Final Project. Group 68

Emna Bouassida 20221740
Antonina Filatova 20221104
Marta Manevska 20220056
Ludovico Toscano 20220044

Abstract

This report reviews the analysis conducted on a customer aggregated data dataset provided by New Supermarkets International which was aimed to reveal data patterns, conduct customer segmentation and predict customer response to the subscription offer. Through data exploration negative entries in the customer spending column were revealed, they stand for 8,3% of data: the initial description of data can't be considered accurate, since it doesn't stand for "amount spent" but shows profit or loss that a certain customer has brought to the company. Clusterization was performed using percentage spent for each category, resulting in four clusters: Perishable Lovers, Beverage Lovers, Beverage and Canned Lovers, Frozen and Other Lovers. To maximize profit, it is recommended to target the top 15% of customers, with a predicted acceptance rate of 52% for the subscription offer.

Introduction

The objectives of data engineering are to find new selling patterns and develop new focused programs for the company New Supermarkets International in order to maximize the profit. We are using the SAS software to get smart with the database and provide an analytical project for further decision making.

The dataset provided by the company includes aggregated customer data identified by CustId. The dataset contains demographic information such as education, marital status, age, gender, dependents, and income. Additionally, it includes behavioral data such as frequency, recency, and internet usage. The dataset also includes different lines of business, including perishables, beverages, frozen, canned, and others. Lastly, the dataset includes the Net Promoter Score (NPS).

Exploratory and Descriptive Analysis

Variables

First step working with SAS software was changing the roles and levels of the variables from the collected data. For the levels of the variable, we change customer ID to Nominal level and Dependents to Binary Variable because we have 0 and 1 meaning if the customer has dependents in the household or not.

We changed the role of the Target variable to a rejected variable just for the purposes for the exploratory analysis. The role of CustomerID as ID role and Frequency as Interval role. NPS represents Adapted Net Promoter Score from 1 to 5, we consider it to be an ordinal variable, since it is a score and the distance between points is not equal.

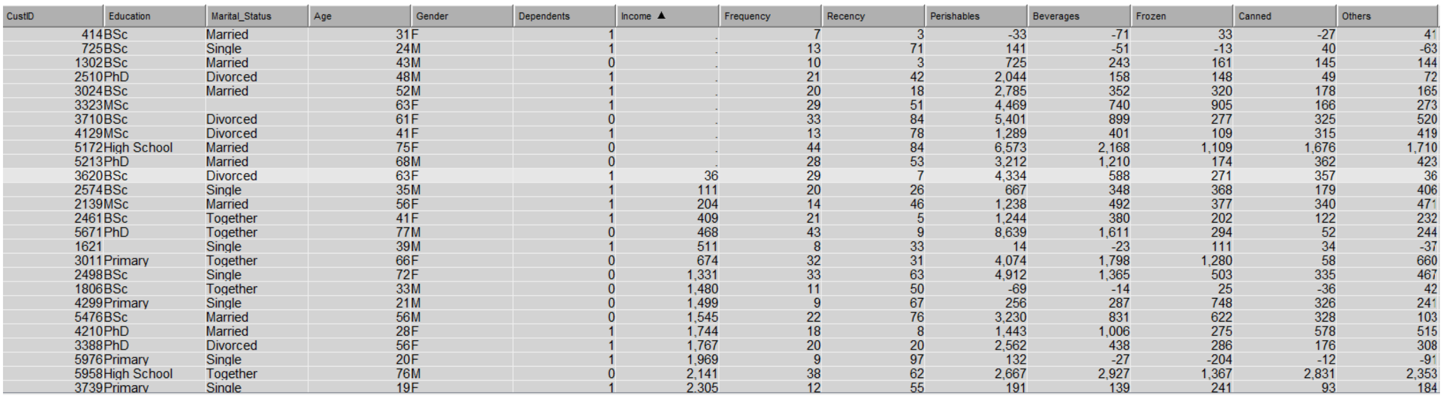
We considered performing feature engineering, such as analyzing spending on the internet or calculating average spend per visit. However, it is not feasible. The internet sales data represents lifetime values, while the amount spent on the line of business pertains to 2022, and the frequency data corresponds to 2021

Negative values & discrepancies

Values of Recency go from 1 to 365, they show customers that joined over the past year.

We discovered negative values in observations of the product categories, that they were not expected since the document with the explanation of the features says that these are the amount spent on each line of businesses. Since we have 8,3% of negative values in all the columns we believe that the documentation was wrong, and that these features are representing customer value brought to the company (it can stand for product returned when negative or products bought on discount, meaning no profit for the supermarket), so the decision was made to keep these observations. Hence, the first business suggestion is to provide the correct data in order to allow better analysis.

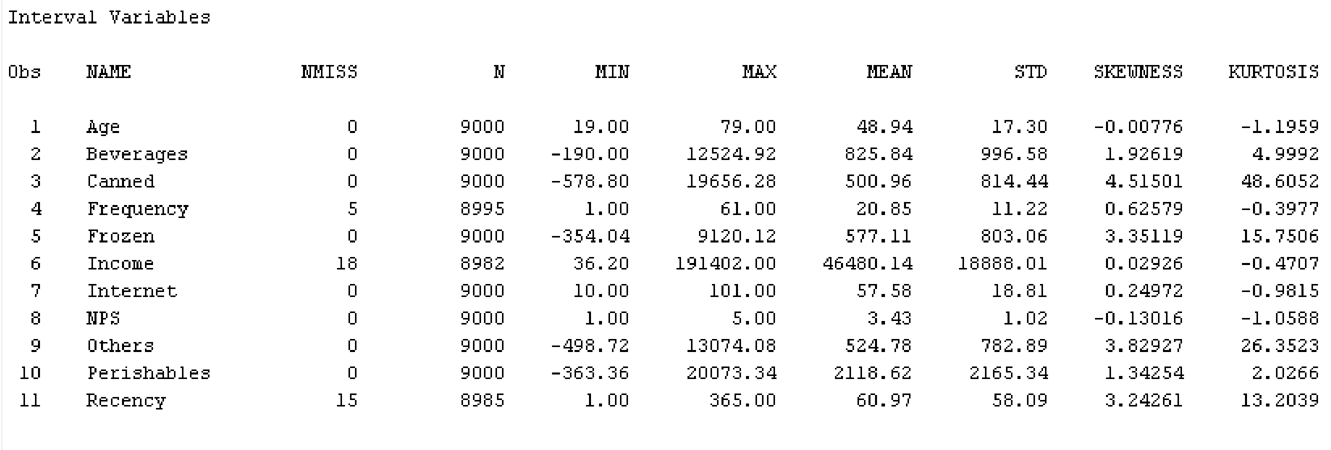
Going to the node graphical exploration there is an example table and we can see that some customers have low income and the amount of money they spend is more than what they earn. We don't consider these values to be errors. It is possible that these customers are financially supported by their family members or partners, otherwise they intentionally indicate income that is lower than the real one. For example the income is 36 and the amount spent on a line of a product is much higher, for example perishables is 4,334. Later we do feature engineering which will be the sum of money spent on all the categories.



CustID	Education	Marital_Status	Age	Gender	Dependents	Income ▲	Frequency	Recency	Perishables	Beverages	Frozen	Canned	Others
414	BSc	Married	31	F	1	.	7	3	-33	-71	33	-27	41
725	BSc	Single	24	M	1	.	13	71	141	-51	-13	40	-63
1302	BSc	Married	43	M	0	.	10	3	725	243	161	145	144
2510	PhD	Divorced	48	M	1	.	21	42	2,044	158	148	49	72
3024	BSc	Married	52	M	1	.	20	18	2,785	352	320	178	165
3323	MSc		63	F	1	.	29	51	4,469	740	905	166	273
3710	BSc	Divorced	61	F	0	.	33	84	5,401	899	277	325	520
4129	MSc	Divorced	41	F	1	.	13	78	1,289	401	109	315	419
5172	High School	Married	75	F	0	.	44	84	6,573	2,168	1,109	1,676	1,710
5213	PhD	Married	68	M	0	.	28	53	3,212	1,210	174	362	423
3620	BSc	Divorced	63	F	1	36	29	7	4,334	588	271	357	36
2574	BSc	Single	35	M	1	111	20	26	667	348	368	179	406
2139	MSc	Married	56	F	1	204	14	46	1,238	492	377	340	471
2461	BSc	Together	41	F	1	409	21	5	1,244	380	202	122	232
5671	PhD	Together	77	M	0	468	43	9	8,639	1,611	294	52	244
1621		Single	39	M	1	511	8	33	14	-23	111	34	-37
3011	Primary	Together	66	F	0	674	32	31	4,074	1,798	1,280	58	660
2498	BSc	Single	72	F	0	1,331	33	63	4,912	1,365	503	335	467
1806	BSc	Together	33	M	0	1,480	11	50	-69	-14	25	-36	42
4299	Primary	Single	21	M	0	1,499	9	67	256	287	748	326	241
5476	BSc	Married	56	M	0	1,545	22	76	3,230	831	622	328	103
4210	PhD	Married	28	F	1	1,744	18	8	1,443	1,006	275	578	515
3388	PhD	Divorced	56	F	1	1,767	20	20	2,562	438	286	176	308
5976	Primary	Single	20	F	1	1,969	9	97	132	-27	-204	-12	-91
5958	High School	Together	76	M	0	2,141	38	62	2,667	2,927	1,367	2,831	2,353
3739	Primary	Single	19	F	1	2,305	12	55	191	139	241	93	184

Missing values

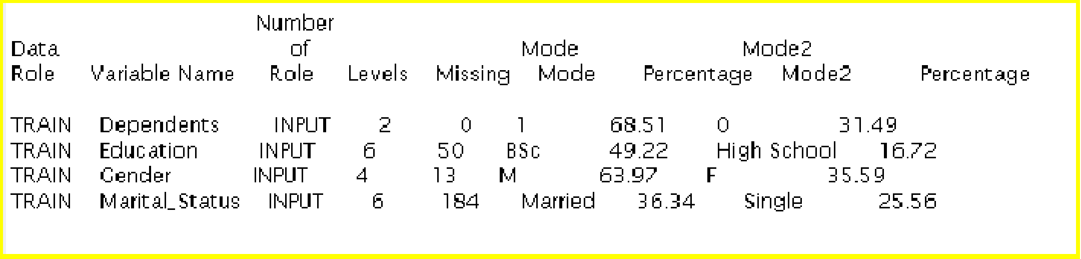
We can see that there are missing values for Frequency, Income and Recency. Education, Marital_Status and Gender show missing values as well. At this point the rows were not deleted. Moreover, we will consider excluding them in case they will be used for clustering and classification.



```
Interval Variables

Obs   NAME          NMISS      N      MIN        MAX       MEAN        STD    SKEWNESS   KURTOSIS

  1   Age             0      9000     19.00      79.00      48.94      17.30   -0.00776    -1.1959
  2   Beverages       0      9000   -190.00   12524.92     825.84     996.58    1.92619     4.9992
  3   Canned          0      9000   -578.80   19656.28     500.96     814.44    4.51501    48.6052
  4   Frequency       5      8995      1.00      61.00      20.85      11.22    0.62579    -0.3977
  5   Frozen          0      9000   -354.04    9120.12     577.11     803.06    3.35119    15.7506
  6   Income         18      8982     36.20  191402.00   46480.14   18888.01    0.02926    -0.4707
  7   Internet        0      9000     10.00     101.00      57.58      18.81    0.24972    -0.9815
  8   NPS             0      9000      1.00       5.00       3.43       1.02   -0.13016    -1.0588
  9   Others          0      9000   -498.72   13074.08     524.78     782.89    3.82927    26.3523
 10   Perishables     0      9000   -363.36   20073.34    2118.62    2165.34    1.34254     2.0266
 11   Recency        15      8985      1.00     365.00      60.97      58.09    3.24261    13.2039
```



```
                          Number
Data                        of                    Mode               Mode2
Role     Variable Name     Role    Levels   Missing    Mode    Percentage    Mode2          Percentage

TRAIN    Dependents        INPUT      2        0       1         68.51       0              31.49
TRAIN    Education         INPUT      6       50       BSc       49.22       High School    16.72
TRAIN    Gender            INPUT      4       13       M         63.97       F              35.59
TRAIN    Marital_Status    INPUT      6      184       Married   36.34       Single         25.56
```

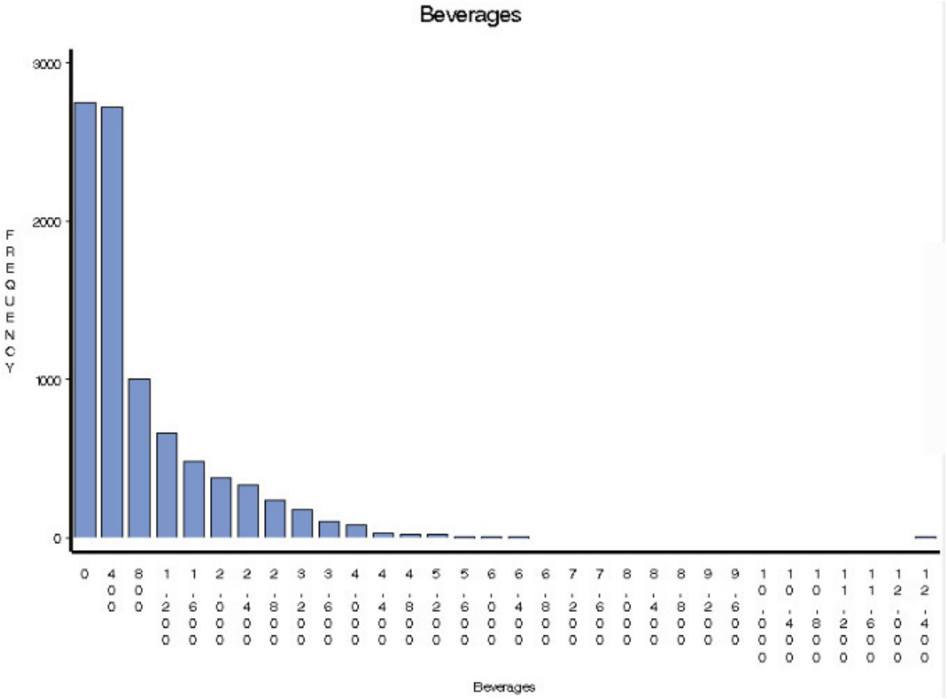
Distribution of original values

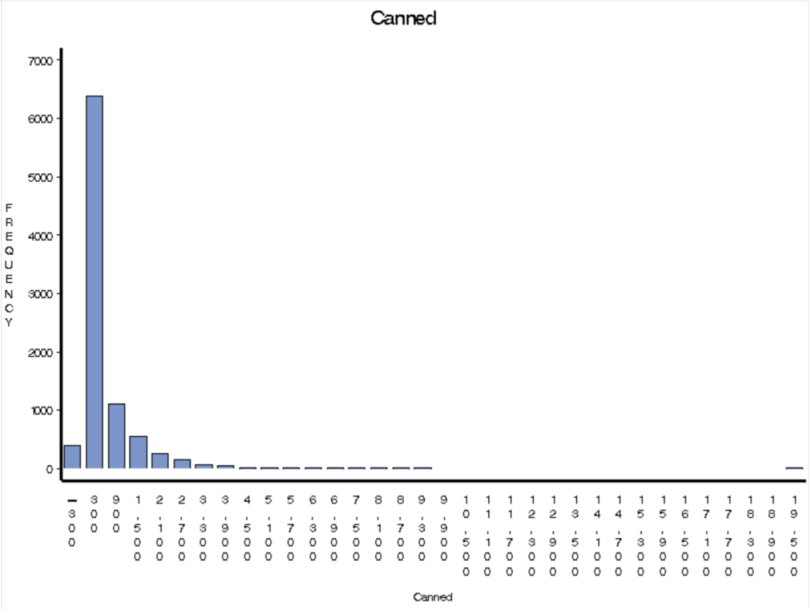
The treatment that we did on these features will be used only for the product clustering that we made with absolute values, but as we will see later since it was not so insightful so we used percentages spent in each line of business. For the classification task we used these features without treatment because all the outliers were on the test dataset, which is not a problem since the outliers have to be treated only for the training dataset. The only row that we removed everywhere is on the internet because it has 101% of internet purchases which probably is a mistake.

Since deleting data is not always recommended we decided to transform the outliers because the k-means will bring relatively close values together anyways, as long as the distance between them is not drastically huge.

Taking this into consideration, we transform them in that sense. For example, the top 2 spenders of “Beverages” show values of 12525 and 6230. Since the first value is twice higher than the second, and we cannot make a cluster for each customer, the 1st one needs to be changed so that it doesn’t affect the rest of the clustering, we can transform the first value to another one closer to the second, which will make them classified in the same cluster, as top spenders.

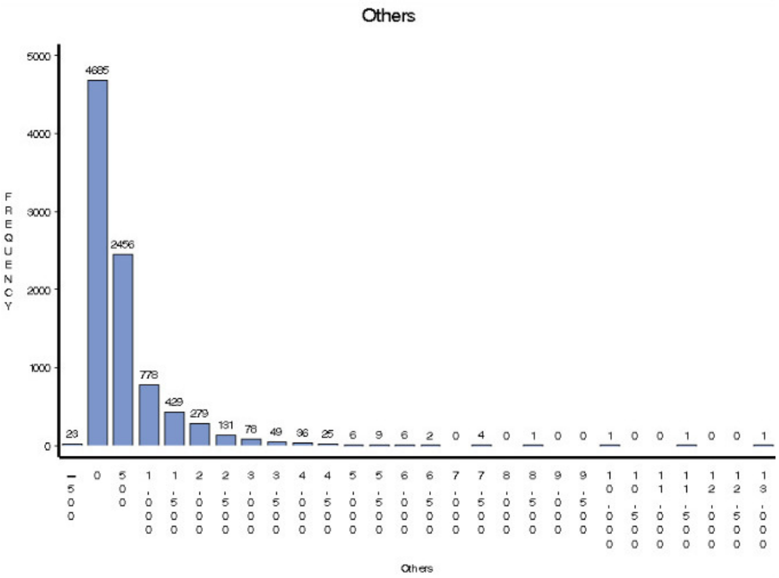
Further on, we checked the distribution of the data, and we found some outliers, in Beverages, Canned, Perishables, Others and Income columns.
Beverages and Canned goods have only one value which is two times higher than the second highest.



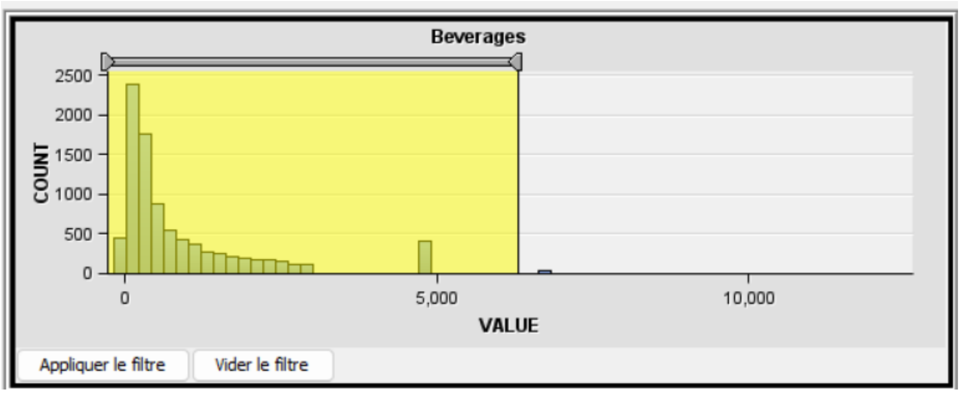


In Canned there is only one outlier of 19,500. The maximum value set in the filter was 9500.

For “others” there is a huge difference between the 9th value and the first 8 ones, as you can see on the graph. The 8 highest values were dropped.

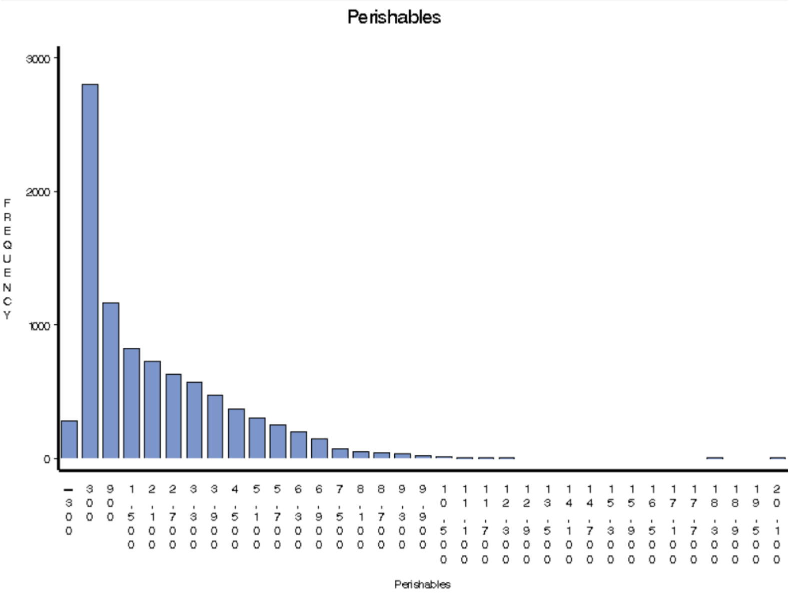


For Perishables the two highest values are 150% higher than the third one.

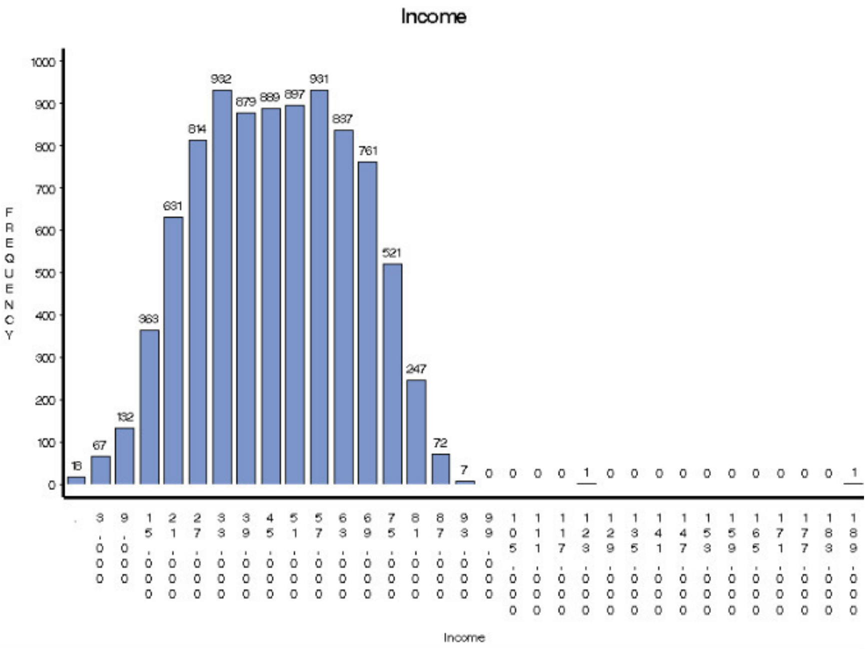


Using the multiplot histogram, we see that it complies with the values that we see in the Excel file, which show only one outlier of 12,400 for the "beverages" variable. However, when we use the Filter node, we notice that the indication of the values is quite different, in the sense that we don't see the outlier of 12,400, but also, we see "fake" outliers, because there are omitted values that we don't see in the scale above.

As a consequence, if we rely on the filter node we would cut more observations than what needs to be cut. Taking this into consideration, we set the maximum value to be 6300.



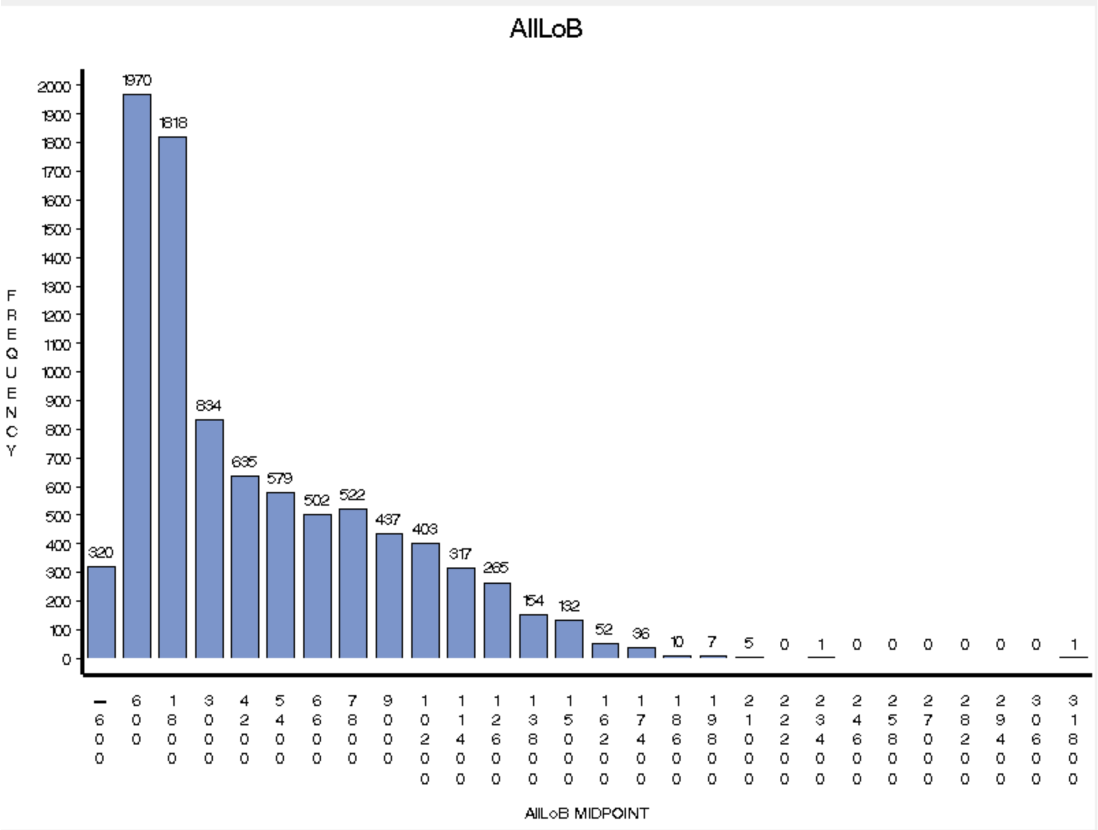
For Perishables, we have two outliers, so the limit was set to 12,600.



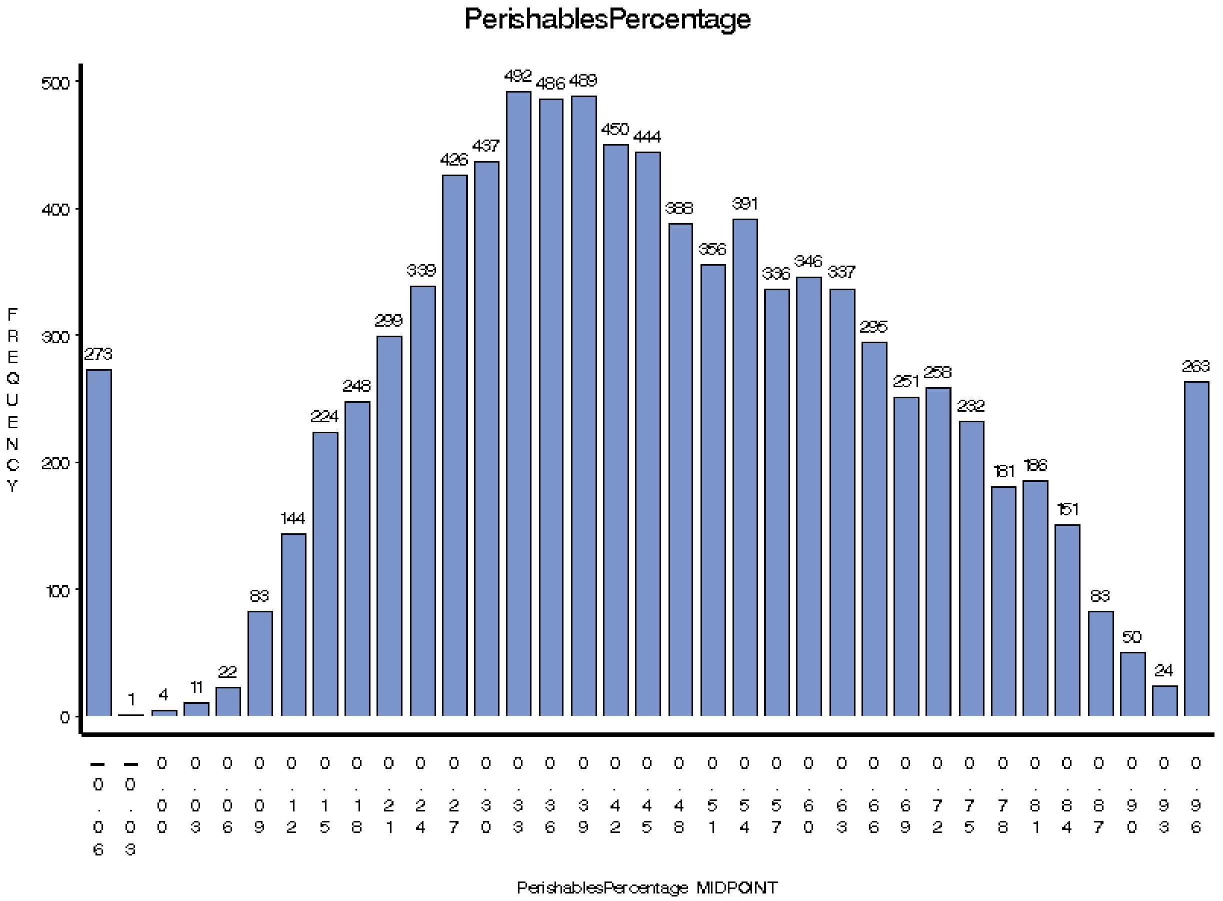
For income we transform two outliers.

Distribution of the engineered values

The feature is the sum of all the lines of business called AllLoB (all lines of business), present one outlier but we will not treat it because it belongs to the test dataset.



The other new features which represent how much a customer has spent on each line of business in percentage have many outliers, this is because having negative and positive values lead to a scale that is not between 0% and 100% but can be for example between -300% and 500%. This bigger difference between the smallest and bigger percentage is more likely to create outliers which is what happened in our case. What we decided to do was to transform them like we did for the previous clustering since we will use them for the clustering again. The result for all the features is pretty similar, now on the head and the tail of the graph a lot of values.



Correlation

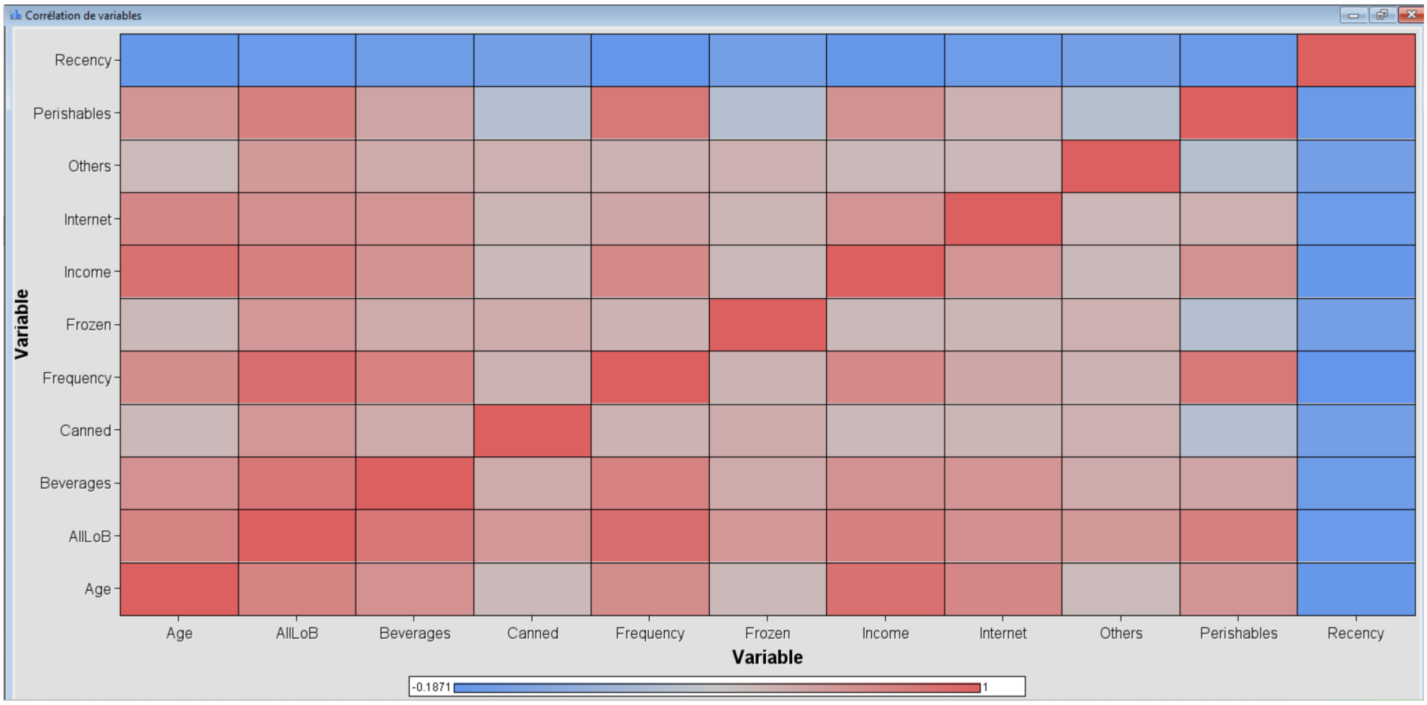
Looking at correlation is important for two reasons. It can give us some insight about existing patterns in data and later on it will be useful to understand which variables need to be included in the clustering or classification.

Correlation matrix below shows that Income and Age have a high correlation of 90%, which was expected since usually the older people are more likely to cover a higher job position. Also internet and frequency have a correlation over 70% with each other and with income and age. Beverages that show a correlation over 70% with all the variables mentioned before.

We can see that from the table below. For example, the variable Perishables is much more correlated with the variables Frequency and Income, than with Beverages and Canned.

In this matrix you can also see that we have AllLoB which is the sum of all the lines of business, it is highly correlated with the income variable. That means also that this variable is highly correlated with the ones correlated with income.

Later in order to perform the classification we had to deal with missing values, and in this case the correlation data turned out to be useful. For example Income has some missing values. Sometimes this happens because rich people don't want to reveal their income so knowing that income is correlated with age and AllLoB we took a look at what we were removing before to perform the classification. What we discovered was that most likely in this case the excluded observations were not about rich people, at least not people that behave like the other rich that we have on the dataset. Is important to try to understand why we have missing values, in case we want to delete the observation. For example, it is a problem if that observation belongs to our richest customer. Same applies when we want to impute the values, we can't impute the mean if we know that the missing income is about rich customers.



Customer Segmentation

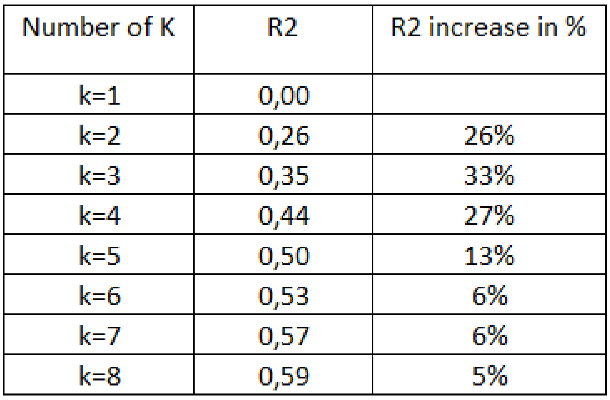
We tried to capture three different types of insight with clustering, one was how different demographics behave, one is to cluster using the absolute values "spent" by customers and the third one to see how much they spend in percentage on each line of business.

The demographics clustering wasn't insightful, because after trying to choose different variables we only got two clusters one with people with low income and another one with people with high income. And as we will see later the cluster made with the absolute amount spent already captures this difference.
Clustering by the absolute amount spent we found 3 distinct clusters, the table below:

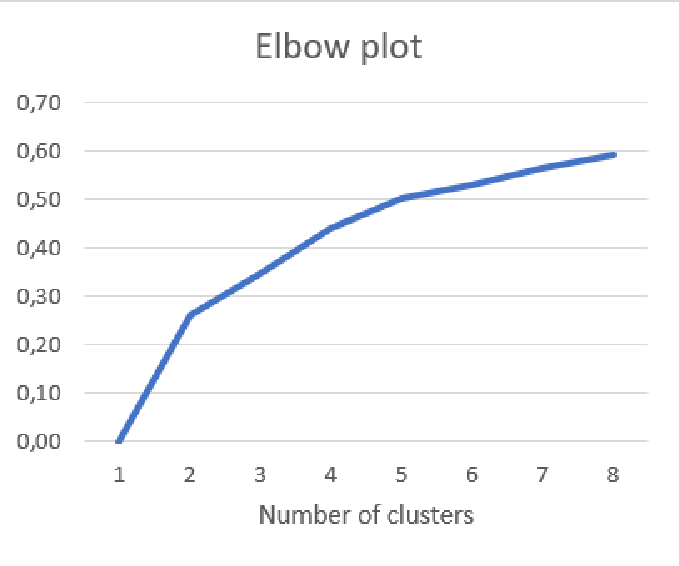
Segment Id	*Beverages*	*Canned*	*Frozen*	*Others*	*Perishables*	*Name*
1	*260.76*	*168.07*	*241.76*	*203.98*	*864.99*	*Low spenders*
2	*1452.89*	*624.87*	*697.10*	*645.04*	*4682.86*	*Perishable lovers*
3	*2451.55*	*1945.96*	*2062.86*	*1879.41*	*3329.32*	*High spenders*

However, this approach introduces bias in defining clusters by attributing customers to clusters based on whether they are high spenders or low spenders, instead of focusing on their preference to certain product categories. Similar results can be obtained with RFM analysis, for instance. Anyway we made a discovery on the profiling node, which is that high spenders are used to having a higher NPS.

Thus we decided to perform clustering based on the percentage of spendings on each line of business.
To choose the optimal number of clusters created by K-means we made an elbow plot showing R2 for each cluster. Below we can see the elbow plot for the percentage clustering, at this point the plot is quite ambiguous and doesn’t provide a clear choice for the number of K, so we move to the next step.



Number of K	R2	R2 increase in %
k=1	0,00	
k=2	0,26	26%
k=3	0,35	33%
k=4	0,44	27%
k=5	0,50	13%
k=6	0,53	6%
k=7	0,57	6%
k=8	0,59	5%



Below there are tables for each clustering with the means of variables. Comparing them,we can choose which clustering has the most distinct differences between clusters.

Average statistics of clusters for k=2

Segment Id	*Beverages Percentage*	*Canned Percentage*	*Frozen Percentage*	*Others Percentage*	*Perishables Percentage*	*Name*	*Cluster frequency*
1	*14%*	*4%*	*10%*	*6%*	*65%*	*Perishable lovers*	*4214*
2	*16%*	*14%*	*22%*	*17%*	*29%*	*Perishable haters*	*4786*

Average statistics of clusters for k=3

Segment Id	*Beverages Percentage*	*Canned Percentage*	*Frozen Percentage*	*Others Percentage*	*Perishables Percentage*	*Name*	*Cluster frequency*
1	*11%*	*3%*	*8%*	*4%*	*72%*	*Perishable lovers*	*2874*
2	*20%*	*12%*	*16%*	*11%*	*40%*	*Beverage lovers*	*4277*
3	*9%*	*14%*	*31%*	*25%*	*18%*	*Canned, Frozen & Other lovers*	*1849*

Average statistics of clusters for k=4

Segment Id	*Beverages Percentage*	*Canned Percentage*	*Frozen Percentage*	*Others Percentage*	*Perishables Percentage*	*Name*	*Cluster frequency*

1	*9%*	*1%*	*6%*	*2%*	*79%*	*Perishable lovers*	*1686*
2	*19%*	*7%*	*13%*	*9%*	*51%*	*Beverage lovers*	*3590*
3	*18%*	*16%*	*20%*	*18%*	*27%*	*Beverage and canned lovers*	*3398*
4	*-22%*	*5%*	*70%*	*24%*	*13%*	*Frozen and other lovers*	*325*

Average statistics of clusters for k=5

Segment Id	*BeveragesPercentage*	*CannedPercentage*	*FrozenPercentage*	*OthersPercentage*	*PerishablesPercentage*	*Name*
1	*10%*	*0%*	*2%*	*2%*	*82%*	*Perishable lovers*
2	*17%*	*7%*	*10%*	*7%*	*58%*	*Beverage & Perishable lovers*
3	*19%*	*8%*	*20%*	*18%*	*35%*	*Others and beverages lovers*
4	*17%*	*20%*	*21%*	*17%*	*24%*	*Canned & others & beverage lovers*
5	*-38%*	*2%*	*77%*	*14%*	*30%*	*Frozen lovers*

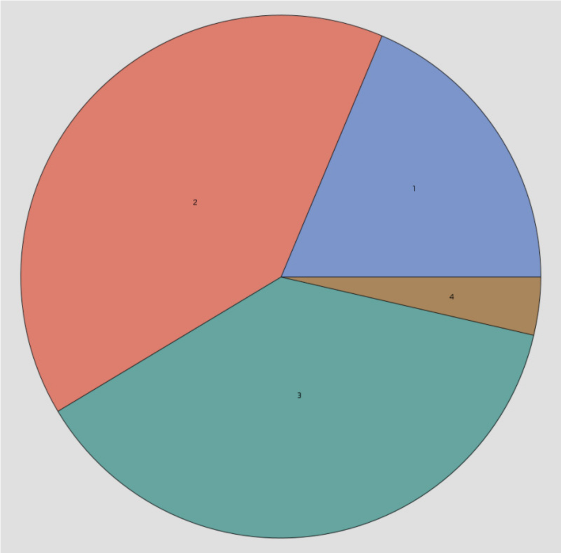
Dividing customers into 2 clusters is not very insightful, since the clustering is done only based on their purchase of perishables. Three clusters show more buying patterns, but four clusters provide a more distinguished grouping of customers. In reality, it is possible to go up to even 7 clusters, but we chose to stop at k=4 firstly in order not to spend more money, time and effort than needed, but also because it is already giving distinct clusters with different results, based on the variables chosen, which will allow us to do a more specific and effective targeting. For example, cluster 3 and 4 are lovers of 2 products simultaneously, but cluster 1 and 2 are mainly lovers of 1 specific product each.

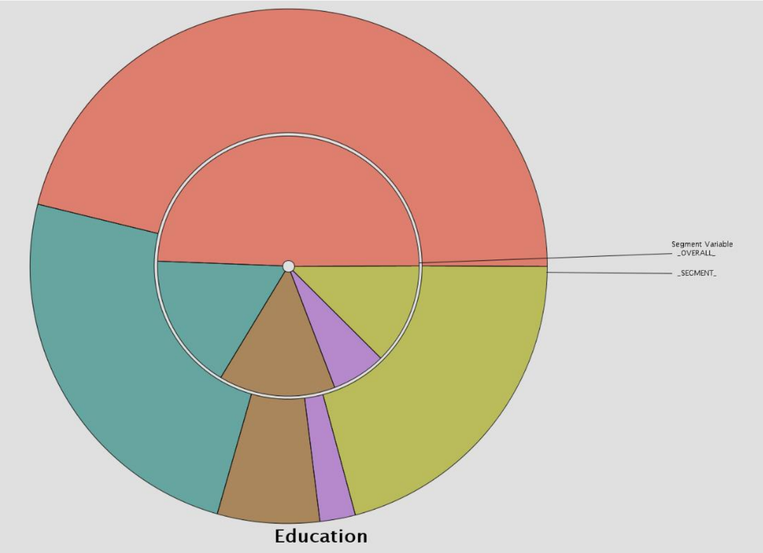
Clusters description

Chart below shows the distribution of the clusters, we can see that clusters 3 and 2 are the largest and cluster 4 is the smallest. Adding data on the AllLoB amount spent by each cluster reveals their value brought to the company. As expected, two largest clusters 2 and 3 are spending the most in absolute amount. Dividing these amounts by cluster size and looking at the average spent per customer inside the cluster we can see that cluster 2 has the highest spent per customer, followed by cluster 3. While cluster 1 has slightly lower spending per customer, there is a big gap between size of the cluster and the value brought to the company for cluster 4.

All lines of business spent, by each cluster

Cluster	***Total spent***	***Cluster size***	***Total spent, %***	***Cluster size, %***	***Av spent per customer***
Cluster 1	7459465	1686	18,2%	19%	4424
Cluster 2	17218759	3590	42,1%	40%	4796
Cluster 3	16092571	3398	39,3%	38%	4736
Cluster 4	137334	325	0,3%	4%	423
Total	**40908129**	**8999**	**100%**	**100%**	**14379**





Cluster frequency

Education in cluster 3

Cluster 1, **Perishable Lovers,** represents customers who have a strong preference for perishable products, with the majority of their purchases (79%) within the perishables category. They have a relatively lower interest in beverages, canned items, frozen products, and other categories.

Cluster 2, **Beverage Lovers,** comprises customers who have a significant preference for beverages. They also show moderate interest in canned items, frozen products, and other categories. However, their preference for perishable items is relatively lower. This is the largest cluster among the four and it is driving the company earnings up. Beverage Lovers are the most valuable customers.

Cluster 3, **Beverage and Canned Lovers**, represents customers who prefer both beverages and canned items. They also show some interest in frozen products and other categories, while their preference for perishable items is relatively lower. This is the second biggest cluster.

Cluster 4, **Frozen and Other Lovers**, consists of customers with a strong preference for frozen products and other categories. Interestingly, they show a negative percentage for beverages, most likely they buy these products only on discount. Their preference for canned items and perishable products is relatively moderate. This cluster is relatively small compared to the others. Importantly, this cluster is bringing the least value to the Supermarket.

The only significant difference for other features between clusters found on profile node was related to the Education variable in cluster 3. Compared to total customers, Beverage and Canned lovers are more numerous for primary school, are two times less graduated from Master degree and three times less from PhD compared to the total sample. Overall, we can say that cluster 3 have a lower education degree.

To create a complete marketing strategy based on the clusters, they have to be enriched with behavioral data.

Predictive Modeling

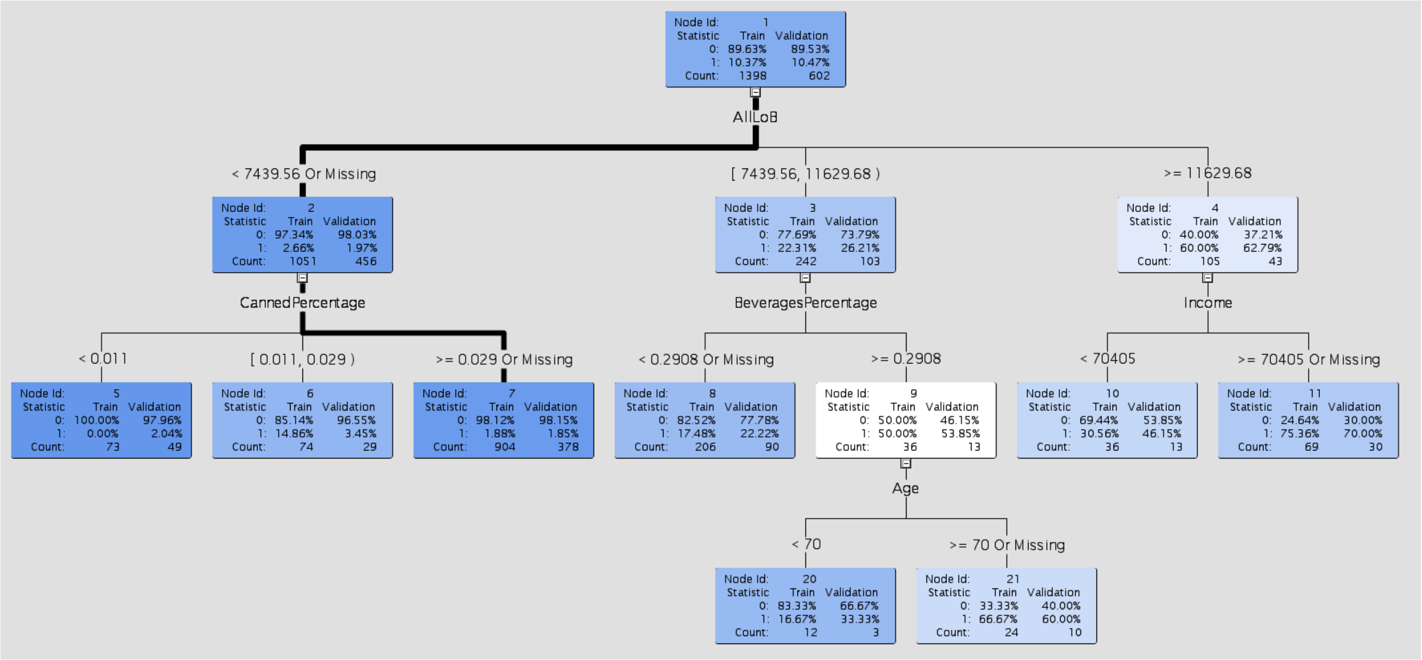
Missing values for the classification

We did this list in order to decide how to test which feature to keep and which to remove based on missing value, chi-square and worth. Then of course we test everything.
Features with missing values can't be used for classification, however, not all features are needed for classification, as some provide low signal. So here is the review of the features and their importance:

1. Education has missing values and low chi square score -> can be removed from the analysis
2. Marital status has missing values and low chi square score -> can be removed from the analysis
3. Gender has missing values and low chi square score -> can be removed from the analysis
4. Income has missing values but high worth -> so we will remove the missing values
5. Frequency has high worth and five missing values -> since they are few we removed them
6. Recency has missing values and low variable worth -> so we can delete this column.
7. AllLoB has the highest worth and no missing values -> so we can keep it.

Besides deleting the missing values we also considered imputing them, but since they are very few (4 of the variable frequency) they will not impact the overall result so we just delete them. Considering the feature selection that we did we excluded some features with missing values, and also most of the missing values belong to the test dataset.

Decision tree graph

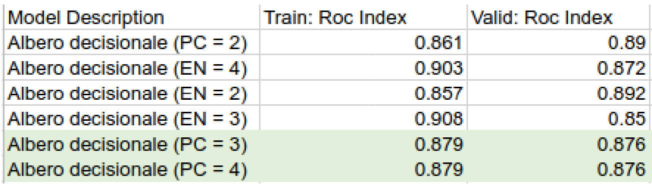


Even if we didn't decide to use the tree as we said below, watching how the tree makes the decision can be helpful to understand what are the most important variables for the business. In this case we have AllLoB which is also the one with the highest worth. We think that AllLoB is so important because the more you spend at our supermarket the more likely you like (according to the product cluster) the supermarket, and the more likely you are likely to buy the subscription. For the other features we didn't find a good interpretation supported by our findings.

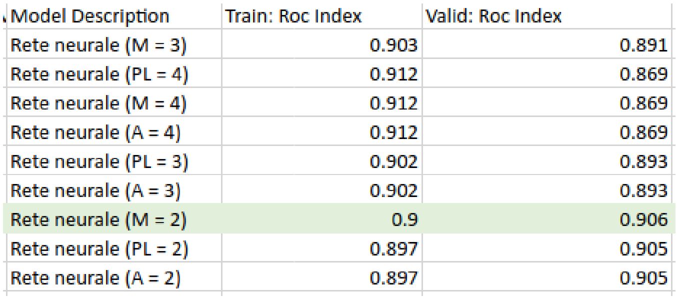
Modeling and model comparison

Two different algorithms we used for the classification task, with different selection criteria then we compared them together in order to choose the best one. The features that we are using are the original ones, plus the one engineered for making the clusters and a categorical one for each cluster.

As expected the neural network has worked better than the decision tree, because looking at the roc we can notice that the tree has lower values. Initially the trees were overfitting but then we changed the depth and now they are less likely to overfit. We also have taken a look at the other performance measures like, cumulative lift, cumulative % response and cumulative % captured response but the neural network was always doing better. Looking at the neural network roc we can see that networks with more than 2 hidden layers are overfitting, that means that they memorize the data and they will poorly perform with new data. So we decided to go for 2 hidden layers.

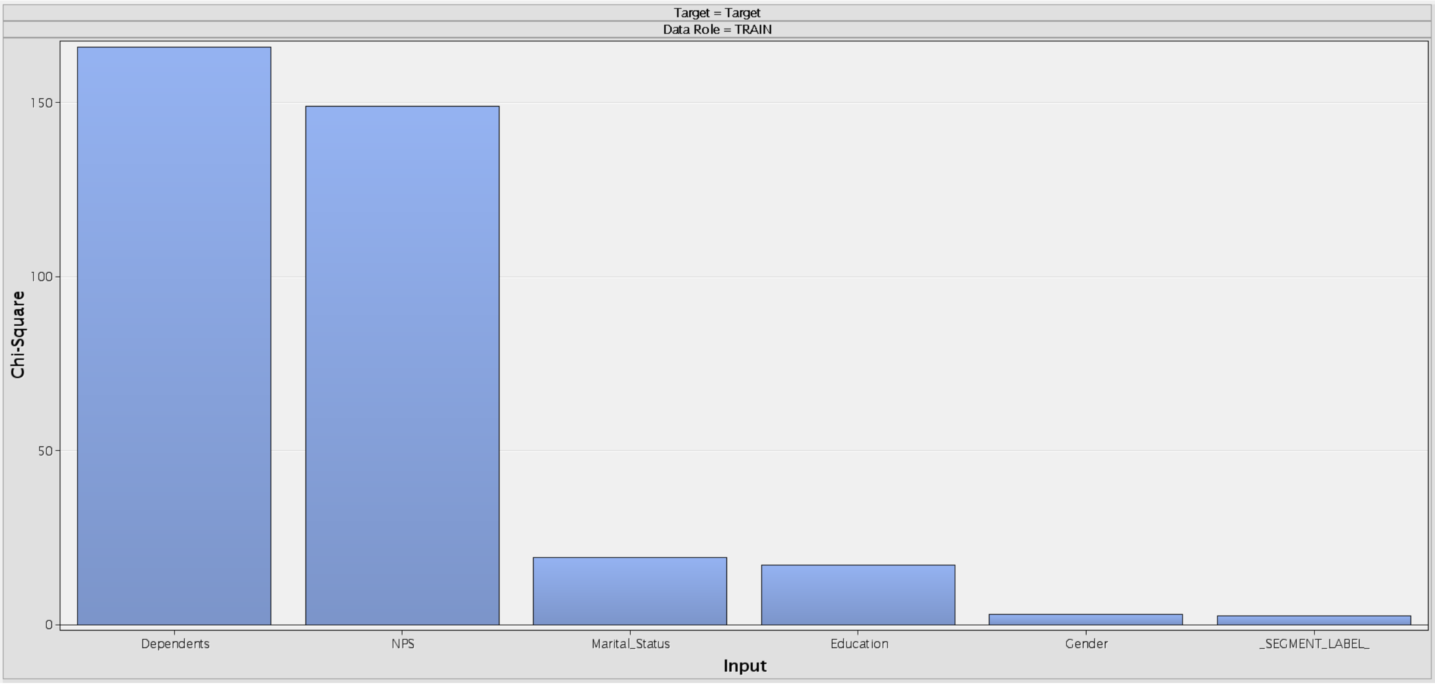


Model Description	Train: Roc Index	Valid: Roc Index
Albero decisionale (PC = 2)	0.861	0.89
Albero decisionale (EN = 4)	0.903	0.872
Albero decisionale (EN = 2)	0.857	0.892
Albero decisionale (EN = 3)	0.908	0.85
Albero decisionale (PC = 3)	0.879	0.876
Albero decisionale (PC = 4)	0.879	0.876

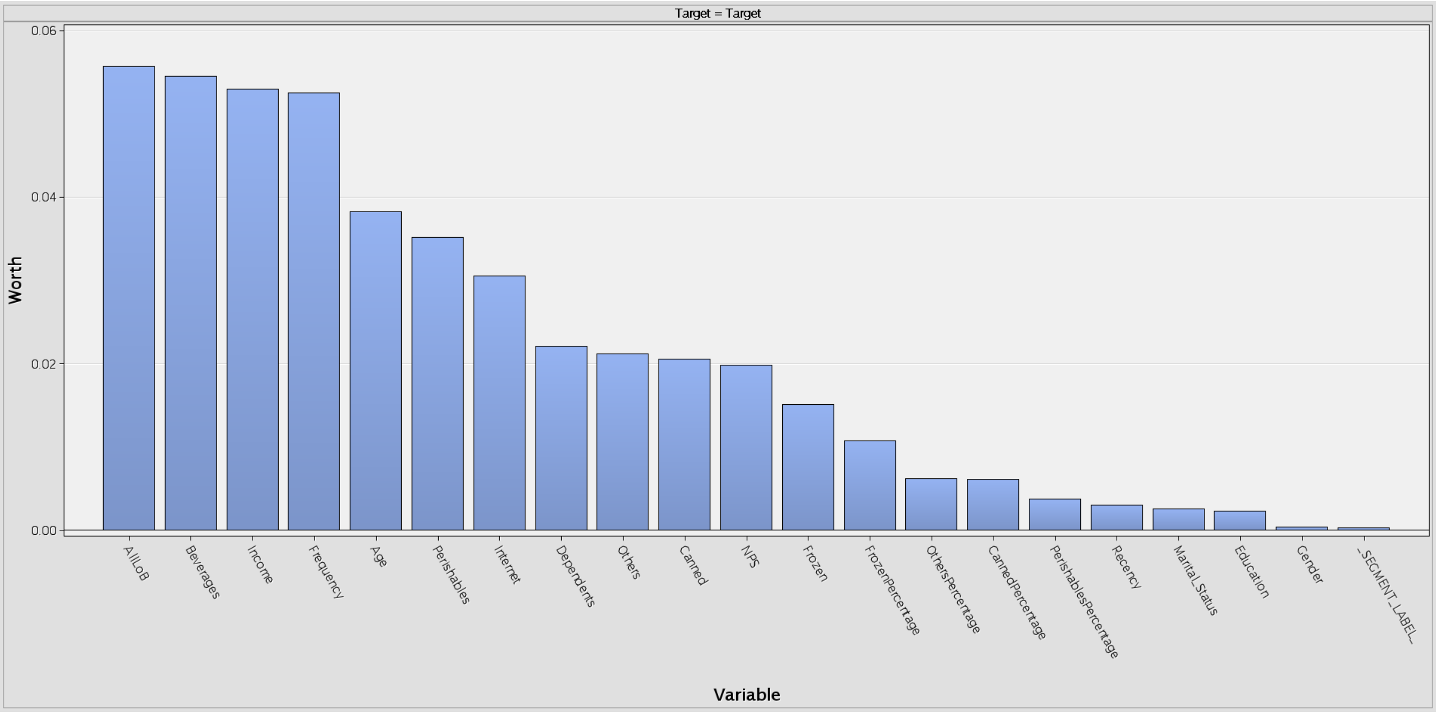


Model Description	Train: Roc Index	Valid: Roc Index
Rete neurale (M = 3)	0.903	0.891
Rete neurale (PL = 4)	0.912	0.869
Rete neurale (M = 4)	0.912	0.869
Rete neurale (A = 4)	0.912	0.869
Rete neurale (PL = 3)	0.902	0.893
Rete neurale (A = 3)	0.902	0.893
Rete neurale (M = 2)	0.9	0.906
Rete neurale (PL = 2)	0.897	0.905
Rete neurale (A = 2)	0.897	0.905

Then we performed feature reduction with the neural network, based on the worth given by the software. We haven't done this with the tree because it is able to automatically detect the importance of each variable.

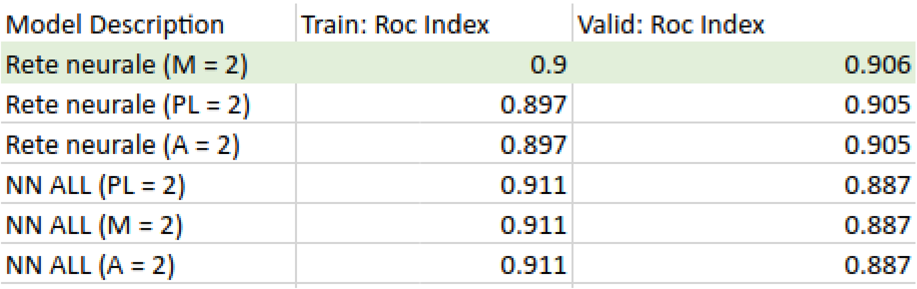


We quickly notice that in this case the segment label is not a good categorical variable, since it has a low chi-square.



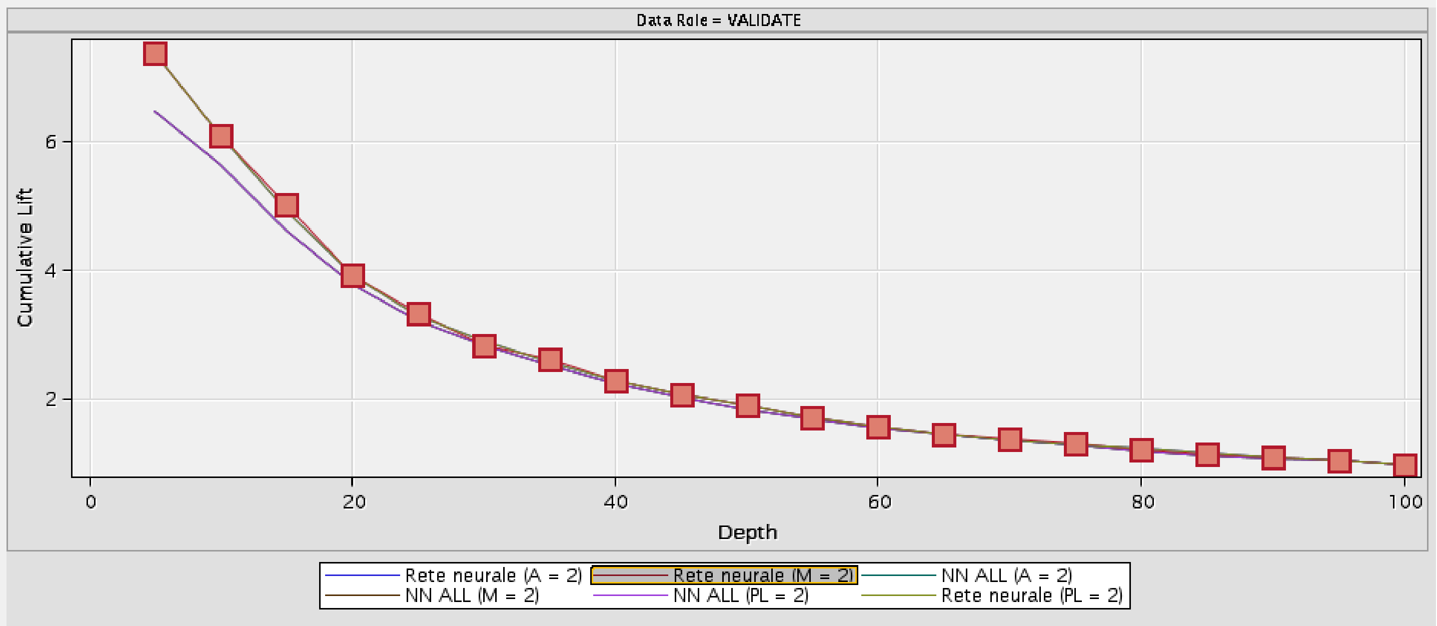
Variables as Marital_Status, Education, Gender and Recency have the smallest variable worth. We are going to remove them to find out if the model works better with feature reduction. Then we see that the new variable AllLoB is the one with the highest worth, and since it has a correlation of 81% with income we decided to remove income. We tested before and after this removal the ROC is improved, not only but income has some missing values that AllLoB doesn't have. On the table below can be seen that the model with features excluded has better performance: higher ROC index for validation and less difference between train and test.

The features that we excluded are the one with low worth, segment label and income, we tried to remove also other features but in the end this was the best feature selection performed. We can also see that the selection criteria used by the neural network were not making the difference in this case.

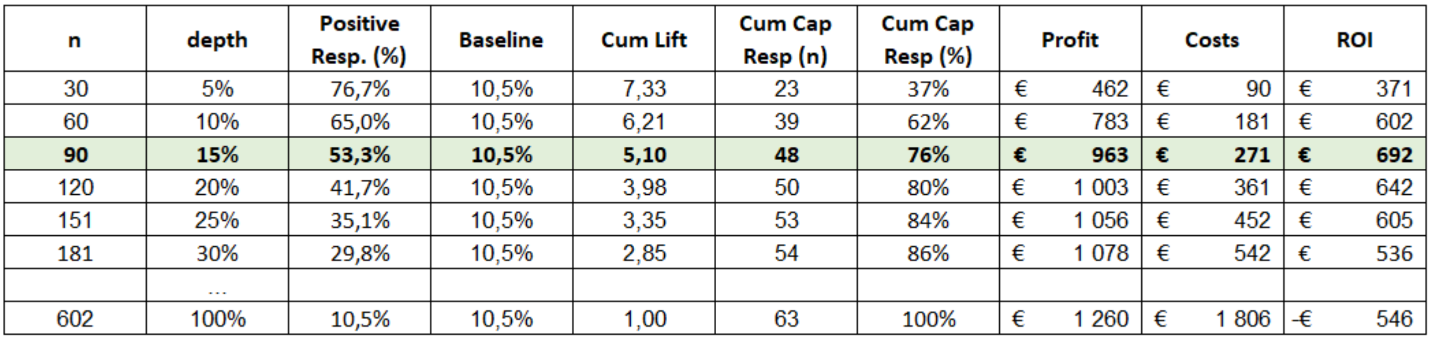


Model Description	Train: Roc Index	Valid: Roc Index
Rete neurale (M = 2)	0.9	0.906
Rete neurale (PL = 2)	0.897	0.905
Rete neurale (A = 2)	0.897	0.905
NN ALL (PL = 2)	0.911	0.887
NN ALL (M = 2)	0.911	0.887
NN ALL (A = 2)	0.911	0.887

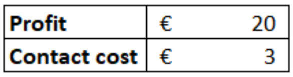
The graph shows cumulative lift and the neural network with features excluded called “Rete neurale (M=2)” shows the best performance, in this case all the lines of the neural network with excluded features are overlapping between themselves as well the ones with all the features. So there is no difference in choosing one selection criteria instead of another one. The same pattern is repeated for the other two graphs that we analyzed, that are cumulative % response and cumulative % captured response.



Cost and profit calculation



n	depth	Positive Resp. (%)	Baseline	Cum Lift	Cum Cap Resp (n)	Cum Cap Resp (%)	Profit	Costs	ROI
30	5%	76,7%	10,5%	7,33	23	37%	€ 462	€ 90	€ 371
60	10%	65,0%	10,5%	6,21	39	62%	€ 783	€ 181	€ 602
90	**15%**	**53,3%**	**10,5%**	**5,10**	**48**	**76%**	**€ 963**	**€ 271**	**€ 692**
120	20%	41,7%	10,5%	3,98	50	80%	€ 1 003	€ 361	€ 642
151	25%	35,1%	10,5%	3,35	53	84%	€ 1 056	€ 452	€ 605
181	30%	29,8%	10,5%	2,85	54	86%	€ 1 078	€ 542	€ 536
	…								
602	100%	10,5%	10,5%	1,00	63	100%	€ 1 260	€ 1 806	-€ 546



Profit	€ 20
Contact cost	€ 3

Performing ROI analysis we see that the optimal number of customers that should be contacted is top 15%. Slightly more than half of these customers, precisely 53%, are predicted to accept the subscription. In the analysis conducted we used 602 observations from the validation dataset. In this case, contacting 90 customers the company spends 271 euro and earns 963, which results in 692 euro profit, that is the maximum profit the company gets for the given contact cost and profit per customer.

78% of data is a test dataset (the “?” for the target variable), customers that haven’t been contacted yet: contacting 15% of these customers will show how accurate the prediction is.

Conclusions

Data quality

Finding out the company's earnings or losses with the analyzed customer is important, but it is very valuable to have the actual amount spent by the customer. Calculating the frequency, internet sales, and amount spent for all lines of business within the same year can enable us to create powerful new features. By utilizing this data, we can generate behavioral features that provide deeper insights compared to the existing features we currently possess.

Customer clusters

Clusterization was performed using percentage spent for each category, resulting in four clusters: Perishable Lovers, Beverage Lovers, Beverage and Canned Lovers, Frozen and Other Lovers. Beverage Lovers are the most valuable customers for the Supermarket, so the company should pay close attention to this cluster, making sure they don’t churn.

The suggestion can be made to offer cross-promotion for 3 cluster customers on beverages and canned goods, for example, buying bundles, since these are the categories the majority of customers buy.

We saw as well that customers of the cluster Frozen and Other Lovers tend to be unprofitable for the company in the beverage category, and overall are bringing least value to the company, so it is advised not to offer special promotions for these customers especially in the beverage category.

Subscription

In order to maximize profit, the top 15% of customers should be contacted, 52% of them are predicted to accept the subscription offer.