FOUNDATIONAL ASPECTS OF DATA SCIENCE

Supermart Superstore (SMS)

SAS Miner
Group Project Report

Darija Avrmoska
Viktoriya Mokhovikova
Duarte Fernandes
Vladislav Botnev
Iurii Sheiman

01/12/2023

ABSTRACT: The purpose of our work is to identify the appropriate group of ice cream buyers to whom it is best to advertise. The first stage is the consideration of the dataset and the identification of correlations, the second is the preparation of data, the third is clustering and the final stage is the analysis of the decision tree and predictive models.

KEYWORDS: Supermarket; Data; Client Profile; Ice cream.

1. Introduction

The **Supermart Superstore** is currently a big store with a seven-year history and a large number of customers. To expand the range, it was decided to supply an ice cream machine. Like any new product, this one requires advertising and promotion. It was decided to analyze customer data before launching and sending emails to save money. This document will provide a qualitative analysis of the data and identify suitable ice cream-loving customers. Because the company will only **send emails to** those who are **most likely** to make a purchase.
The result of our project will be

- the choice of the right strategies to increase profit by selling the appropriate product,
- finding potential customers,
- marketing strategies

promoting the products. Then SMS will make sure the products are **in the right hands**.

2. Methodology

This section is the first and important stage in the work, because raw data is bad for working with them due to outliers, missing values and other various factors.

2.1 Data Pre-processing

To begin with, import the file and set the desired role and level of variables.
IceCreamMachine – TARGET – BINARY, because we want to explore and predict it.

Age	**Input**	**Interval**
Canned	**Input**	**Interval**
Custid	**ID**	**Interval**
Dayswus	**Input**	**Interval**
Drinks	**Input**	**Interval**
Educ	**Input**	**Nominal**
Freq	**Input**	**Interval**
Fresh	**Input**	**Interval**
Frozen	**Input**	**Interval**
IceCreamMachin	**Target**	**Binary**
Income	**Input**	**Interval**

Kidhome	**Input**	**Binary**
LemonCookies	**Rejected**	**Binary**
Monetary	**Input**	**Interval**
NPS	**Input**	**Interval**
Pastry	**Input**	**Interval**
Perdeal	**Input**	**Interval**
Recency	**Input**	**Interval**
Teenhome	**Input**	**Binary**
WebPurchase	**Input**	**Interval**
WebVisit	**Input**	**Interval**

Figure 1 - Initial Variable Configuration

(NOTE: customer ID number (CUSTID); number od days as a customer (DAYSWUS); customers' age (AGE); academic degree (EDUC); household income (INCOME); 1 = child under 13 lives at home (KIDHOME); 1 = child 13-19 years lives at home (TEENHOME); number of purchases in the past 12 months (FREQ); number of days since last purchase (RECENCY); total sales in the past 12 months (MONETARY); % purchases bought on discount (PERDEAL); % of purchases spent on medicines (MEDICINES), skin products (SKIN), hair products (HAIR), beauty products (BEAUTY) and accessories (ACCESS); % of purchases made on the website (WEBPURCH); average visits to the website per month (WEBVISIT); adapted net promoter score (NPS))

The SAS Miner program has several tools for data visualization:

1) **Multiplot:** The tool allows you to build multiple graphs at the same time to compare models, variables, or analyze results.

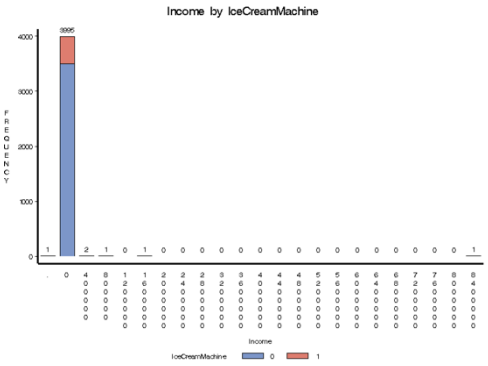

Figure 2 – Income by IceCreamMachine

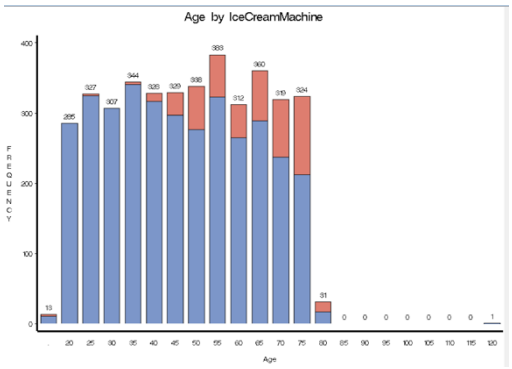

Figure 3 – Age by IceCreamMachine

As we can see in Figure 3, the distribution of ice cream purchases relative to age. People over 35-40 are more likely to buy ice cream.

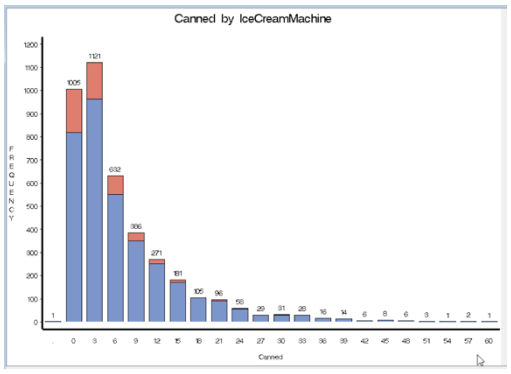

Figure 4 – Canned by IceCreamMachine

Distribution of the purchase of ice cream relative to the purchase of canned. People who buy little canned food are much more willing to buy ice cream.

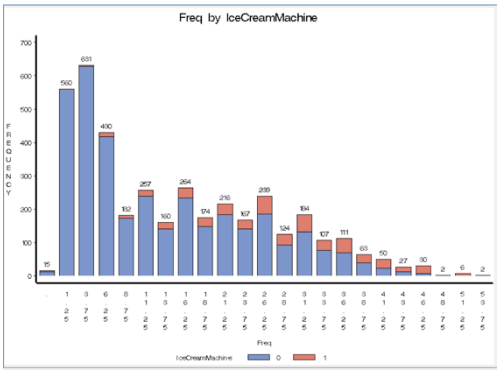

Figure 5 – Frequency by IceCreamMachine

Distribution of ice cream purchases relative to the frequency of purchases. People who shop more often tend to buy ice cream.

2) **Graph explore:** This tool allows you to visualize the relationship between features within a single object. The most frequent buyers tend to be older and, usually, are not very keen on buying products at discounts.

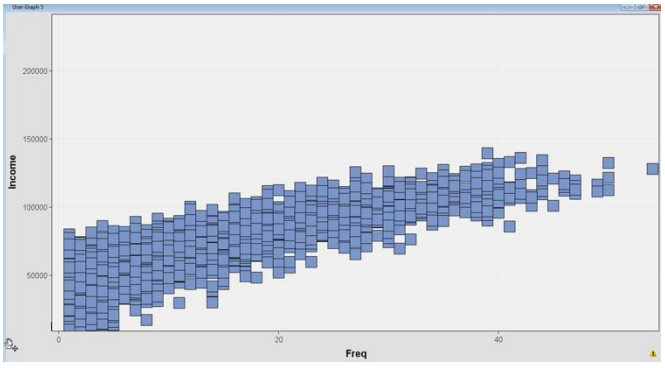

Figure 7 – Income by frequency

In this graph, a discrete function shows people's income at a certain purchase frequency.

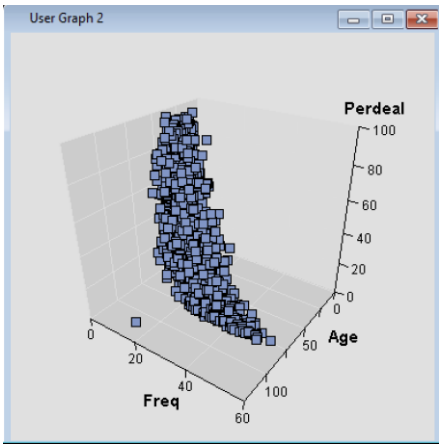

Figure 6 – 3D Scatter Plot (Age, Freq, Perdeal)

3) **StatExplore:** we see the impact of each variable on the target variable in the summary statistics. In our case, the most relevant are: **Frequency, Monetary, Age, Perdeal, Income.** NPS, Teenhome and Educ as the most irrelevant ones.

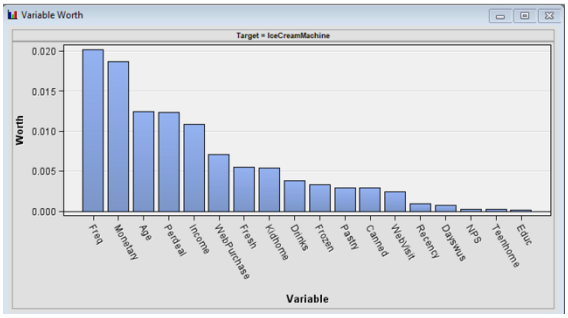

Figure 8 - Variables Worth

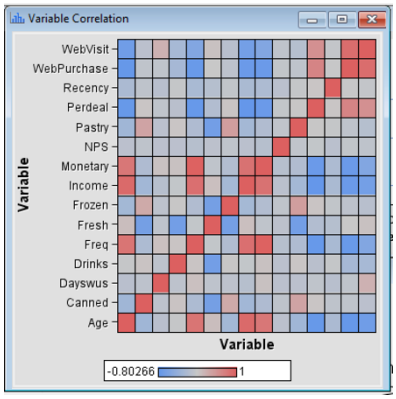

Figure 9 – Variable Correlation (Variable Clustering)

4) **Variable clustering:** We use this tool to check the correlation of variables. How can one observe Monetary is related to Income, WebVisit to WebPurchase.
5) In node **Metadata** we changed the role of IceCreamMachine to Rejected, because we want to make clusters in the future and there shouldn't be a target variable. The variable NPS was rejected due to the high number of missing values.
6) **Installing filters:** There may be values in our raw data that are very out of line. These are **outliers**. For further work with the data, we will remove them from the sample in order not to give a big impact on any one value, or greatly unbalance the results. In the **Filter node**, we go into the interval variables configure the **yellow area** for each of them and click the "Apply Filter" button. Values outside of it will be deleted. In our case, we removed 48 variables. This is (48/4001) = **1.2 %** of the total date. We will also set all the smallest values to 0, because there can be no negative numbers.

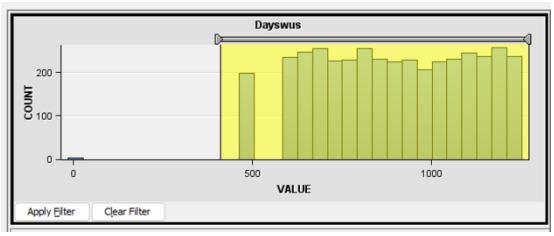

Figure 10 - Outliers in Dayswus

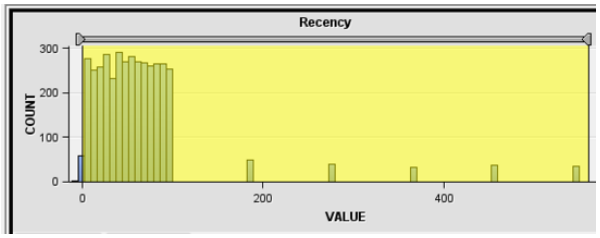

Figure 11 - Outliers in Recency

Data Role	Filtered	Excluded	DATA
TRAIN	3953	48	4001

Figure 12 - Number of excluded

	Limit	Limit
Freq	.	
Fresh	.	
Frozen	0	49.271
Income	0	25840
Monetary	0	24976.
Pastry	0	50.047
Perdeal	.	
Recency	0	559.42
WebPurchase	.	
WebVisit	0	10

Figure 13 - Boundaries

7) **Replacement**: In this part, we are doing an artificial substitution of some variables and boundaries. The boundaries stetted to compress some of the values that stand out from the rest of the dataset are reported in Fig. 10. Additionally, the Educ variable values were transformed into years: “High School” corresponds to 12 years, “BSc” to 15 years, “MSc” to 17 years and “PhD” to 20 years.

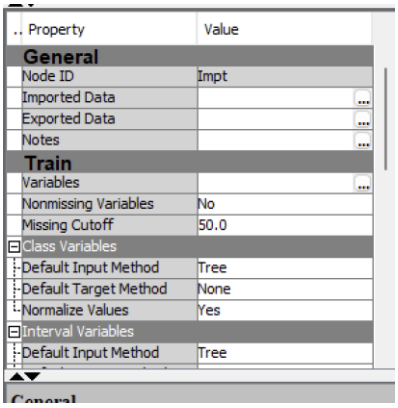

Figure 14 - Impute

Age	Default	Default	.	90	Default
Canned	Default	Default	.	.	Default
Dayswus	Default	Default	.	.	Default
Drinks	Default	Default	.	.	Default
Freq	Default	Default	.	.	Default
Fresh	Default	Default	.	.	Default
Frozen	Default	Default	.	.	Default
Income	Default	Default	.	200000	Default
Monetary	Default	Default	.	20000	Default
Pastry	Default	Default	.	.	Default
Perdeal	Default	Default	.	.	Default
Recency	Default	Default	.	.	Default
WebPurchase	Default	Default	.	.	Default
WebVisit	Default	Default	.	.	Default

Figure 13 - Upper boundaries

8) **Impute:** Imputation (to “treat” missing values). Arter the Impute, the changed variables are called IMP_PEP_Variable. The problem of missing values in the data requires separate consideration. We can put in the data that this value is skipped. And we chose a method to fill in the missing values: Tree.

9) **Transform Variables:** There are a lot of variables in the dataset. We can create other metrics from them. This helps to take a broader look at the variables and the relationship.

<u>New Variables:</u>

PART MONETERY = IMP_REP_Monetery / IMP_REP_Income

- shows how much of the income a person spends on purchases in this store

AVGPURCHCOST = IMP_REP_Monetery / IMP_REP_Freq

- shows the average purchase receipt for a person

TOTALWEBVIS = (IMP_REP_Dayswus / 30) * IMP_REP_WebVisit

- shows us how many webvisits a person makes per month an average

EduInt = IMP_REP_Educ * 1

MonCanned = IMP_Canned * IMP_REP_Monetary

- shows us how much money have been spent for canned for last year

MonDrinks = IMP_Drinks * IMP_REP_Monetary

MonFresh = IMP_Fresh * IMP_REP_Monetary

MonPastry = IMP_Pastry * IMP_REP_Monetary

MonFrozen = IMP_Frozen * IMP_REP_Monetary

After cleaning the dataset with filters, replacements, imputing missing values and creating new variables, it was once again analyzed to identify changes or new insights. TOTALWEBVIS and EduInt is useless variable for us. The most important you can see on the graph 15.

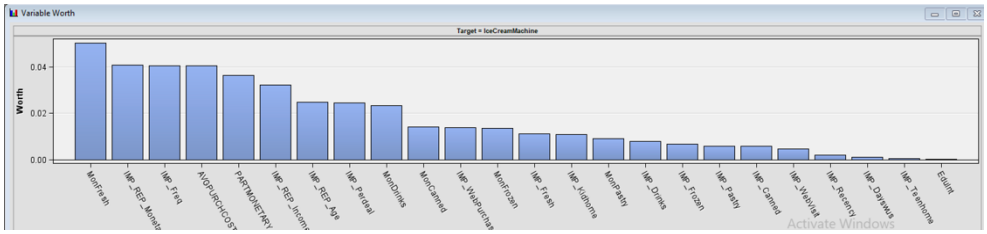

Figure 15 - Worth

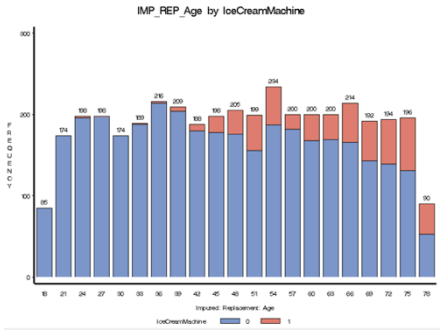

Figure 16 - Age

The distribution relative to the age and frequency of purchases remained similar.

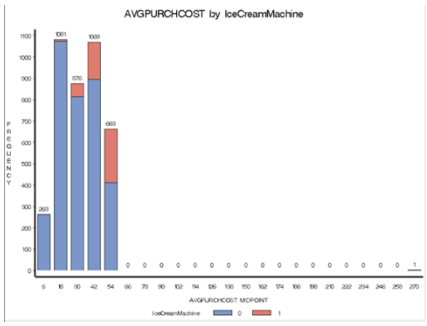

Figure 17 - AVGPURCHCOST

The new variable shows the average purchase receipt. The graph shows that people with a large average check buy ice cream.

2.2 Clustering

After acquiring filtered data and new variables we can make clusterization of the dataset to segment customers by characteristics. Formed clusters will help Supermart Superstore to identify the different groups of clients. Clustering analysis is also done to study demographic characteristics of the groups.

Creation of segmentation

While making clusters it's important to define variables that will be considered in creating segmentation. The variables our team chosen are (Custid, IMP_Pastry, IMP_Frozen, IMP_Canned, IMP_Drinks, IMP_Fresh, IMP_Perdeal) these variables will suit the best to split customers in relation to the products.

- Clusterization is done with the following properties:
- Interval Standardization – Standardization
- Seed Standardization method – Princomp

Then we had to choose the best number of clusters (find the best k), to make it we made 6 different clusterization with different k and inspected RSQ of each of them.

Number of clusters	RSQ
2	0,31756
3	0,45887
4	0,52450
5	0,56636
6	0,58689
7	0,62497

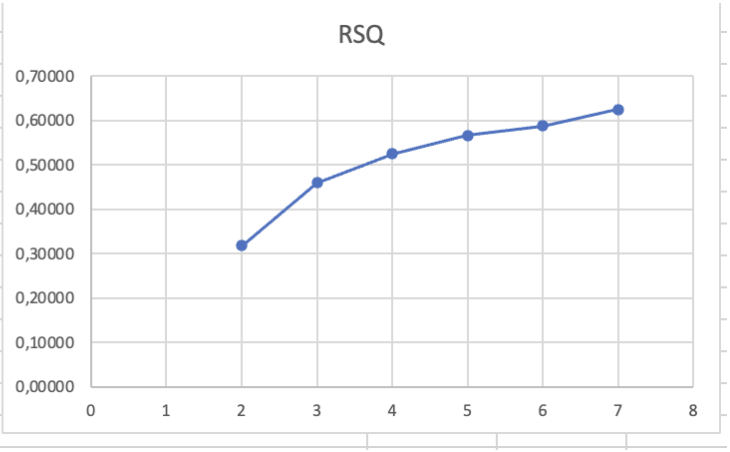

Figure 18 - Cluster Number x RSQ

Segment Id	Frequency of Cluster	Imputed: Canned	Imputed: Drinks	Imputed: Fresh	Imputed: Frozen	Imputed: Pastry	Imputed: Perdeal	Name
1.0	1249.0	2.155324259407511	16.353883106485203	77.12249799839873	2.176941553242616	2.1473178542834184	23.818254603682938	Healthy
2.0	924.0	4.7218614718614695	31.65800865800866	53.92965367965368	4.923160173160194	4.7564935064935066	61.02164502164503	Discounters
3.0	1215.0	7.368740457629347	39.270332959221854	38.117054517644526	7.666424823981436	7.601813450714719	11.6090426683993	Drinks
4.0	598.0	18.518394648829428	26.717391304347824	18.55852842809365	18.344481605351152	17.856187290969917	48.145484949832785	CFP

Figure 19 - Clusters

Clusters were named according to the dominant characteristic of the segment:

- "Healthy eaters" – The group that mainly buys fresh food, on average over 77% of their Monetary spend on fresh food. (32%)
- "Bargain hunters" – The group that mostly buys products only on sale, on average 61% of their purchases were made with discounts. (23%)
- "Drink lovers" – The group that mostly buys drinks. (30%)
- "CFP – Students" – The group that mostly buys canned, frozen, pastry food, before segment profiling, we assumed that this group is mostly students. (15%)

After analyzing segments profiles our team found following insights about groups of customers:

Healthy Eaters	Bargain Hunters	Drink Lovers	Students
Have mostly age between 44 and 59 years old (74%) Mostly have a teen 13-19 y.o. (73%)	Have mostly age between 29 and 44 years old (80%) Mostly have kid <13 y.o (82%)	Have mostly age between 59 and 74 years old (73%)	Have mostly age of 21 (60%) 43% of group have only High School education
Have slightly more than average income between 66k and 93k (74%) Spend on average on 1 purchase (26 –43) (66%) Buys products more frequent than average (11-24) purchases per year (66%)	Have slightly less than average income between 39k and 66k (78%) Spend on average on 1 purchase (9-26) (81%) Don't buy products frequently 4 (76%)	Have greater than average income between 93k and 124k (63%) Spend on average on 1 purchase (43 –112) (64%) Buys products more frequent than average (24-44) purchases per year (60%)	Have lover than average income between 12k and 39k (67%) Don't buy products frequently 4 (75%) Spend on average on 1 purchase (9-20) (69%)

		Prefer not to make website purchases (10-19) (54%)	
Prefer to buy mostly fresh food (77% of purchases spend on fresh food) and don't buy others	Prefer to buy products with discount (61%) Besides that, group buying fresh food and drinks (53% and 31%)	Prefer to buy mostly drinks (39% of purchases spend on drinks) also buys a lot of fresh food (39%). Interestingly this group don't focus on discounts and have lowest number of purchases made on discount (11%) among other groups	Prefer to buy canned, frozen, pastry food (18%, 18%, 17%). Being top 1 in these categories among other segments. This group Is top 2 in buying products on discount with (48%)

2.3 Prediction

In the Data Partition Node, the dataset was divided in two: 70.0% of the individuals constitute the Train set, and 30.0% the Validation set

1) We compared 3 different trees, each with Probchisq as the Nominal Target Criterion, Depth = 6 and a different number of maximum branches (from 2 to 4) and obtained following results:

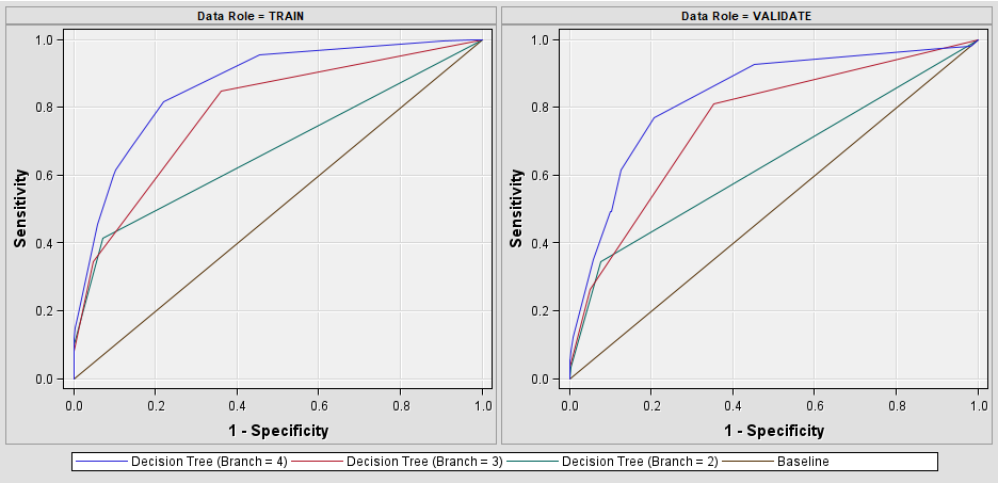

Figure 20 - ROC Chart, Model Compression Node (Decision Tree)

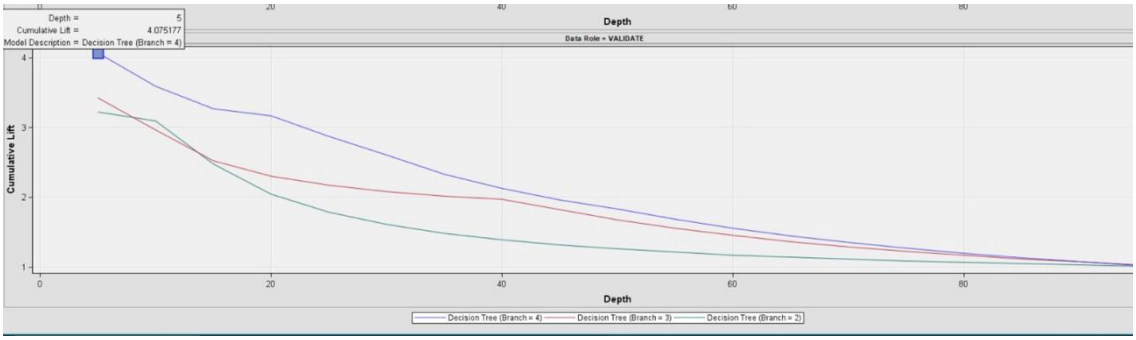

Figure 21 - Cumulative Lift Chart, Model Compression Node (Decision Tree)

The X-axis shows 1-Specificity (or False-Positive Frequency), and the Y-axis shows Sensitivity (or True-Positive frequency). Therefore, the higher the curve, the better. It was concluded that the "Decision Tree (Branch = 4)" with **4** maximum branches had the highest sensitivity overall.

Examples: If the company selects 5% of people according to the model's instructions, the results will be 4,07% times better than if you select 5% of random people from the database. (Fig. 21)

2) **Neural Networks:** We compared 4 different Neural Networks, each with Misclassification (the model that has the smallest misclassification rate) as the Model Selection Criteria and a different number of hidden units (from 2 to 5), and obtained the following results:

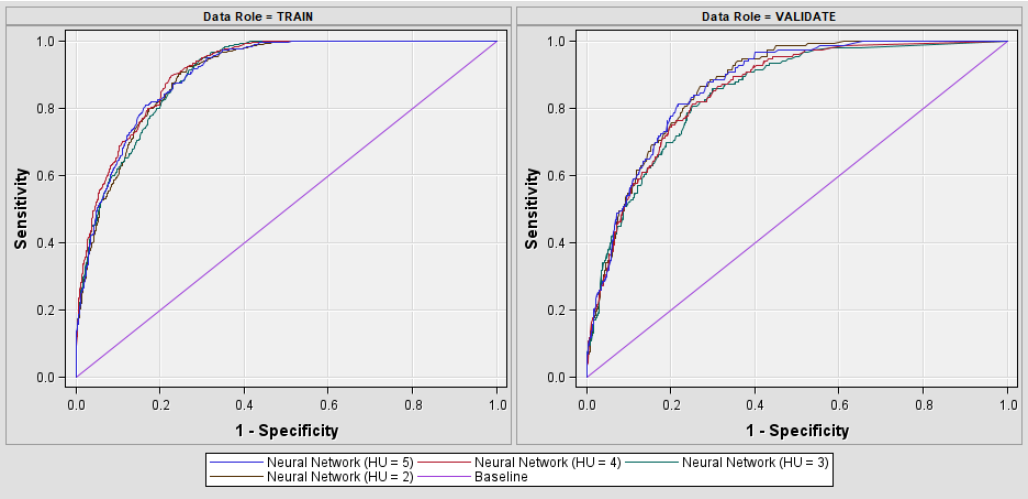

Figure 22 - ROC Chart, Model Compression Node (Neural Network)

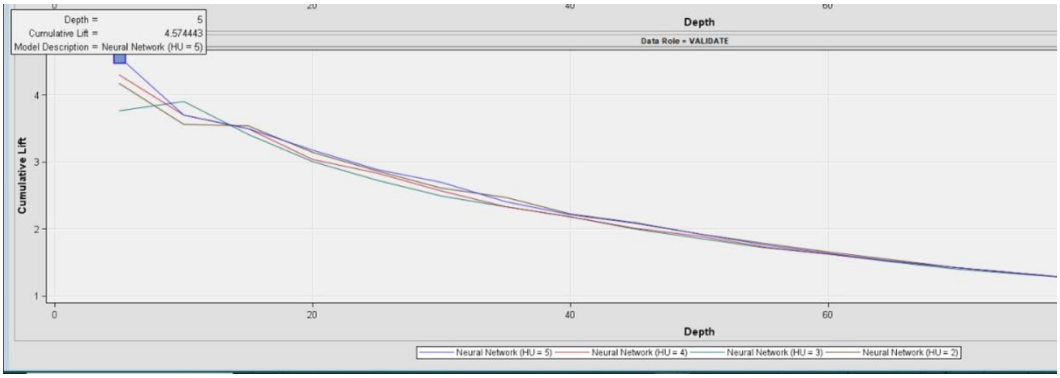

Figure 23 - Cumulative Lift Chart, Model Compression Node (Neural Network)

Although the models appear to be similar, we chose to use "Neural Network (**HU = 5)**" with 5 hidden units as it presents the highest sensitivity and consistency overall and has better Cumulative Lift. Example: If the company selects 10% of people according to the model's instructions, the results will be 3,70% times better than if you select 10% of random people from the database. (**Appendix Fig. 5**)

Final Comparison: After selecting the best decision tree and neural network, we will analyze them in Model Compression and choose the best. It is shown in **Appendix Fig. 6**

With these results we can conclude that the best prediction model is “Neural Network (HU = 5)”.

3) Conclusion

We have inspected the decision tree. The following conclusions can be drawn: people with very low incomes will not buy ice cream. People with very high incomes, on the contrary, are more likely to buy ice cream. Another factor is that, according to our research, the chance of a customer buying ice cream has a relationship with their tendency to buy fresh food: people who prefer fresh are also more likely to purchase ice cream.

If we were to contact possible customers for the Ice Cream Machine, we would send an email to 10% of the most likely to buy the ice cream from the cluster “Healthy eaters”, as it is characterized by clients that are more likely to buy fresh product. Another option would be to send an email to 20% of the most likely to buy the ice cream from clusters “Drink lovers” and "Healthy eaters”, with the split 40/60.

4) Limitations and Suggestions

It is important to mention that the sample studied is not representative of all customers. Data can be generalized, but it does not show the full picture of the customer base.

APPENDICES

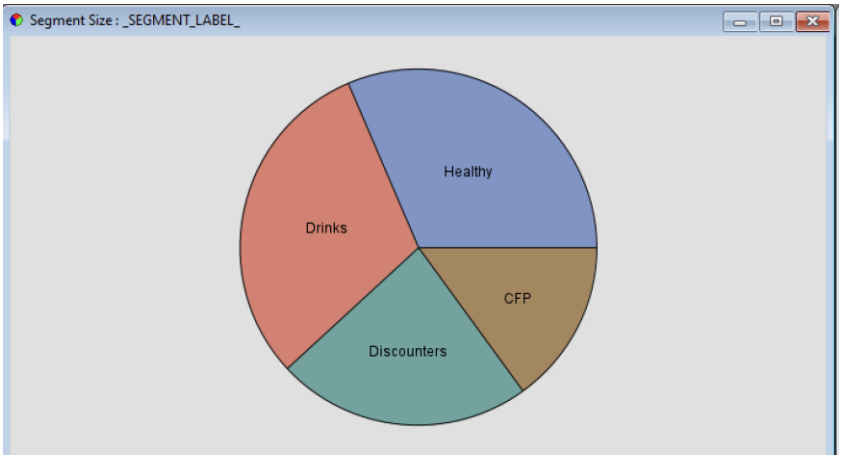

Figure 1 - Clusters

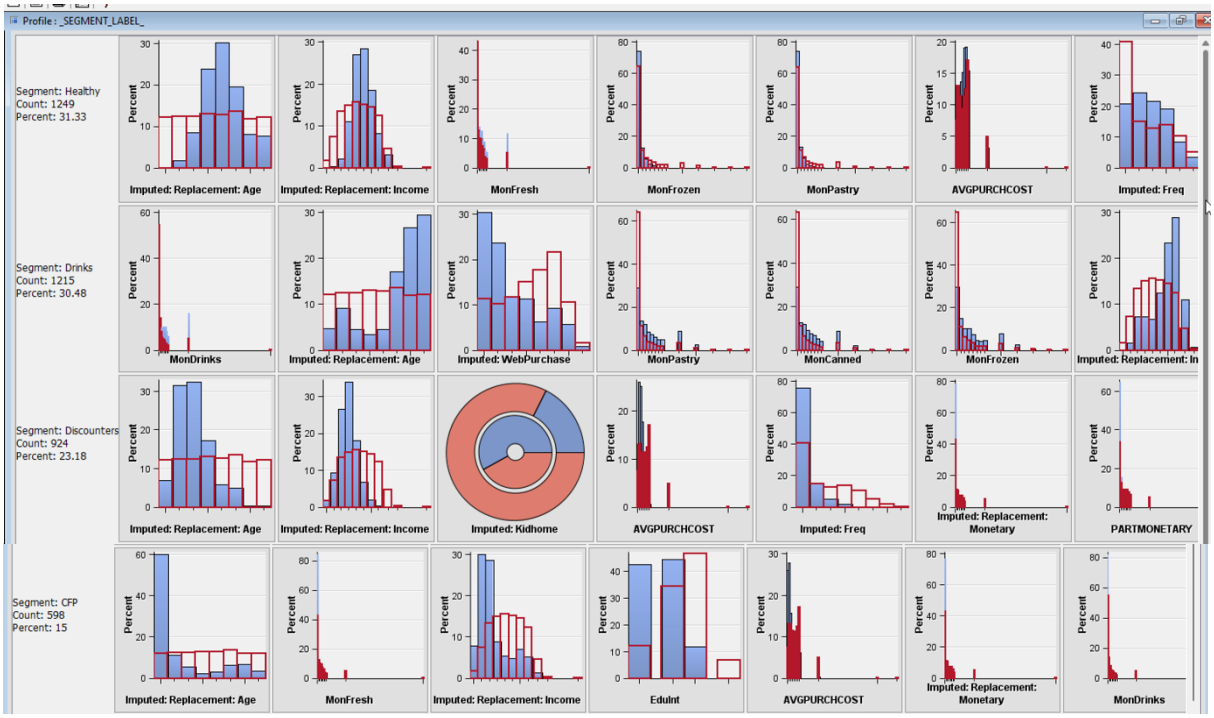

Figure 2 – Segment Profile

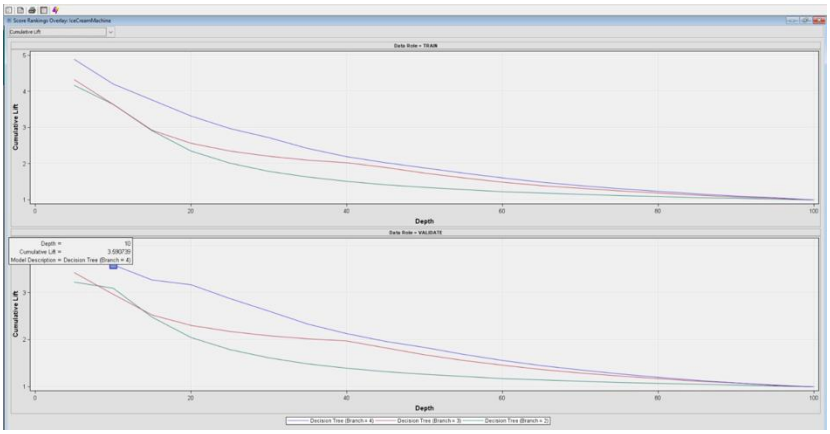

Figure 3 – DT Probchisq4 Cumulative Lift (3,59%, depth = 10)

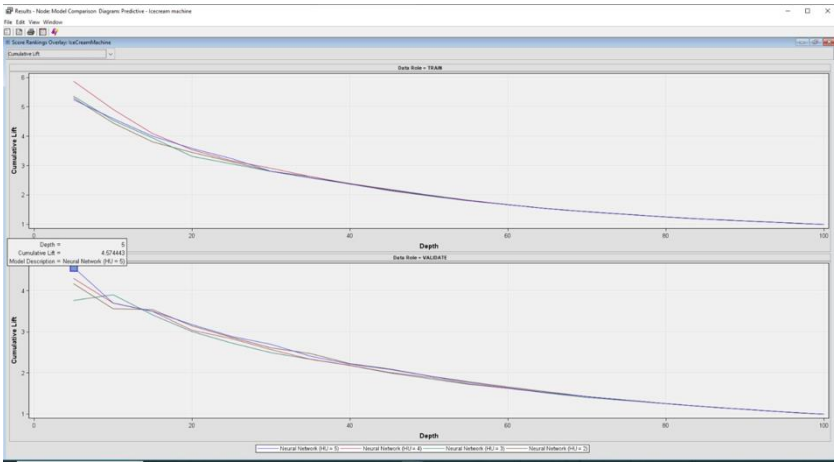

Figure 4 – NN Misclass Cumulative Lift (4.57%, depth = 5)

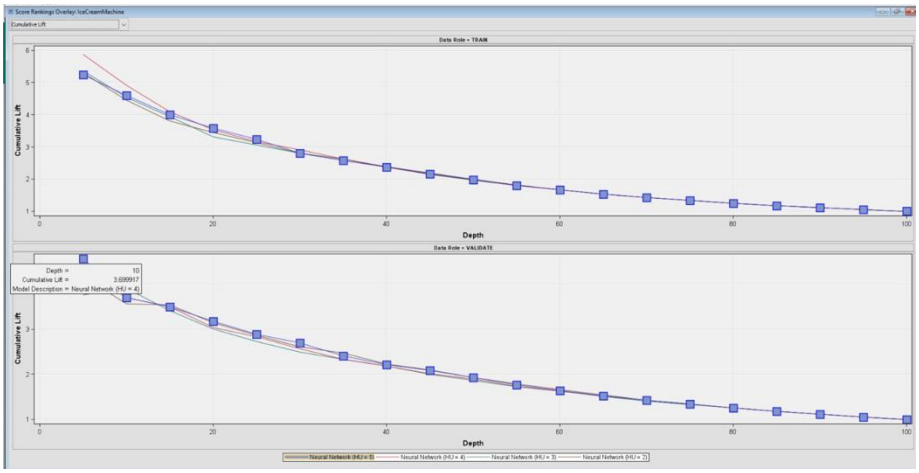

Figure 5 – NN Misclass Cumulative Lift (3,70%, depth = 10)

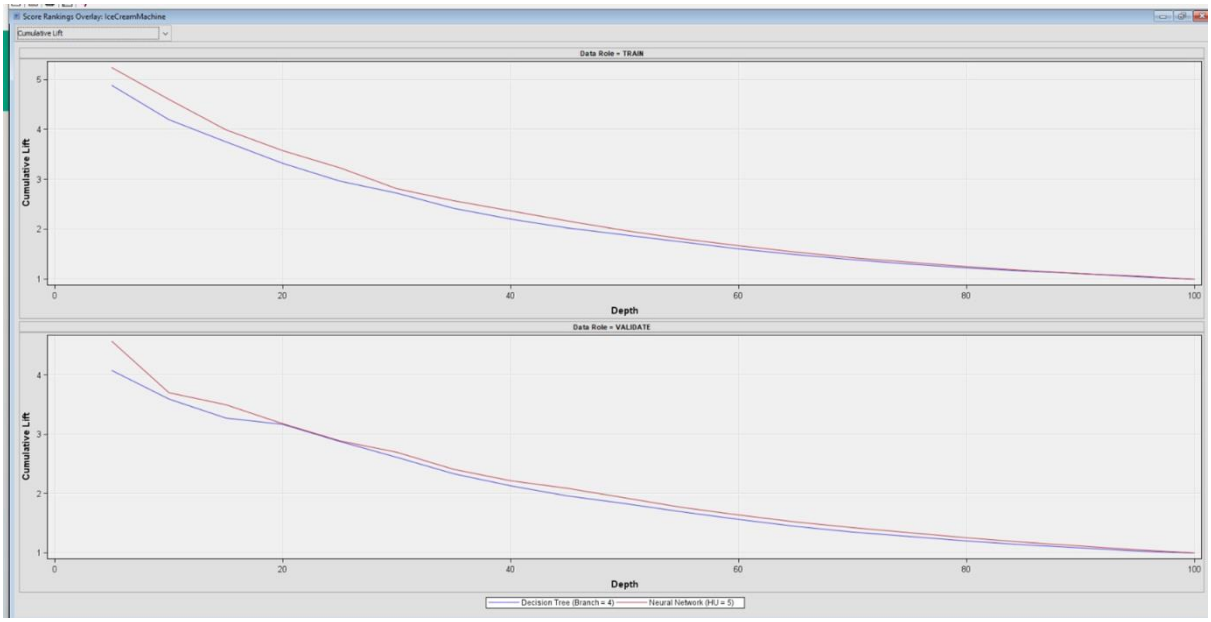

Figure 6 – Final Comparing

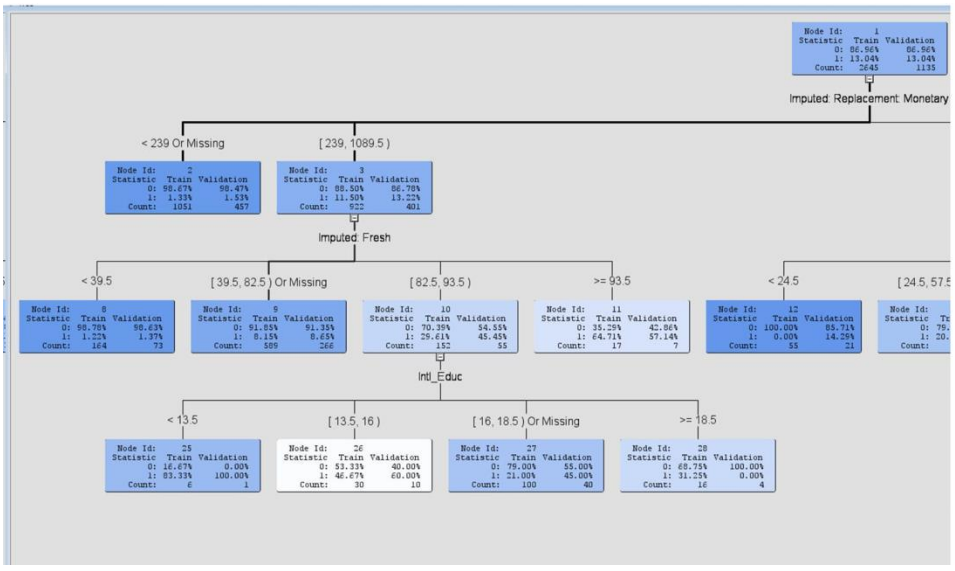

Figure 7 – Decision Tree (Attention to Validation Set)