar{\bar{x}}_2$
LCL$\bar{\bar{x}}_2$
UCL$\bar{\bar{x}}_2$