Università degli studi di Milano

Msc: Data Scicence for Economics

Statistical Learning Module

iFood Data Analyst Case

Pietro Padovese
Matriculation Number: 12356

Abstract

This report contains an analysis of a dataset made available by iFood, the lead food delivery app in Brazil. The company made public such data regarding its customers, their purchasing habits and a promotional campaigns proposed to them, with the objective of evaluating different approaches to extract business insights.
The main goals of this study are to investigate the presence of clusters among customers, build a predictive model that can determine a customer's response to a new promotional campaign and determining the most important variables that drive this choice.
Following the data precopressing phase, various supervised and unsupervised learning techniques are applied in order to achieve the expected goals while evaluating different alternatives and analysing their performance.

Contents

1 Introduction

1.1 Dataset Description

The company iFood operates in the retail food sector. Their products belong to five main categories: wines, rare meat products, exotic fruits, fish and sweet products. These products can also be divided into gold or regular products. The company reaches its customers through three sales channels: physical stores, catalogues and the company's website. Recently, they have launched a pilot campaign involving 2240 customers for whom they recorded the outcome of their proposal. The analysed dataset contains, in addition to the outcome of the campaign, socio-demographic information about the customers, their product and sales channel preferences, and their adherence or non-adherence to five previous campaigns.

Name	**Description**
AcceptedCmp1	1 if costumer accepted the offer in the 1^{st} campaign, 0 otherwise
AcceptedCmp2	1 if costumer accepted the offer in the 2^{nd} campaign, 0 otherwise
AcceptedCmp3	1 if costumer accepted the offer in the 3^{rd} campaign, 0 otherwise
AcceptedCmp4	1 if costumer accepted the offer in the 4^{th} campaign, 0 otherwise
AcceptedCmp5	1 if costumer accepted the offer in the 5^{th} campaign, 0 otherwise
Complain	1 if costumer complained in the last 2 years
DtCustomer	date of customer's enrollment with the company
Education	customer's level of education *(categorical)*
Marital	customer's marital status *(categorical)*
Kidhome	number of small children in customer's household
Teenhome	number of teenagers in customer's household
Income	customer's yearly household income
MntFishProducts	amount spent on fish products in the last 2 years
MntMeatProducts	amount spent on meat products in the last 2 years
MntFruits	amount spent on fruits in the last 2 years
MntSweetProducts	amount spent on sweet products in the last 2 years
MntWines	amount spent on wines in the last 2 years
MntGoldProds	amount spent on gold products in the last 2 years
NumDealsPurchases	number of purchases made with discount
NumCatalogPurchases	number of purchases made using catalogue
NumStorePurchases	number of purchases made directly in stores
NumWebPurchases	number of purchases made trough company's web site
NumWebVisitsMonth	number of visits to company's web site in the last month
Recency	number of days sinche the last purchase
Response	1 if costumer accepted the offer in the last campaign, 0 otherwise *(target)*

The dataset contains a total of 24 explanatory variables plus the response. Of these 24, 6 are binary variables, 2 are categorical variables, 1 represents a date and 15 are numeric variables.

1.2 Outline

The data pre-processing and EDA phases are given in section 2, where we discuss among other things the encoding of variables, how to deal with outliers and null values, variable transformations, how classes are composed, and the correlation between variables. Section 3 shows the results of hierarchical clustering for the identification of clusters among clients. The construction and comparison of predictive models and the identification of the most important variables are dealt with in section 4.

2 Data Preparation and Visualization

2.1 Data Pre-processing

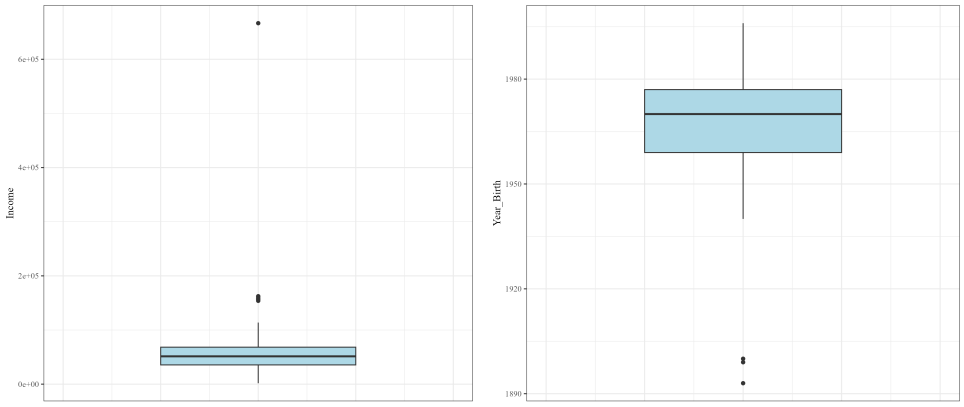
Before proceeding with the analysis, it is necessary to ensure that the data collected are complete and in a suitable format.

Within the dataset, the only variable with missing values is the income variable, which is missing in 24 of the observations. In these cases, instead of removing the observation completely, the median value of Income was assigned.

The search for potential outliers and/or inconsistent values revealed two situations that required more attention:

- For the variable "Year Birht", three observations reported values below 1900. Given the high improbability that these were correct values, they were removed.
- For the variable "Income", one observation was exceptionally large compared to the others, and to prevent it from excessively distorting the analyses, it was removed.

Figure 2.1: Visualization of outliers



(a) Boxplot for Income (b) Boxplot for Year of Birth

The next step was encoding the categorical variables through one-hot encoding.

For the variable 'Education,' which contains the levels 'Basic,' 'Graduation,' 'Master,' '2n Cycle,' and 'PhD,' the 'Master' and '2n Cycle' levels were first combined, as they

represent the same educational level in different schooling systems. 'Graduation' was chosen as the reference category (for which no dummy variable was created) because it is the most common level. 'Basic' was excluded from this choice due to having too few observations (54).

For the 'Marital Status' variable, observations with values 'Absurd' (2), 'Alone' (3), and 'YOLO' (2) were categorized under the 'Single' level, which was chosen as the reference category. Four dummy variables were then created to represent the levels 'Divorced,' 'Married,' 'Widow,' and 'Together'.

As the final step in the data processing, several variables were transformed:

- The variable 'Year Birth' was transformed into the variable 'Age.'
- The 'DtCustomer' variable, representing the date when the customer enrolled with the company, was modified to 'enrollment days,' representing the number of days since enrollment.
- Finally, the dummy variables representing responses to the previous 5 campaigns were replaced by their sum. This decision was made because, lacking information on the previous campaigns, it was considered more relevant for the analysis to have a metric representing the overall tendency of the consumer to accept a proposal.

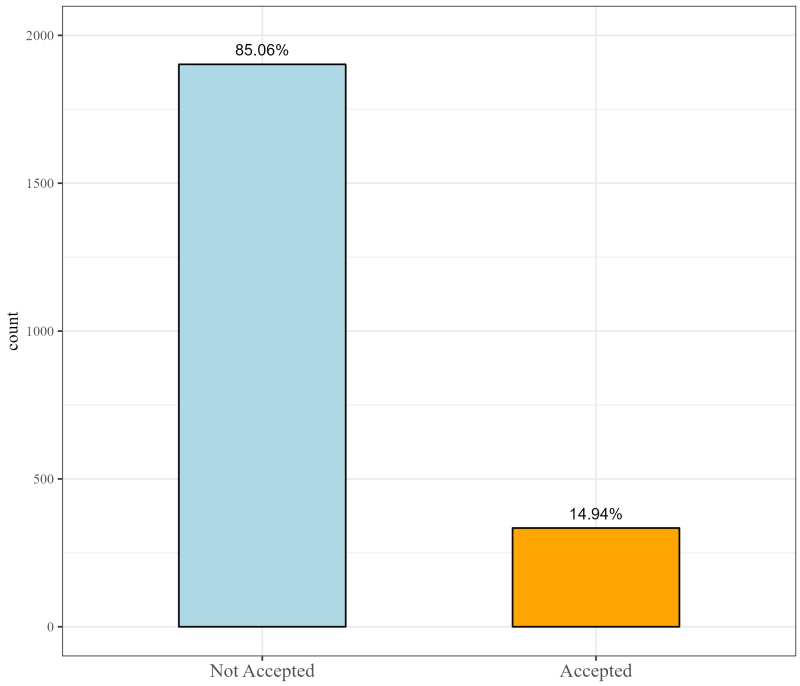
2.2 Exploratory Data Analasys

Since one of the objectives of this paper is to study the response provided by consumers to the pilot campaign offered to them, it is useful to visualize how this response is distributed in the sample.

It is evident from Figure 2.2 that a large majority of the surveyed consumers rejected the campaign offered to them. This not only tells us that there is ample room for improvement but is also a factor that will be taken into consideration when predictive models for this response are constructed in Section 4.

Given the amount of variables available, the visualization options for extracting insights about consumers are many; only a few will be shown here to give an idea.

Figure 2.2: Acceptance Rate

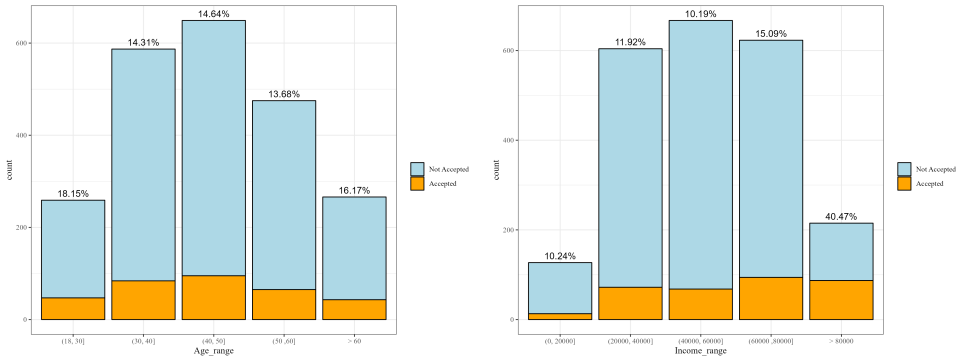


Continuing with the desire to explore the distribution of response to the campaign, one possibility is to visualize it with respect to different consumer groups, listed by age and income, thus also observing how the population is distributed with respect to these variables.

From the barplot in Figure 2.3.a, we can tell that the sample is composed mostly of people between 40 and 50. The graph is almost symmetrical to these central values, with a small prevalence of younger people. Regarding the percentage of acceptance of the campaign all classes are in line with the general level near 15 % with a slightly higher value in the younger category.

Figure 2.3.b follows a distribution similar to the one just described, with most observations concentrated on values between $20000 and $80000. The remainder falls with greater prevalence in the range above this, which also reports a significantly higher acceptance rate (40 %) than all others , a fact that assigns some significance to the possible correlation between income level and campaign outcome.

Figure 2.3: Distribution for Age and Income



(a) Age (b) Income

Other interesting insights may emerge from the analysis of the correlation between the variables. Figure 2.4 shows the correlation matrix that includes only correlations with an absolute value greater than 0.5. The Income variable appears to be the one correlated with the most variables. Not surprisingly, income is correlated with more money spent in each category of product category. Interestingly, it also seems that higher income leads to make more purchases in the store and via catalog, while at the same time to use less the website.

Figure 2.4: Correlation between the variables

These many correlations that Income need attention because including this variable in the construction of certain models (ex. logistic) could lead to multicollinearity problems. Inspecting the Variance Inflation Factor, it takes a value of 5.44, which confirms the suspected multicolinearity.

Since this indicates that the explanatory variable Income, transmits information that is "repeated" by the variables of money spent in each product category, it was chosen to transform the latter into the percentage spent in a specific category relative to the overall amount spent. As the five product variables would become a perfect linear combination, it is necessary to remove one of them, and the choice in this case fell on the MeatProducts category. In addition to the benefit of removing multicollinearity problems, in this way observations with similar values represent consumers with common habits regardless of their spending capacity.
Given the above benefits, this transformation was also applied to the variables containing money spent through each sales channel. In this case, the variable removed was that representing the physical stores.

3 Unsupervised Learning

3.1 Hierachical clustering

In this section, unsupervised learning algorithms will be used to perform clustering analysis on the dataset.

In the case under consideration, the goal of clustering is to identify groups of customers with similar characteristics regardless of their final response on the campaign. An information that is useful for the company to communicate in a more targeted way with its consumers.

Considering the composition of the dataset, which contains both numerical and categorical variables, the choice of clustering algorithm fell on hierarchical clustering, using gower distance as the dissimilarity metric among observations. The gower distance treats numerical and categorical variables differently. For the former it calculates the absolute difference between two points, while for the latter it calculates the proportion of equal categories between the two points; the more categories in common, the closer the observations will be

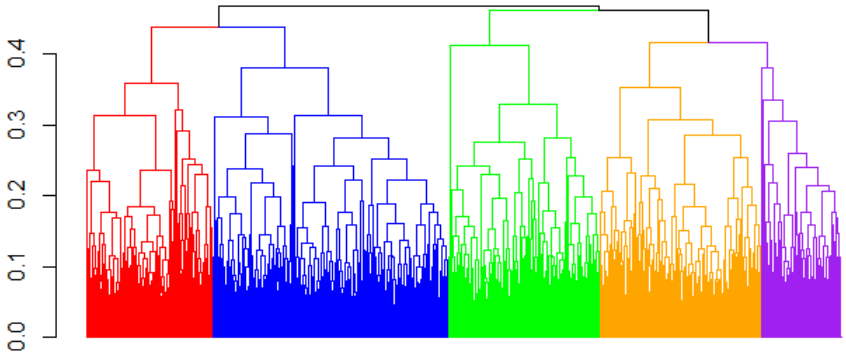
3.1.1 Variable transformation

With the aim of reducing the number of variables so that the algorithm could base its choices only on features that could bring more information to the company, modifications were applied to the variables described in the previous section.

For what concerns marital status, the possible options have been reduced to two: Toghether or Single. In addition, education levels have been merged into a single dummy that takes value 1 if the education level is equal to or higher than a master level, and 0 if lower.

Using the entire dataset initially the algorithm when asked to create a number of clusters greater than 2, created one cluster containing only 4 observations. Once their values were observed, these were classified as outliers, and then removed from the dataset.
After repeating the hierarchical clustering operation, looking at the final dendogram, and studying the different possibilities, it was chosen to divide the consumers into 5 different clusters, each with different characteristics, which will now be examined.

Figure 3.1: Dendogram

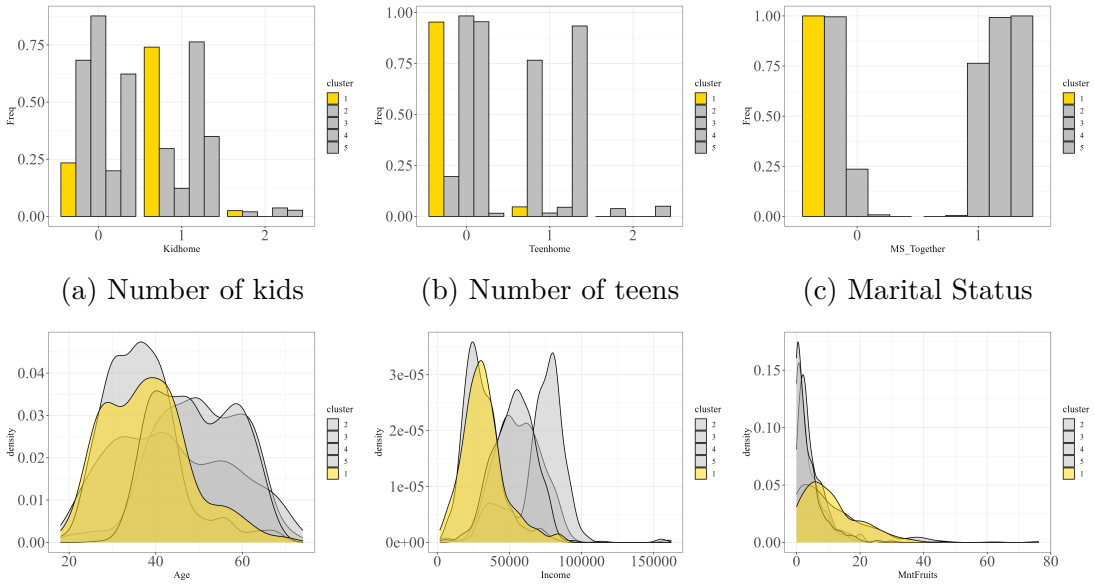


3.2 Cluster Analysis

Cluster 1
The first cluster is overwhelmingly composed of people who are not married or in couples. They have a high probability of having one or more young children, and few consumers in this cluster have teenagers. Their income is in the middle/low range. On Average they are younger than the rest of the population. Compared to the rest of consumers they buy more fruits.

Cluster 2
As cluster 1, this cluster is composed of people not in a relationship. In contrast they have a high probability of having teenage children, tend to be older, and their income is average. They also have a preference for wine.

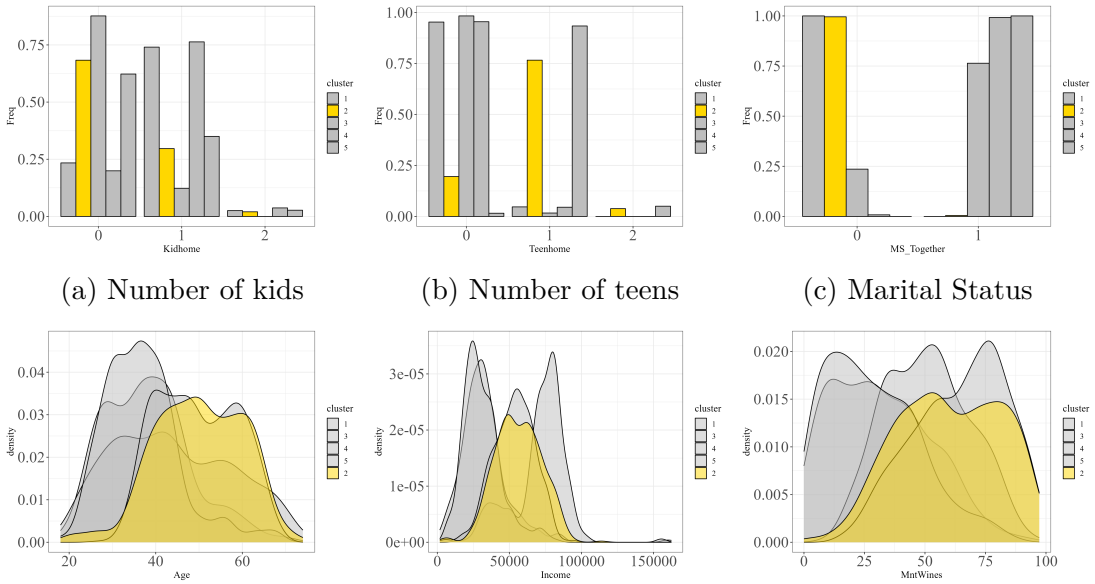
Figure 3.2: Cluster 1



(a) Number of kids (b) Number of teens (c) Marital Status

(d) Age (e) Income (f) Fruits Category

Figure 3.3: Cluster 2



(a) Number of kids (b) Number of teens (c) Marital Status

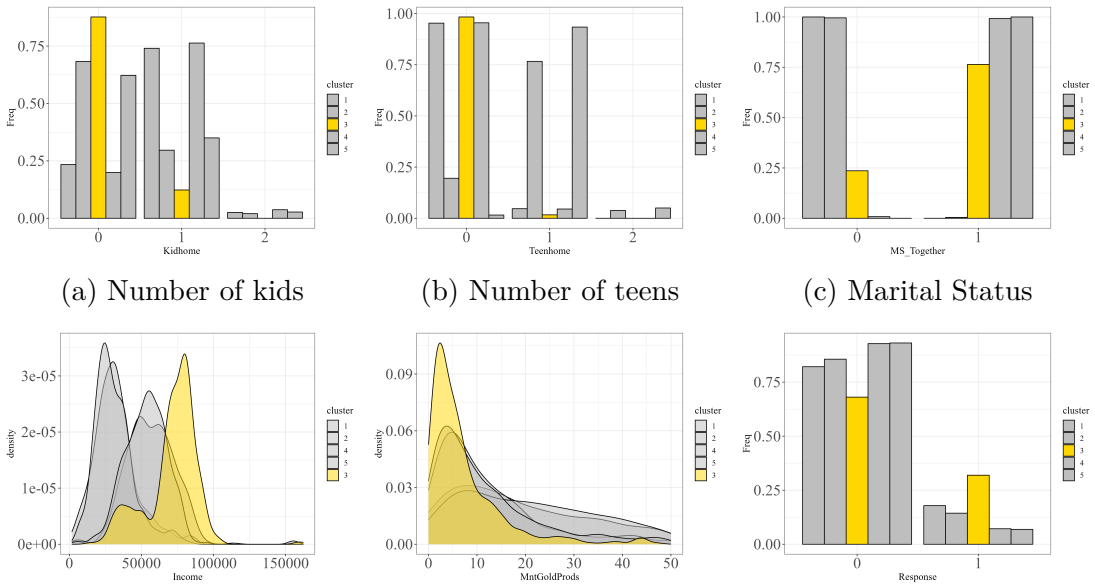
(d) Age (e) Income (f) Wine Category

Cluster 3
This cluster is the first to be composed mostly of couples (although a 25% is made of single people). What they have in common is the low likelihood of having children (both kids and teens) at home. They tend to have a high income, but despite this they buy fewer gold products than the average. Besides that, it is the group that has accepted the promotional campaign most frequently.

Cluster 4
Cluster 4 consists mainly of couples with young children. They are younger and have low incomes. They have preferences for sweet products and fruits, as well as premium products.

Cluster 5
This cluster also consists of couples, who tend to be older and have older children; their income is average. They consume less meat but much more wine.

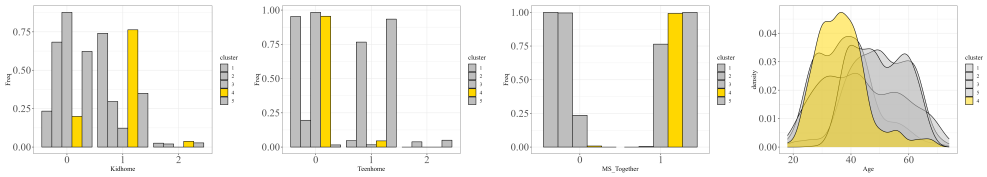
Figure 3.4: Cluster 3



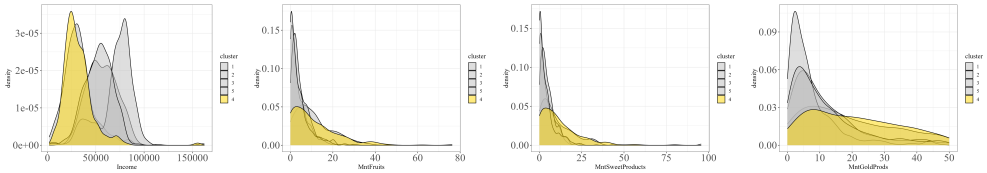
(a) Number of kids (b) Number of teens (c) Marital Status

(d) Income (e) Gold Produces (f) Acceptance Rate

Figure 3.5: Cluster 4

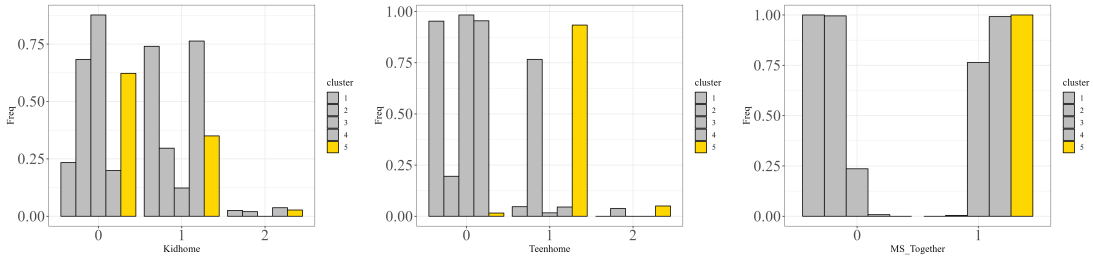


(a) Number of kids (b) Number of teens (c) Marital Status (d) Age

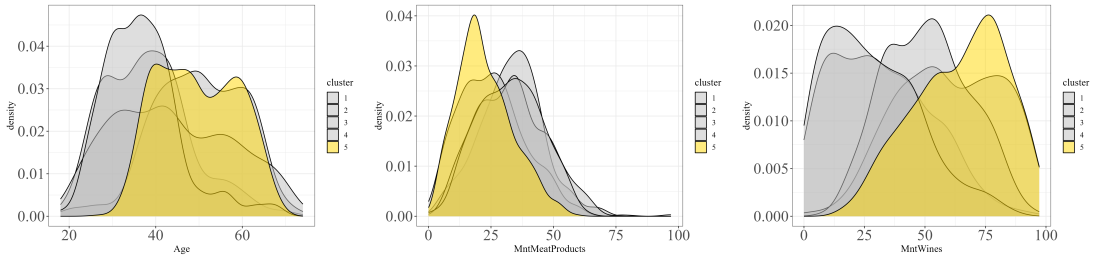


(e) Income (f) Fruits Category (g) Sweet Category (h) Gold Category

Figure 3.6: Cluster 5



(a) Number of kids (b) Number of teens (c) Marital Status



(d) Age (e) Meat Products (f) Wine Category

4 Supervised Learning

In this section, several classification algorithms will be used with the goal of building a predictive model of consumers' response to the campaign. In order to determine which of these models is best, a suitable metric must be selected.
To this end, it is necessary to consider the context in which we are operating. As we are reasoning from the perspective of building a new promotional campaign, one of the goals we can expect to achieve is to generate profits.

The company reports that to contact the 2240 customers surveyed, it spent the equivalent of $4 per person. In total 343 accepted the campaign, generating revenues totaling $3764, or $11 per positive response.

We now introduce the precision measure, which shows us the percentage by which our predictions of observing a positive campaign response are correct, and is calculated in this way:

$$P = \frac{T_p}{T_p + F_p}$$

where T_p is the number of true positive and F_p is the number of false positive.

Using the above data and assuming we want to make a 25% profit, we can write:

$$11P - 4 = 4 * 0.25 \rightarrow P \sim 45\%$$

We can then conclude that for a model to generate profits, it is necessary for it to reach a level of precision at least equal to 45 percent.

Having reached this necessary condition, however, it remains to decide how to choose one model over another. To do this, the recall measure will be used, which is calculated as follows:

$$P = \frac{T_p}{T_p + F_n}$$

where F_n is the number of false negative. The Recall level instructs us on the amount of customers who are likely to respond positively to the proposal that we are able to classify correctly with our model. These customers are a latent source of profit for the company, and it is important to aim to reach as many as possible.

Another factor to consider here is the imbalance of classes, where the minority class, is the one toward which the company has a greater interest in predicting correctly. Instead of using techniques that could balance this the dataset (like over/undersampling), since all models considered produce some probability of belonging to a class, and then classify the observation according to a predetermined treshold, it was decided to make a comparison of results for different levels of this treshold.

4.1 Logistic Regression

The first classification model examined is that of logistic regression. This uses the logistic function to estimate the relationship between the probability that the target variable is equal to 1, conditional on the explanatory variables.

$$\textbf{Logistic Function} : Pr(Response = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Which can be rewritten as:

$$\textbf{Log odds} : \log\left(\frac{Pr(Response = 1|X)}{1 - Pr(Response = 1|X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

The estimated coefficients we can observe in Figure 4.1 show the estimate change of log odds given a one-unit increase in the explanatory variable, and not the change in the probability of observing a positive response.

Evaluating this model on the training set shows that the best decision threshold to use to achieve a precision of 45% and obtain the highest recall level (0.831) is 0.13 (Table 4.1)

Looking at the results in the Figure 4.1 one thing that becomes clear is that many explanatory are classified as not statistically significant. In other words, their impact on the variable we are trying to estimate is limited. This is enough to motivate the study of models that limit the numbers of predictors.

Figure 4.1: Logistic Regression Output

```
Call:
glm(formula = Response ~ ., family = "binomial", data = dtrain)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.744e+00  9.288e-01  -4.031 5.54e-05 ***
Age                  1.514e-02  8.437e-03   1.794  0.07274 .
Income               1.424e-06  8.048e-06   0.177  0.85955
Kidhome              3.391e-01  2.404e-01   1.411  0.15829
Teenhome            -6.936e-01  2.414e-01  -2.873  0.00407 **
Ed_Basic            -1.163e+00  9.265e-01  -1.255  0.20947
Ed_Master            4.337e-01  2.268e-01   1.912  0.05588 .
Ed_PhD               1.317e+00  2.370e-01   5.557 2.74e-08 ***
MS_Married          -1.493e+00  2.403e-01  -6.212 5.23e-10 ***
MS_Together         -1.446e+00  2.669e-01  -5.418 6.04e-08 ***
MS_Divorced          2.590e-02  3.000e-01   0.086  0.93121
MS_Widow            -2.462e-03  4.166e-01  -0.006  0.99528
enrollment_days      4.809e-03  5.505e-04   8.736  < 2e-16 ***
Recency             -3.310e-02  3.471e-03  -9.537  < 2e-16 ***
MntWines            -4.386e-02  7.497e-03  -5.850 4.91e-09 ***
MntFruits           -1.547e-02  1.572e-02  -0.984  0.32516
MntFishProducts     -5.999e-02  1.529e-02  -3.924 8.72e-05 ***
MntSweetProducts    -2.376e-02  1.485e-02  -1.600  0.10969
MntGoldProds         2.015e-03  2.496e-03   0.807  0.41948
NumDealsPurchases    2.692e-02  5.616e-02   0.479  0.63169
NumWebPurchases      4.806e-02  9.440e-03   5.091 3.56e-07 ***
NumCatalogPurchases  5.506e-02  9.269e-03   5.940 2.85e-09 ***
NumWebVisitsMonth    8.593e-02  6.632e-02   1.296  0.19511
Complain            -2.070e-01  1.522e+00  -0.136  0.89182
CmpAccepted          1.844e+00  1.480e-01  12.458  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1514.71  on 1788  degrees of freedom
Residual deviance:  835.17  on 1764  degrees of freedom
AIC: 885.17

Number of Fisher Scoring iterations: 7
```

Table 4.1: Best Threshold: Logistic Regression

Threshold	Precision	Recall
0.11	0.430	0.892
0.12	0.441	0.862
0.13	**0.458**	**0.831**
0.14	0.452	0.800
0.15	0.455	0.769

4.1.1 Stepwise Selection

One possible solution to construct a model containing only a subset of the most important features is stepwise selection. Specifically, in this analysis the results of a Forward Stepwise Selection were examined. This method starting from a model without predictors adds to it, one at a time, the predictor that guarantees the best additional improvement to the fit, and finally selects the model that guarantees the best result measured with a pre-defined metric. The metric used in this case is the Akaine Information Criteria, which measures models by a trade-off between the degree to which the model fits the data and its complexity.

By doing so of the 24 original variables, 9 were removed, and the remaining ones are all reported as significant at a significance level of 0.05.

Figure 4.2: Logistic Regression after Forward Stepwise Selection Output

```
Call:
glm(formula = Response ~ CmpAccepted + enrollment_days + Recency +
    NumCatalogPurchases + NumWebPurchases + Teenhome + MS_Married +
    MS_Together + Ed_PhD + MntWines + MntFishProducts + Ed_Basic +
    NumWebVisitsMonth + Kidhome + Age + MntSweetProducts, family = "binomial",
    data = df)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.7406125  0.6167492  -6.065 1.32e-09 ***
CmpAccepted       1.7702394  0.1267783  13.963  < 2e-16 ***
enrollment_days   0.0045085  0.0004713   9.567  < 2e-16 ***
Recency          -0.0332870  0.0030368 -10.961  < 2e-16 ***
NumCatalogPurchases  0.0582430  0.0077110   7.553 4.24e-14 ***
NumWebPurchases   0.0451358  0.0078404   5.757 8.57e-09 ***
Teenhome         -0.7154846  0.1891730  -3.782 0.000155 ***
MS_Married       -1.3626828  0.1885514  -7.227 4.93e-13 ***
MS_Together      -1.2727793  0.2141315  -5.944 2.78e-09 ***
Ed_PhD            1.0573284  0.1919771   5.508 3.64e-08 ***
MntWines         -0.0391837  0.0060158  -6.513 7.34e-11 ***
MntFishProducts  -0.0598297  0.0134182  -4.459 8.24e-06 ***
Ed_Basic         -2.0744685  0.8561259  -2.423 0.015389 *
NumWebVisitsMonth    0.1103093  0.0473611   2.329 0.019853 *
Kidhome           0.4382205  0.1938596   2.261 0.023790 *
Age               0.0153650  0.0072205   2.128 0.033340 *
MntSweetProducts -0.0207521  0.0122075  -1.700 0.089141 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1885.5  on 2235  degrees of freedom
Residual deviance: 1072.4  on 2219  degrees of freedom
AIC: 1106.4

Number of Fisher Scoring iterations: 6
```

In order to assess whether this reduction in predictors improves predictions over the full model, we need to look at the results obtained with respect to the test set. In this case, the selected threshold level is 0.12, which reports a higher recall level (0.892) than the full logistic regression model, demonstrating an improvement in performance on the test set.

Table 4.2: Best thresholds: Forward Stepwise Selection

Threshold	Precision	Recall
0.10	0.432	0.923
0.11	0.444	0.923
0.12	**0.453**	**0.892**
0.13	0.456	0.877
0.14	0.470	0.862

4.1.2 Shrinkage Methods: The Lasso

One alternative to the method described above in which we selected variables by comparing different models with a different number of parameters, is to fit a single model containing all predictors and a penalty that shrinks the coefficient estimates toward zero. The two most widely used techniques for this purpose are Ridge and Lasso, where the latter has the advantage of forcing some of the coefficients to be exact zero as long as the tuning parameter lambda is sufficiently large.

Lambda is often choose as the value that minimizes the cross-validation errors. Another common practice is not to choose this value, but the highest lambda that is at most one standard deviation away from the minimum performance. This allows a larger penalty to be used, increasing the chance of seeing coefficients taken to zero, while still obtaining similar results. In this case the second approach has been used.

Figure 4.3: Lambda values

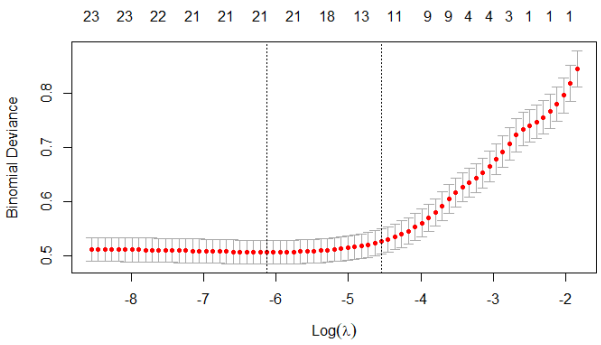


Table 4.3: Logistic Regression with Lasso Output

	Estimate
Intercept	-2.022
Teenhome	-0.218
Ed Basic	-0.074
Ed PhD	0.612
MS Married	-0.751
MS Together	-0.742
enrollment days	0.663
Recency	-0.639
MntWines	-0.215
MntFishProducts	-0.081

After selecting lambda, we can use it to estimate the model with Lasso. The result shown in Figure 4.3, contains only 9 variables, thus operating a greater reduction than forward stepwise selection.

This further reduction, however, did not lead to better performance; in fact, for the selected threshold level of 0.17, the recall level is significantly lower than for the previous models analyzed.

Table 4.4: Best thresholds: the Lasso

Threshold	**Precision**	**Recall**
0.15	0.388	0.800
0.16	0.413	0.800
0.17	**0.455**	**0.785**
0.18	0.442	0.708
0.19	0.451	0.708

4.2 Discriminant Analysis

In this section we analyze techniques that, instead of directly modelling PR(Y = k | X = x), like in the logistic regression, try to model the distribution of the predictors X separately in each of the response classes. Only after this the Bayes's theorem it is applied to turn these distributions into estimates for Pr(Y = k | X = x).
The key point in this approach to obtain good results is to use an estimate of the density function of X for different classes which approximate well the reality.
We will study two classifiers that use different estimates: the Linear Discriminant Analysis and the Quadratic Discriminant Analysis.

4.2.1 Linear Discriminant Analysis

In the Linear Discriminant Analysis, when we have more than one predictor, it is assumed that $X = (X_1, X_2, \ldots, X_p)$ is drawn from a multivariate Gaussian distribution, with a class-specific mean vector and a common covariance matrix.

What it tries to do is to find a linear transformation of the original features to maximize class separation while minimizing within-class variance.

The Figure 4.4 shows the coefficients of the first discriminant function, indicating the direction and strength of the relationship between the features and the discriminant function. These coefficients are of particular interests since they show insights into the importance of original features in the discriminant process.

Figure 4.4: Components of the first discriminant function

	LD1 <S3: AsIs>
CmpAccepted	1.2760774475957
MS_Together	-0.6889537220404
Ed_PhD	0.6281087005025
MS_Married	-0.6258450457468
Complain	0.3491463646896
Ed_Basic	-0.3383488647614
Teenhome	-0.2246094504575
Ed_Master	0.1893923883414
Kidhome	0.1164373842343
MS_Widow	0.0749103257095
MS_Divorced	-0.0529912424200
NumWebVisitsMonth	0.0322062291918
NumCatalogPurchases	0.0220656425803
MntFishProducts	-0.0216958653177
MntWines	-0.0198339591277
NumWebPurchases	0.0197495658598
Recency	-0.0141823769739
MntSweetProducts	-0.0124376944445
MntFruits	-0.0091867738503
NumDealsPurchases	0.0054269697406
Age	0.0048352439622
enrollment_days	0.0019272729739
MntGoldProds	0.0013481574329
Income	-0.0000003943382

For what concern the results of the predictions made with the LDA model, we can see that it generally performs poorer than the logistic regression, a sign that may indicate that the dataset does not meet the LDA assumptions.

Table 4.5: Best thresholds: LDA

Threshold	Precision	Recall
0.09	0.423	0.800
0.10	0.444	0.800
0.11	**0.468**	**0.800**
0.12	0.486	0.785
0.13	0.485	0.754

4.2.2 Quadratic Discriminant Analysis

QDA is similar to LDA, but it makes different assumptions about the underlying data distribution. QDA, unlike LDA, assumes that each class has its own covariance matrix, resulting in a more flexible decision boundary.

Since QDA needs to estimate a separate covariance matrix for each class, it needs to estimate much more parameters, and usually requires larger amount of data in order to obtain good results. This may be one of the reasons that might explain the results that show QDA having much difficulty finding a balance between accuracy and recall.

Table 4.6: Best thresholds: LDA

Threshold	Precision	Recall
0.69	0.449	0.477
0.70	0.449	0.477
0.71	**0.456**	**0.477**
0.72	0.463	0.477
0.73	0.463	0.477

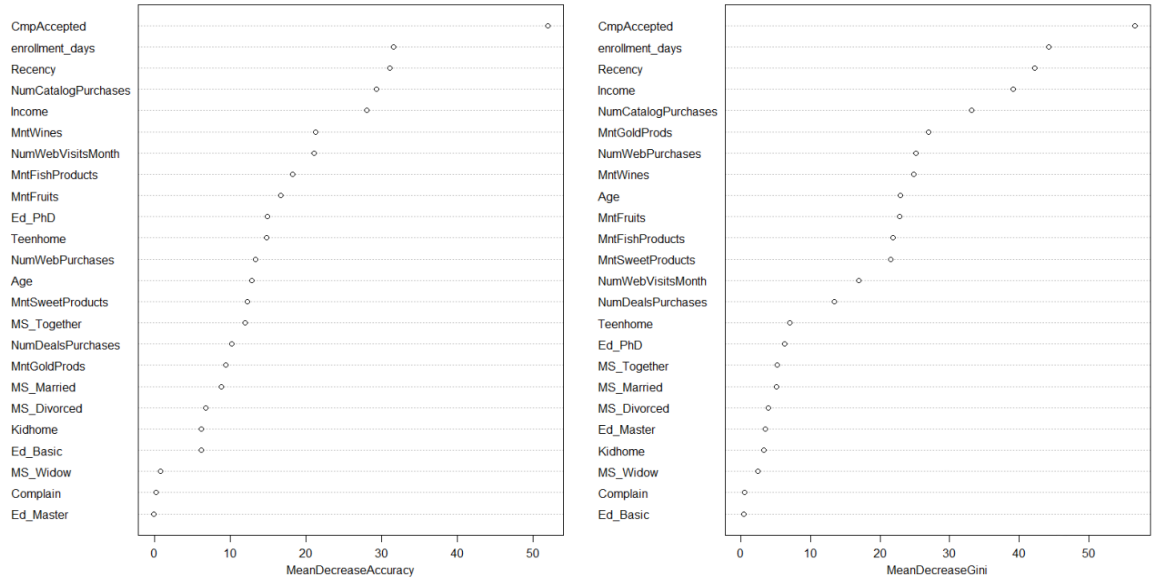
4.3 Random Forests

The Random Forest is a method used both for regression and classification which is built upon the structure of decision trees. The Random Forest creates different sample using bagging, which implies creating subsets of the training data by random sampling with

replacement the original sample, and use them to train individual decision trees. In addition to this, it selects a random subset of features that each decision tree can consider, which promotes diversity among trees. For classification problem then it combines the predictions of individuals trees using a majority vote.

The Fig 4.5 shows the importance that the algorithm assign to each feature based on two metrics. The left panel shows the the Mean Decrease in Accuracy or, in other words how much accuracy the model losses by excluding each variable. On the right panel the Gini Importance is reported, which is the average of how each feature decrease the impurity of the split.

Figure 4.5: Random Forest: Feature Importance



Despite being a powerful ensemble learning algorithm, also in this case the results of Random Forests do not outperform the simpler logistic model. The algorithm for a best threshold of 0.21, indeed obtain a level of recall of just 0.785

Table 4.7: Best thresholds: Random Forest

Threshold	Precision	Recall
0.19	0.401	0.785
0.20	0.425	0.785
0.21	**0.451**	**0.785**
0.22	0.459	0.769
0.23	0.462	0.738

4.4 Feature Importance

Besides making prediction, the models we have seen carry with them other important information. One of these concerns feature importance, that is, how relevant they consider a variable to be in determining a customer's response. Comparing the results of different models with respect to this information can be useful for the company to focus its efforts in working on one consumer characteristic rather than another.

- In the case of logistic regression with stepwise selection, the order in which the variables are selected to be included in the model represents their impact in explaining the target variable and thus can be viewed as a ranking of importance.
- In the case of LDA we have already mentioned that the coefficients of the discriminant function indicate the importance a variable has in distinguishing the two classes.
- As for the random forest, it too reports two possible orderings of how a predictors affects the final response.

Table 4.8 reports the five most important features for each model. We can observe that the *CmpAccepted* variable is the most important one in all of them, meaning that the previous tendency that a customer had towards accepting this kind of campaign plays a crucial role in determining whether or not he will accept a new one.

Other two variables on which logistic stepwise and random forest agree are *Recency* and *enrollment days*, respectively the number of days since the last purchase and how long a customer has been part of the company. Their importance is also supported by the fact that both of these variables have been included in the Lasso Model. We can interpret as an hint that customer which are more "active" and loyal to the company are more willing to accept the campaign.

Table 4.8: Feature Importance: different models

Stepwise Logistic	**LDA**	**RandomForest**
CmpAccepted	CmpAccepted	CmpAccepted
enrollment days	MSTogether	enrollmentdays
Recency	EdPhD	Recency
NumCatalogPurchases	MSMarried	NumCatalogPurchases
NumWebPurchases	Complain	Income

There is also a bit less strong indication that customer who prefer catalogue purchases may be more tempted to positively answer to a promotional campaign.

5 Conclusion

In conclusion to this analysis the most significant aspects are reported:

- iFood consumers can be divided into five clusters that classify them primarily according to their family unit, which then implies differences in terms of both products purchased and channels used, information that can lead to the company offering them more effective promotions and advertising.
- the best model for predicting consumer response is the stepwise logistic model, which, in addition to meeting the level of accuracy required to make a 25 percent profit, can identify 89 percent of consumers who responded positively to the campaign. This means being able to reach nearly 9 out of 10 consumers who are predisposed to like such a proposal, thus increasing their engagement with the company.
- the predisposition to accept these offers, the time spent with the company, and the fact that they have made recent purchases are the variables that more than others impact this choice, thus making them the aspects on which the company should be most interested when planning to whom to offer the next promotional campaigns