NOVA INFORMATION MANAGEMENT SCHOOL
MARKETING ENGINEERING & ANALYTICS

NEW SUPER MARKETS

INTERNATIONAL

JUNE 2024 | FINAL REPORT

GROUP AC

BUKET DANIS	20230152
MUHAMMET EMIN IMIR	20231378
TAUSIF AHMAD	20231030
TERESA MARTINS	20221146
MARIA LEONOR SILVA	20230210

Table Of Contents

Table Of Contents

Table Of Contents

1.ABSTRACT

This paper aims to analyze data from the New Supermarkets International chain in order to recommend various marketing measures and assess the feasibility of offering free home deliveries on a monthly basis.

We had to do an exploratory and predictive analysis in order to make some inferences, which would help us better segment the customers of the supermarkets and forecast service subscriptions.

After defining the segment, we were able to offer marketing recommendations based on the features of each category and determine that 20% of the customers would be interested in subscribing to the service; these are the customers we should reach out to.

2.INTRODUCTION

New Super Markets International operates a chain of grocery stores dedicated to providing convenient and affordable options for customers across Portugal. This project is geared towards comprehensively understanding their data warehouse to enable effective customer segmentation and the development of impactful marketing strategies.

To achieve this objective, our approach commenced with rigorous data preprocessing, laying the foundation for subsequent customer segmentation and predictive modeling. Leveraging the capabilities of the SAS Enterprise Miner platform, we conducted detailed analyses to extract valuable insights into customer behavior.

These insights served as the bedrock for formulating innovative strategies aimed at both expanding the customer base and increasing the frequency of visits from existing clientele. Ultimately, this endeavor presents the culmination of our findings and recommendations derived from an exhaustive analysis of the NEW SMI customer database.

3.EXPLORATORY /DESCRIPTIVE ANALYSIS

3.1 BUSINESS UNDERSTANDING

Our exploration began with an in-depth analysis of the business data using SAS Enterprise Miner. Initially, we opted to set aside the target variable, reserving it for the predictive phase of our project. Moving forward, we meticulously examined each variable to uncover unique insights into their individual data behaviors.

During our scrutiny, we unearthed intriguing patterns within variables linked to the products sold, specifically Canned, Frozen, Beverages, Perishables, and Others (refer to Figure 1 in the Appendix). Notably, we detected negative minimum values in these variables, indicating instances of product refunds. Despite this anomaly, we made the deliberate decision to retain these values due to their significance in our analysis.

Additionally, we encountered an unusual minimum value within the Income variable (refer to Figure 1). Another noteworthy irregularity surfaced in the Internet variable (refer to Figure 1), where we observed a maximum value of 101%, a clear impossibility given the context of working with percentages.

Furthermore, it's essential to acknowledge the presence of missing values in the Recency, Frequency, Income, Education, Gender, and Marital Status variables (refer to Figure 1 and Figure 2).

Leveraging the Multiplot node's capabilities, we were able to generate a compelling visual narrative of our dataset. These insightful graphs acted as a key to unlocking a deeper comprehension of the consumer landscape.

By simultaneously depicting the interplay of various factors, the multiplot visualization unveiled patterns and trends that would have likely remained obscure in a siloed analysis. This multifaceted approach empowered us to glean nuanced insights that would have otherwise been missed.

3.EXPLORATORY /DESCRIPTIVE ANALYSIS

3.1 BUSINESS UNDERSTANDING

Upon scrutinizing the visualizations, a compelling observation emerged: all product categories – Beverages, Canned Goods, Frozen Foods, Others, and Perishables – exhibited outliers (Figures 3-7). These outliers represent data points that deviated significantly from the expected or typical patterns within each category. Delving into these outliers could yield valuable insights into specific product categories, their potential impact on consumer behavior, and the overall dataset analysis.

Demographically, the store's customer base skewed male, with a percentage nearly double that of females. Additionally, customer income distribution exhibited a semi-homogeneous pattern, indicating a relatively even spread across income levels (Figure 8). However, the presence of outliers with significantly higher incomes was notable, suggesting individuals with substantial purchasing power who could potentially influence market dynamics. We further analyzed these outliers to assess their potential contribution to understanding the overall consumer landscape and identifying opportunities for targeted marketing strategies.

Our analysis also revealed a relatively low inclination towards online purchasing among the customer base, with less than 50% opting for this mode of shopping. This observation suggests a preference for traditional retail experiences or a potential reluctance to engage in online transactions.

Examining the Recency variable, we discovered that the majority of customers were repeat patrons (Figure 9). However, we considered a duration of over two months between visits to be excessive. This finding highlights the potential need to enhance customer retention strategies and encourage more frequent visits.

On the other hand, we tried to understand the correlation between variables using Variable Clustering. If it is very red, it is close to 1. If it is very blue, it is close to -1. (Figure 10)

3.EXPLORATORY /DESCRIPTIVE ANALYSIS

3.1 BUSINESS UNDERSTANDING

Delving deeper into variable relationships, we unearthed several intriguing correlations within the dataset. Strong correlations emerged between Perishables and Frequency (Figure 11) and Beverages and Frequency (Figure 12). These findings imply that customers who shop more frequently are also more likely to spend on perishable goods and beverages. This could be attributed to factors such as larger household sizes, a preference for fresh food preparation, or simply a higher overall consumption rate. Notably, Income exhibited a remarkably strong positive correlation (>0.9) with Age (Figure 13), suggesting that older customers tend to have higher incomes. In general, income tends to increase with age. This is generally associated with an increased likelihood of earning higher income as career progresses and experience is gained. This view is compatible with the correlation here.

By examining the scatter plots depicting these correlations (Figures 11-13), we observed a clear direct proportional relationship between Income and Age, as well as between Frequency and purchases of Perishables and Beverages. This visualization reinforces the notion that as customers age and potentially experience career growth, their spending habits in these categories may also increase.

Interestingly, the scatter plots also confirmed the presence of outliers previously identified in the individual variable analysis. These data points deviate significantly from the general trends, potentially representing unique consumer segments with distinct purchasing behaviors. Analyzing these outliers in more detail could provide valuable insights into niche markets or customer profiles that require targeted marketing strategies.

3.EXPLORATORY /DESCRIPTIVE ANALYSIS

3.2 DATA PREPROCESSING

After completing the data analysis phase, the next step was addressing various issues that we encountered; including missing values, outliers and dealing with strange values we found during our business analysis.

In terms of dealing with the outliers detected, we decided to filter out values with minimal occurrences that deviate from the overall distribution: Beverages, Canned, Frozen, Income, Others and Perishables (Figures 14 to 19). This filtering resulted in the exclusion of 589 of 9000 observations (Figure 20).

Regarding the missing values, we found them in variables Recency, Frequency, Income, Education, Gender and Marital Status. Our decision was to impute using the Tree method. This imputation method is crucial for preserving our information integrity and help us handle the missing data patterns, ensuring the imputed values were derived considering the dataset of the variable itself. It was the method chosen because of its ability to handle the missing values without the need for explicit imputation, preserving data structure and reducing bias.

Dealing with strange values, our groups' approach was to replace the minimum value of Income to 9840 euros, since it is Portugals minimum wage (anually). We also put the Internet maximum percentage to 100%, since it is impossible to be higher. (Figure 21)

Additionally, we also opted to create the variable "NEW_FREQUENCY" so it would only contain positive values, ensuring its whole positivity. We used the formula SQRT x IMP Frequency^2. This change allowed us to improve the interpretability of the data and facilitate more accurate results. We also created the variable "ON_STORE" to analyse the number of purchases in store, since we already had the variable showing us the purchases that were made online. (Figures 22 and 23)

4.CUSTOMER SEGMENTATION

4.1 RSQUARED – ELBOW METHOD

Based on the Elbow Method, we chose Cluster 6 as the best option for customer segmentation. Here's why:

The Elbow Method involves plotting the R-squared values against the number of clusters (K) and identifying where the curve starts to flatten. This point indicates the optimal number of clusters, as it balances homogeneity within clusters and heterogeneity between clusters.

In our case, the R-squared values for each cluster are:

Cluster 3	**0.5015**
Cluster 4	**0.5389**
Cluster 5	**0.5533**
Cluster 6	**0.6108**
Cluster 7	**0.6103**

Cluster 8	**0.6628**
Cluster 9	**0.6720**
Cluster 10	**0.6784**
Cluster 11	**0.7097**

Table:1 RSQ Values

Upon examining the R-squared values, we notice a significant jump from Cluster 5 to Cluster 6, with an increase of 0.0575 (0.6108 - 0.5533). This indicates that Cluster 6 captures a substantial amount of variation in the data, making it a more distinct and homogeneous cluster.

In contrast, the increase in R-squared values from Cluster 6 to Cluster 7 is minimal (0.6108 - 0.6103 = 0.0005), suggesting that adding another cluster does not significantly improve the model's explanatory power. This is why we chose Cluster 6 as the optimal number of clusters.

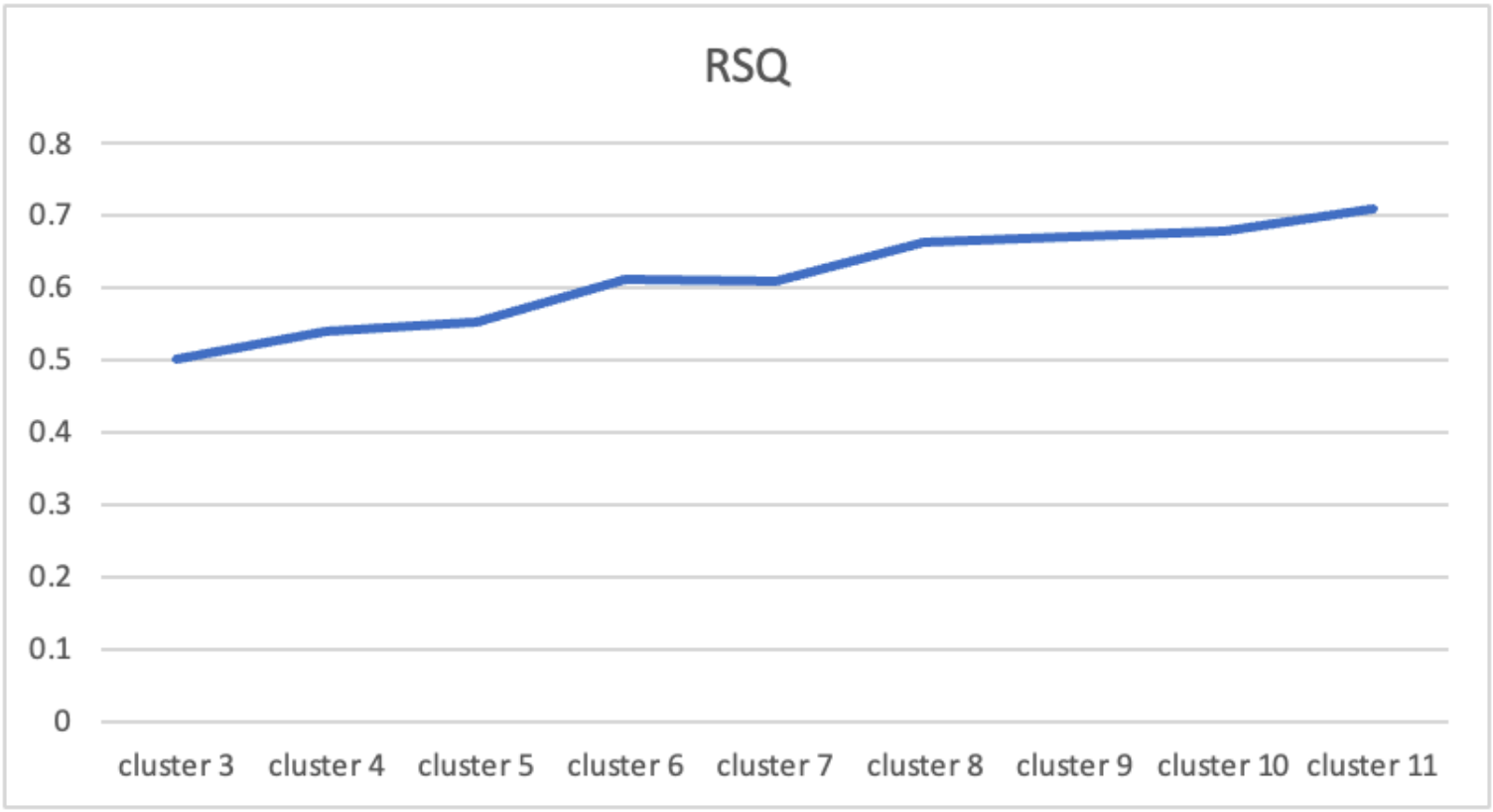
4.CUSTOMER SEGMENTATION

4.1 RSQUARED – ELBOW METHOD

Comparing Cluster 6 with Cluster 5 and 7:
Cluster 5 has an R-squared value of 0.5533, which is lower than Cluster 6's value of 0.6108. This indicates that Cluster 6 is a more distinct and homogeneous cluster, capturing more variation in the data.

Cluster 7 has an R-squared value of 0.6103, which is very close to Cluster 6's value. However, the minimal increase in R-squared value from Cluster 6 to Cluster 7 suggests that adding another cluster does not significantly improve the model's explanatory power. Therefore, we chose Cluster 6 as the optimal number of clusters.

In the context of this project, choosing Cluster 6 allows us to identify six distinct customer segments, each with unique characteristics and behaviors. This enables NEW SMI to develop targeted marketing strategies and improve customer relationships. By selecting Cluster 6, we can create more focused programs and better differentiate customers, ultimately driving business growth and revenue.



Elbow Graph

4.CUSTOMER SEGMENTATION

4.1 RSQUARED – ELBOW METHOD

Using the "Cluster" node, we changed the value of the 'Maximum number of clusters' to obtain the values of the R Square (RSQ) of each model. To perform customer segmentation, we will examine our data from two perspectives: product usage and customer value, in the same cluster. In order to ensure an unbiased analysis, we have chosen to standardize the variables and normalize their distributions. This approach eliminates any potential bias caused by the varying scales of the variables. When analyzing the graph, we pointed out that 6 clusters would be the best option to carry out campaigns that are more targeted at the customer to achieve a better conversion rate.

The elbow method is a technique used to determine the optimal number of clusters for a dataset. It involves plotting the value of a metric for different numbers of clusters. The optimal number of clusters is often considered to be the "elbow" point of the plot, where the metric begins to increase more slowly or even flattens out. The plot provided shows the R-squared value for different numbers of clusters. The R-squared value is a measure of how well the model fits the data.

A higher R-squared value indicates a better fit. The elbow point in the plot appears to be around cluster 6. This means that the model's fit improves significantly when increasing the number of clusters from 5 to 6, but the improvement is less significant when increasing the number of clusters beyond 6. Therefore, 6 clusters could be a good choice for this dataset, but it is not definitive. Further evaluation with other metrics and considerations should be taken into account.

4.CUSTOMER SEGMENTATION

4.2 MEAN STATISTICS

Segment Id	Age	Beverages	Canned	Frozen	Income	Recency	Frequency	On Store	Others	Perishables
1	**70.45**	2484.67	**2680.28**	**1829.02**	**69049.65**	49.73	**35.34**	5.87	1319.40	3318.93
2	55.45	740.00	409.90	487.96	54727.01	50.61	25.89	**11.16**	432.51	3373.19
3	35.55	200.93	140.23	212.51	32038.10	52.14	12.14	6.62	175.40	541.99
4	68.07	2035.25	1174.30	1612.66	67723.45	51.46	33.22	7.00	**2211.51**	3417.73
5	70.43	**2057.55**	748.67	810.38	66898.56	51.04	32.18	5.42	706.90	**4594.05**
6	31.63	114.17	80.98	163.90	27688.29	**311.53**	8.31	4.89	130.12	243.89

Table:2 Mean Statistics

Examining the mean statistics for Cluster 6 reveals a distinct customer segment within NEW SMI's customer base. They have the highest income value (27688.29) and the highest recency value (311.53) among all segments, which indicates that these customers are relatively wealthy and have made recent purchases. This cluster likely represents younger shoppers (average age 31.63) with a lower income (average $27,688.29) than other segments. They tend to shop less frequently (average frequency: 8.31) and spend less across all product categories, but they are still a valuable customer base because of their high income and recent purchases.

The online sales percentage (4.89%) suggests a potential preference for online shopping compared to some other segments. In essence, Cluster 6 captures budget-conscious younger customers who shop online and make smaller, less frequent purchases. The segment's low values in other aspects like beverages, canned goods, frozen foods, and perishables indicate these customers might not be frequent buyers in these categories. Their preferences lean differently from those of other products and services. To further understand this segment, we can analyze the "Others" category.

4.CUSTOMER SEGMENTATION

4.3 INPUT MEANS PLOT

The Input Means Plot shows the average values of each variable for each cluster. Cluster 6 has the highest average values for the following variables:

Replacement Internet: variable has a higher normalized mean in a certain category (represented by a 6 blue), it indicates that this category has a higher average normalized value for the internet replacement variable compared to others. which means these clusters buy more over the internet.

ON_STORE: This indicates that Cluster 6 customers are more likely to purchase items in-store. This could be due to a preference for physical shopping, the need to see and touch products before buying them, or a lack of access to online shopping.

Age: This suggests that Cluster 6 customers are older than average. This could be due to a number of factors, such as income level, lifestyle, or simply a preference for established brands.

An input means plot typically visualizes the average values (means) for each variable used in the cluster analysis. In this case, the plot likely shows the mean spending for different product categories (perishables, beverages, canned goods, frozen food, others) within Cluster 6 of your customer segmentation.

4.CUSTOMER SEGMENTATION

4.4 SEGMENT ANALYSIS

To examine this cluster in greater detail, we had to examine each of its six segments;

Segment 1, represents approximately 4.5% of our customer base. Notably, they prefer online shopping ('Internet') and products related to specific age groups (old people). As well they purchase fewer canned, beverage, and frozen items. That means they prefer to use healthy food more.

For segment 2, which represents almost more than 25% of our customer base, exhibits intriguing patterns. Customers span different life stages, with a significant portion falling within the 30–40 age range. Notably, there's a discernible difference in spending habits related to beverages. Variables like "New Frequency," "on the store," "Income," "Internet," and "Canned" all exhibit high values within this segment. Moreover, customers allocate substantial amounts to beverages and canned goods as well the customer also prefers to buy partially perishables. Also in this segment, the customer prefers to buy online more frequently.

Segment 3, representing a substantial 47.59% of our customer base, exhibits intriguing patterns. Customers span different life stages, with most falling within the 20–35 age range. Notably, they spend significantly on canned goods, perishables, beverages, internet purchases, and frozen products. The abundance and diversity of consumer products will naturally lead to frequent shopping. Naturally, customers shop frequently in this segment. As well as high-income levels within this segment, this provides an ideal context for increased consumption.

Segment 4, representing a substantial 5.9% of our customer base, stands out with unique characteristics. These customers likely fall into an age group whose average age is 45–50 and exhibit significant spending on "others." While their income is high, their preference for a wide selection of products makes them a promising target for loyalty programs. Notably, their high values for the Internet and new frequency indicate comfort with technology and openness to new experiences. In summary, Segment 4 offers strategic potential for targeted marketing efforts and enhanced customer satisfaction.

4.CUSTOMER SEGMENTATION

4.4 SEGMENT ANALYSIS

Segment 5 demonstrates robust consumer behavior, characterized by several key factors. Firstly, their high income and frequent spending at the store reflect financial stability. Secondly, the internet and new frequency values are exceptionally high, indicating comfort with technology and openness to new experiences. Additionally, the medium average spending on perishables suggests a preference for fresh and healthy food choices. Notably, this segment also includes a significant number of dependents, suggesting family-oriented lifestyles. Overall, Segment 5 represents established consumers who value quality, health, and family.

Finally **Segment 6**, the bar chart shows the distribution of recency values for a segment of customers, likely indicating the number of days since their last purchase. The segment is labeled as "Segment 6" with 329 customers, representing 3.75% of the total population. The majority of customers in this segment had a recent purchase (within the first few days), with a spike in the first bar. However, there's also a large proportion of customers with much higher recency values, indicated by the bar at the far right, which means passing the time from the last purchase in the graph there is a high population, which means after the first purchase there is too much time passed, which means they are not buying something often. It highlights the importance of understanding recency patterns to tailor retention strategies effectively.

Focusing on engaging customers who exhibit longer gaps between purchases could enhance loyalty and drive repeat business. The response cannot be improved by seeking information; therefore, web searches are not necessary.

4.CUSTOMER SEGMENTATION

4.5 MARKET SUGGESTION

We might suggest the following marketing initiatives by taking into consideration the analysis conducted for each area.

Inside **segment 1,** can be offered a wider variety of fresh produce, pre-cut and pre-washed vegetables, and pre-marinated meats to cater to their convenience needs. on the other hand can be developed online meal kit options with recipes designed for seniors, considering portion sizes, dietary restrictions, and ease of preparation. Clearly can be marked healthy choices on the online store with labels or icons to make it easier for them to find what they're looking for. As well the company can be partnered with the local farms to offer fresh, seasonal produce directly through online platform. In conclusion can be provided online resources with healthy recipe suggestions, nutritional information, and tips for senior meal planning.

Building on **segment 2** unique characteristics, we can develop a targeted marketing strategy focused on convenience and value. Since this segment exhibits high online buying frequency and a preference for canned goods, consider offerings are;

- Curated beverage and canned food selections delivered directly to customers' doors, reducing shopping trips and ensuring consistent supply.
- Bundle deals and discounts on frequently purchased canned goods and beverages.
- Implementing a loyalty program awarding points or discounts for repeat online purchases, encouraging them to make the company their primary online store.

Segment 3 presents a tremendous opportunity for growth due to their high spending power, diverse purchase habits, and frequent shopping trips. Here's a strategic approach to maximize their value;

- Leveraging their frequent store visits by implementing targeted promotions and product recommendations based on past purchases and loyalty programs. This personalized touch can enhance their shopping experience and encourage bigger basket sizes.

4.CUSTOMER SEGMENTATION

4.5 MARKET SUGGESTION

Since they purchase a variety of products thats why can be considered expanding selection of frozen meals, household goods, and personal care items. This allows them to complete most of their shopping needs in one location, saving time and effort.
Can be capitalized on their high income by offering high-quality private label brands for canned goods, frozen items, and perishables. This caters to their desire for variety and value while potentially increasing company profit margins.

Despite their smaller size, **Segment 4** offers a compelling possibility because of their high income, varied purchasing patterns ("Others"), and willingness to try new things.

Can be created a tiered loyalty program with additional advantages such as invitations to special events, early access to promotions, and personalized shopping experiences. This meets the high income needs of the clientele and honors their brand loyalty. Additionally, specific customer support channels like phone numbers or chatbots that are available via the website or app can be offered. By doing this, you may satisfy their possibly higher expectations and forge better bonds with your customers.

The business can draw in and keep this important market by emphasizing curated discovery, tailored suggestions, an exclusive loyalty program, and high-touch customer support. By satisfying their varied tastes and creating a feeling of exclusivity, this may raise their basket size and increase brand advocacy.

Segment 5 as Customized Marketing with an Emphasis on Family, Convenience, and Quality;

Segment 5 offers an excellent chance to develop devoted clients who value family, health, and quality. It is possible to implement family-friendly shopping zones with broader aisles, kid-friendly product displays, and even play areas inside the store. This makes shopping more enjoyable overall and fits well with their family-oriented lifestyles.

4.CUSTOMER SEGMENTATION

4.5 MARKET SUGGESTION

A loyalty program that gives out points or discounts for buying nutritious products like fruits, vegetables, and whole grains can be put into place. This promotes brand loyalty, encourages healthy choices, and is consistent with their health-conscious ideals.

Company can satisfy Segment 5's need for quality, convenience, and family well-being by emphasizing family-friendly shop experiences, easy access to healthy products, a "Healthy Choice" award system, and recipe ideas. By using this strategy, the business may maintain its position as the customers' go-to grocery store and increase long-term customer value.

Despite its lower size (3.75%), **Segment 6** offers a tactical chance to re-engage disengaged consumers and spark their interest in your business. Here's a focused strategy to bring these potentially important clients back;

Can be Implemented automated promotions triggered by the anniversary of their last purchase. This re-engages them and reminds them of the products they previously enjoyed.

Utilizes historical purchase information to recommend related or recently bought products. This highlights items they might find interesting and gives their purchasing experience a more personalized touch.

Customers in Segment 6 may also be asked for feedback in order to determine why their frequency of purchases has decreased. This insightful information will assist you in recognizing and resolving any possible problems or areas that require development, which will ultimately result in a more satisfying shopping experience and higher client retention.

5. PREDICTIVE MODELLING

In the second stage of our project we employed predictive modelling for NEW SMI to examine customer data and predict their inclination to subscribe to certain services. The major goal of this project is to enhance monthly subscription rates for home deliveries. We are building a model that forecasts prospects' actions in relation to subscribing to this service, as we try to target a particular group of customers. Using predictive models, we were able to pinpoint potential clients that would make our campaigns more effective, increase our business scope as well as improve customer satisfaction.

Through this research the company has been equipped with tools necessary for efficient targeting of customers, cost reduction, increase in profitability and optimization of ROI by attracting high profit Customer segments. It provided useful pointers for designing directed and efficient campaigns according to each segment of consumers. To achieve the best possible return on investment (ROI), we utilized both non-parametric (decision trees) and parametric (neural network) models in predictive modelling.

We began the analysis by preparing data identically as done in project segmentation phase till Metadata node. The Multiplot and Stat Explore nodes were used in order to make several investigations into the data following the definition of percentages for the split datasets. This was done by looking at the distribution of variables depending on the Target variable in the Multiplot node, to better understand who training dataset's customers are.

Therefore: Most number of customers is between 44 and 56 years old with highest concentration at 50 years. The majority of clients in all age brackets fall under "Target 0," though "Target 1" seems to have a stronger presence at ages: 50, 54, 58 and 64. As customers' age exceeds beyond 56, their frequency reduces but as they grow older from sixty-eight to 76 years old, "Target1" increases. (figure 26)

5. PREDICTIVE MODELLING

The data shows that the most frequent purchases happen when customers reach the 150-beverage mark. In "Target 0," there were 537 purchases, and there's also a significant but smaller number of purchases in "Target 1." We can also see peaks in purchases at 450, 750, and 1,050 beverages, with "Target 0" consistently dominating. However, as the quantity of beverages increases, the frequency drops sharply, indicating that high-volume purchases are less common. Interestingly, "Target 1" has a presence across all levels of beverage purchases, although it's minor. This suggests that most customers tend to buy beverages in smaller quantities.(figure 27)

The largest group of customers consists of those with a Bachelor's degree (BSc). In "Target 0," there were 703 customers from this group, and there's also a notable number in "Target 1." The second largest group is made up of high school graduates, with 218 customers in "Target 0" and a smaller but still significant number in "Target 1." Customers with a Master's degree (MSc) also have a substantial presence, followed by those with primary education and PhDs. (figure 28)

The most influential variables, as seen in figure 23, are "NEW_FREQUENCY," "Beverages," and "IMP_REP_Income," each with a significant impact on the target classification. Following these are "Perishables" and "Age". Variables like "REP_Internet," "Others," and "Canned" have moderate importance, and variables like "Dependents," "ON_STORE," "IMP_Marital_Status," "IMP_Recency," "IMP_Education," and "IMP_Gender" have less but still relevant influence.

5. PREDICTIVE MODELLING

5.1. DECISION TREE

The aim of this was to try out different combinations of parameters for the Decision Tree nodes. We were looking for the optimum combination that would provide us with the most reliable results when using Decision Tree model. To this end, we experimented with six Decision Tree models.

For comparing these models, Model Comparison node was used. The results that came from this node were critically analyzed. It is from here that two key measures such as specificity (X), which implies a true negative rate, and sensitivity (Y) – a true positive rate were established and defined. Our starting point was to consider ROC (Receiver Operating Characteristic) curve. The ROC chart gave us an overall evaluation on the decision tree models for predicting a particular variable. It helped us in assessing the performance of each model in detail.

In section for TRAIN dataset, several decision tree models (EN 2, EN 3, EN 4, PC 2, PC 3, PC 4) are represented by lines of different color. Plots of EN 3, EN 4 and PC 3 model have the best performance as they almost touch the top left corner of the plot. This suggests that these models are able to tell apart one target category from another with great sensitivity (true positive rate) and specificity (low false positive rate). (figure 29)

Section for VALIDATE dataset consists of the same decision tree models used for evaluating validation dataset. Again looking at the graphs above it can be seen that EN 3, EN 4 and PC 3 perform better than other models based on its values close to the upper-left edge of this graph. The curves of PC2 and EN2 also performed slightly less well than those of leading models but still go beyond the baseline indicating their effectiveness somehow.

The fact that training and validation sets show a similar trend in performance for EN3, EN4 and PC3 implies that they generalize well without overfitting to training data. (Figure 30)

5. PREDICTIVE MODELLING

5.1. DECISION TREE

In the Cumulative Lift, decision tree EN 2 has the top performance at a depth of 5, meaning that reaching out to the 5% of clients identified by this model will lead to results 8.55 times better than contacting a random 5% of clients. (figure 31)

Looking at the Cumulative Percentage Response, we can see that the same decision tree (EN 3) shows the best values. If we contact 5% of clients that the model suggests, 82,37% of them will subscribe to the service. (figure 32)

Stepping to the Cumulative Percentage Captured Response, the decision tree EN 3 is the best one. According to this graph, if we contact the top 5% the model gives us, we will have 43,30% of the total customers that will subscribe to the service. In agreement with this analysis, we came to the conclusion that the best decision tree is the EN 3. (figure 33)

Examining decision tree EN3 (figure 34) reveals that New Frequency is the variable with the most discriminatory power. For values of New Frequency below 29.5 or otherwise missing, the tree indicates a reduced likelihood of people subscribing to the service while if it's above or equal to 35.5, then there is a considerable increase in the probability of subscription.

Imputed Replacement Income is next. The higher a household's Imputed Replacement Income, the more likely its constituents are to subscribe. Indeed, if this income is $69,9767.0 and greater, there is an abrupt jump in subscription chance as per Node Id: 13 – subscribers made up for 82.19% on training data and 79.17% on validation data.

Finally, ON_STORE (proportion of purchases conducted in physical stores) makes yet another segmentation point. It's amazing how those who spend more money on in-store purchases appear less willing to sign up for this service. This contradicts earlier results and may indicate overfitting in certain areas, such as Node Id: 25 where splitting on ON_STORE between10.215 and 15.765 gives perfect classification accuracy for non-subscribers thereby suggesting possible over fitting .

5. PREDICTIVE MODELLING

5.2. NEURAL NETWORK

Neural Networks provides a model that predicts values depending on variables. With each iteration, it continuously learns and improves. In order to fix this issue, we included a Metadata node in the diagram prior to adding the Neural Network nodes. This allowed us to choose which variables would be used as input in the model. We developed nine Neural Networks, with three dedicated to each of the three Model Selection Criterion choices.

In our analysis of cumulative lift, we observed that neural network M4 performs well. However, in terms of cumulative % response and cumulative % captured response, it does not significantly outperform neural networks A4, M2, PL4, A2, and PL2, just like we see in figure 35,36 and 37.

5.3. BEST MODEL IDENTIFICATION

Having come across the best Decision Tree and Neural Network, we applied the same methodology to identify the supreme of all models. We also analyzed both options – Decision Tree and Neural Network to establish the one to choose that had more significance.To make this comparison easier, we added a third Model Comparison node to assess both models.

This will enable us evaluate various crucial metrics of Tree6 and Neural5 so as to establish which model is better than the other:

On validation set, Tree6 has much lesser misclassification rate of 0.061329 compared with 0.083475 for Neural. This implies that Tree4 has an advantage in fairly predicting what class new data belong into. Similarly on train data, Tree4 shows a misclassification rate of 0.056493 against Neural5's rate of 0.066031.

In terms of total squared errors on validation set, it is again superior over Neural5 at a value of 60.16397 versus 74.71423 (figure 38). Lower values yield improved precision in predictions made by Tree6.The average squared error over the validation set is lower for tree3 at 0.051247 as opposed to neural5's value at 0.063641

5. PREDICTIVE MODELLING

5.3. BEST MODEL IDENTIFICATION

Taking into account all of these measurements, Tree6 (the decision tree model) shows better performance compared to Neural5 (the neural network model). Tree6 consistently demonstrates reduced error rates and misclassification rates on both training and validation sets, suggesting improved predictive accuracy and generalization to the validation data.

After calculating the ROI for each set of people, we can affirm that the depth of 20% is the one we should select, which is the one with the higher return (226,77€).

6. CONCLUSION

The research is a thorough study of consumer data from the New Supermarkets International (NSMI) chain with the goal of informing marketing strategies and determining whether a free monthly home delivery service is feasible. SAS Enterprise Miner is used in the study to do predictive modeling and in-depth research.During the data preprocessing phase, issues such as odd values, missing values, and outliers were addressed. Outliers, rare values, and deviations from the dataset have been cleaned.

The Elbow Method was used to estimate the ideal number of segments for client segmentation. After looking at the R-squared values, it was found that Cluster 6 was the most effective technique to separate clients into six groups. Because every segment has distinct traits and behaviors, there are chances to create marketing tactics that are specifically aimed at them.

The features of the segments are taken into consideration while making marketing recommendations. For instance, Segment 1 suggests health-conscious goods and internet resources for senior citizens. For Segments 2, 3, and 4, strategies concentrated on value and ease of purchasing; for Segment 5, strategies prioritized family values and quality; and for Segment 6, strategies aimed at reviving lost interest.

Predictive modeling was used to forecast users' propensity to sign up for the service. Decision trees and neural networks were among the algorithms used to identify the most likely subscribers. The Decision Tree EN3 model was selected after the best models were compared, as it outperformed the other models. The model with better performance indicated that we should contact 20% of the database, which consists of 117 customers of our dataset for an ROI of 226,77€.

As a result, the report provided essential information to optimize NSMI's marketing activities and increase customer satisfaction. Analysis of the data has provided useful insights for the development of targeted marketing campaigns and subscription services.

7.APPENDIX

FIGURE 1 : SUMMARY STATISTICS - INTERVAL VARIABLES

INTRVLSTAT

Ordered Inputs	Data Role	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Abs C.V.	Coefficient of Variation	Sign
1	TRAIN	Canned	187.28	0	9000	-578.8	19656.28	500.9628	814.4369	4.515007	48.60521	INPUT	Canned	1.625743	1.625743	+
2	TRAIN	Others	234.24	0	9000	-498.72	13074.08	524.7823	782.8938	3.829273	26.35234	INPUT	Others	1.491845	1.491845	+
3	TRAIN	Frozen	269.04	0	9000	-354.04	9120.12	577.1094	803.0634	3.351195	15.75055	INPUT	Frozen	1.391527	1.391527	+
4	TRAIN	Beverages	388.96	0	9000	-190	12524.92	825.8441	996.5773	1.926191	4.999239	INPUT	Beverages	1.206738	1.206738	+
5	TRAIN	Perishables	1382.48	0	9000	-363.36	20073.34	2118.619	2165.338	1.342539	2.026574	INPUT	Perishables	1.022052	1.022052	+
6	TRAIN	Recency	53	15	8985	1	365	60.96572	58.08754	3.242609	13.20391	INPUT	Recency	0.95279	0.95279	+
7	TRAIN	Frequency	18	5	8995	1	61	20.85359	11.2156	0.625792	-0.39767	INPUT	Frequency	0.537826	0.537826	+
8	TRAIN	Income	46560	18	8982	36.2	191402	46480.14	18888.01	0.029258	-0.47074	INPUT	Income	0.406367	0.406367	+
9	TRAIN	Age	49	0	9000	19	79	48.93822	17.29591	-0.00776	-1.19593	INPUT	Age	0.353423	0.353423	+
10	TRAIN	Internet	55	0	9000	10	101	57.58433	18.80724	0.249724	-0.98151	INPUT	Internet	0.326603	0.326603	+
11	TRAIN	NPS	4	0	9000	1	5	3.433	1.017876	-0.13016	-1.05877	INPUT	NPS	0.296497	0.296497	+

FIGURE 2 : SUMMARY STATISTICS - CLASS VARIABLES

Class Variables

Data Role	Variable Name	Level	CODE	Frequency Count	Type	Percent	Level Index	Role	Label	Plot
TRAIN	Dependents	1	1	6166	N	68.51111	2	INPUT	Dependents	1
TRAIN	Dependents	0	0	2834	N	31.48889	1	INPUT	Dependents	1
TRAIN	Education	BSc	1	4430	C	49.22222	2	INPUT	Education	1
TRAIN	Education	High School	3	1505	C	16.72222	3	INPUT	Education	1
TRAIN	Education	MSc	0	1305	C	14.5	4	INPUT	Education	1
TRAIN	Education	Primary	4	1116	C	12.4	6	INPUT	Education	1
TRAIN	Education	PhD	2	594	C	6.6	5	INPUT	Education	1
TRAIN	Education		5	50	C	0.555556	1	INPUT	Education	1
TRAIN	Gender	M	1	5757	C	63.96667	3	INPUT	Gender	1
TRAIN	Gender	F	0	3203	C	35.58889	2	INPUT	Gender	1
TRAIN	Gender	Other	2	27	C	0.3	4	INPUT	Gender	1
TRAIN	Gender		3	13	C	0.144444	1	INPUT	Gender	1
TRAIN	Marital_Status	Married	2	3271	C	36.34444	3	INPUT	Marital_Status	1
TRAIN	Marital_Status	Single	1	2300	C	25.55556	4	INPUT	Marital_Status	1
TRAIN	Marital_Status	Together	0	2119	C	23.54444	5	INPUT	Marital_Status	1
TRAIN	Marital_Status	Divorced	4	679	C	7.544444	2	INPUT	Marital_Status	1
TRAIN	Marital_Status	Widow	3	447	C	4.966667	6	INPUT	Marital_Status	1
TRAIN	Marital_Status		5	184	C	2.044444	1	INPUT	Marital_Status	1

FIGURE 3 : BEVERAGES

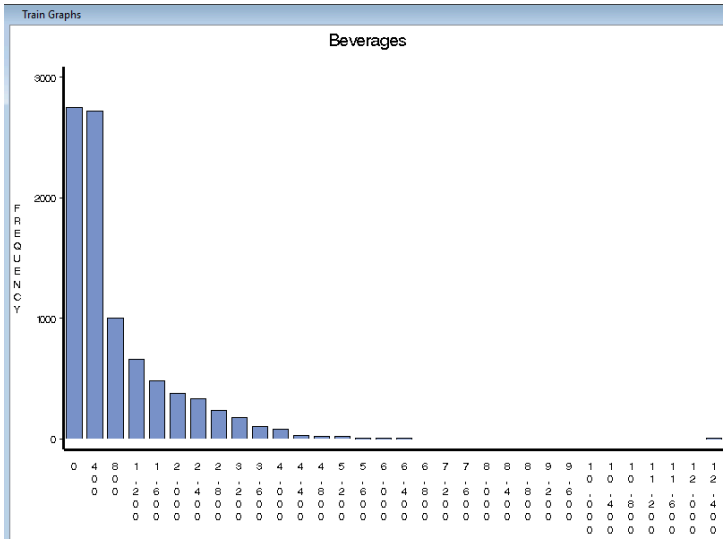
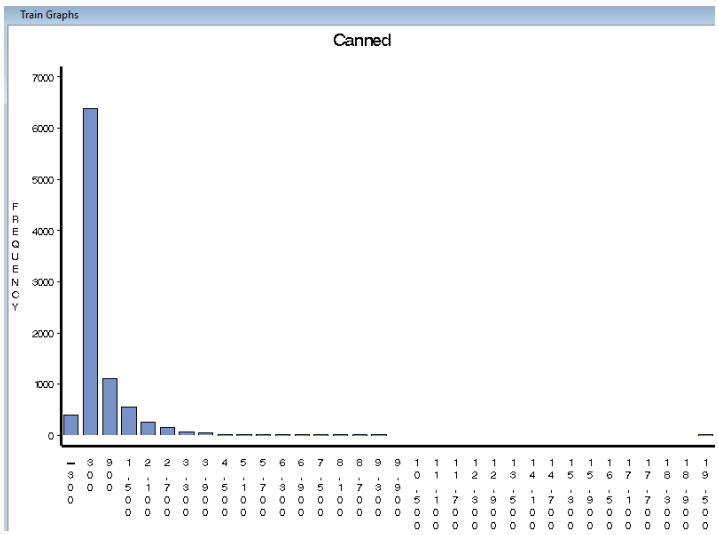


FIGURE 4 : CANNED



7.APPENDIX

FIGURE 5 : FROZEN

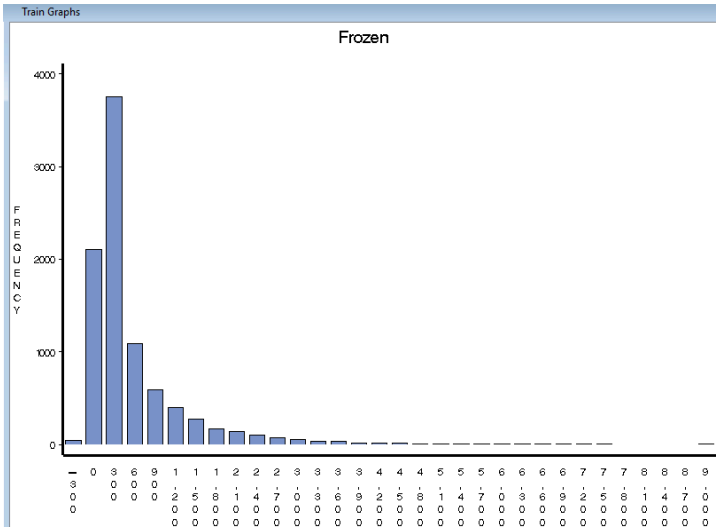


FIGURE 6 : OTHERS

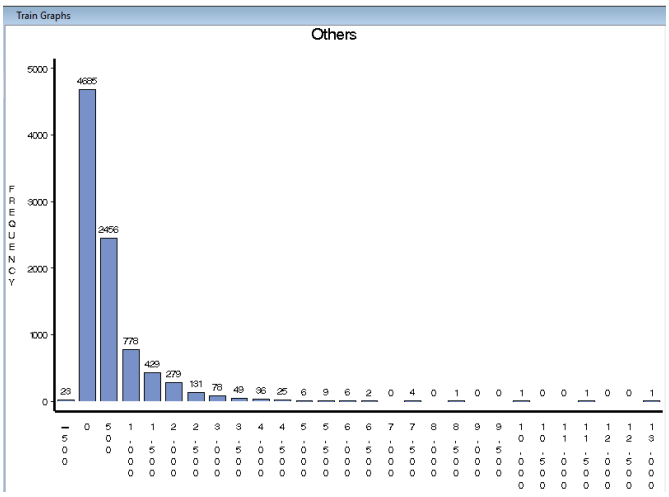
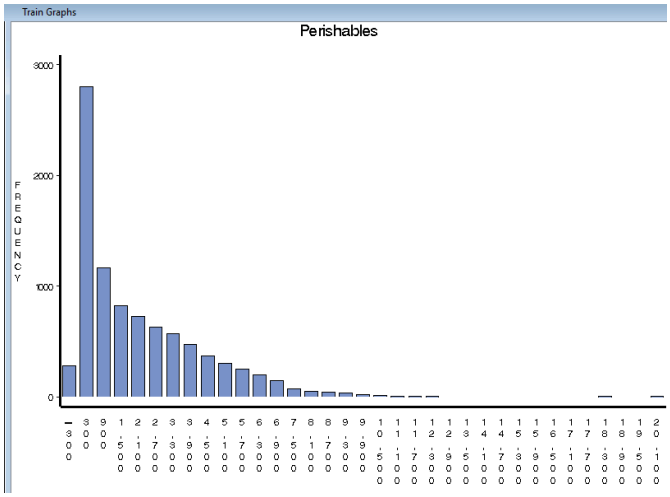


FIGURE 7 : PERSHABLES



7.APPENDIX

FIGURE 8 : INCOME

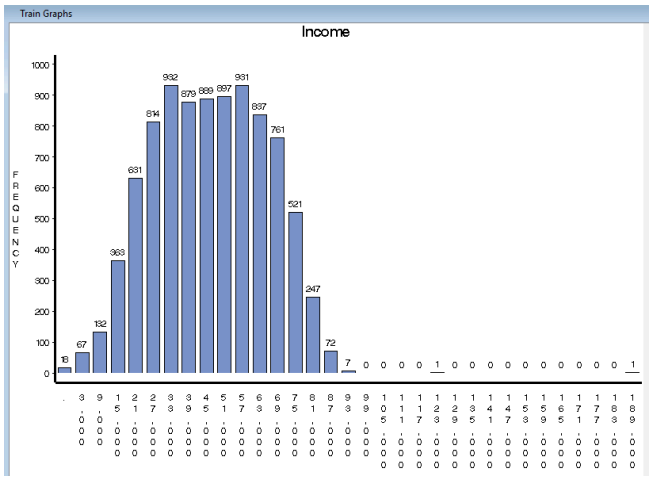


FIGURE 9 : RECENCY

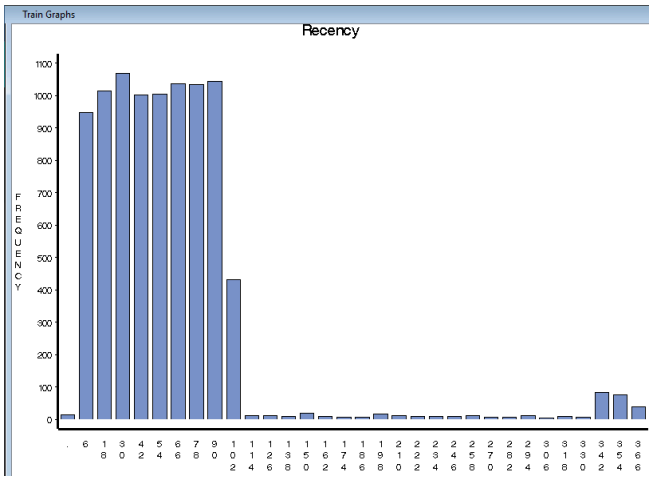
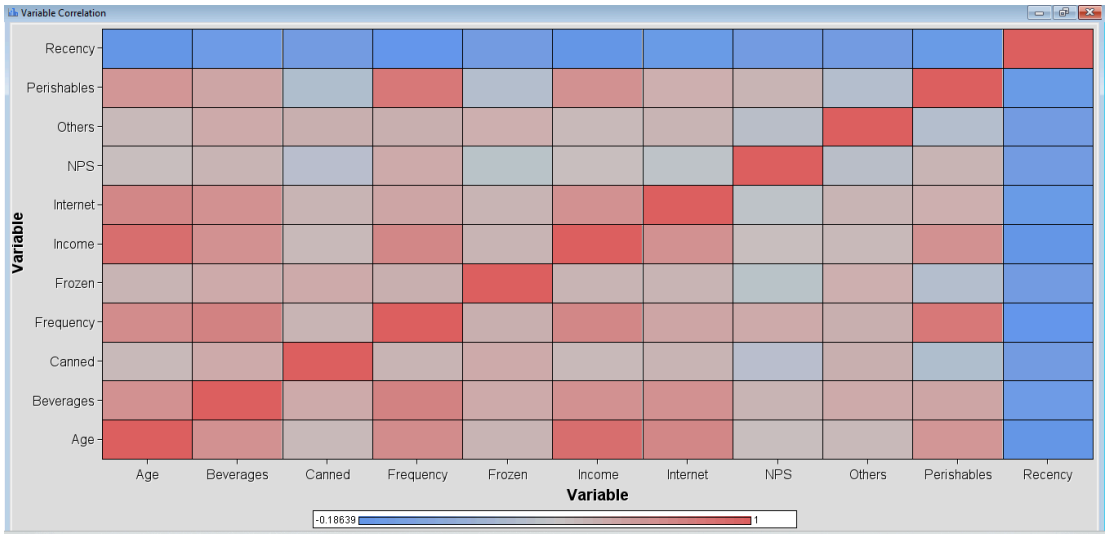


FIGURE 10 : VARIABLE CORRELATIONS



7.APPENDIX

FIGURE 11 : PERİSHABLES AND FREQUENCY

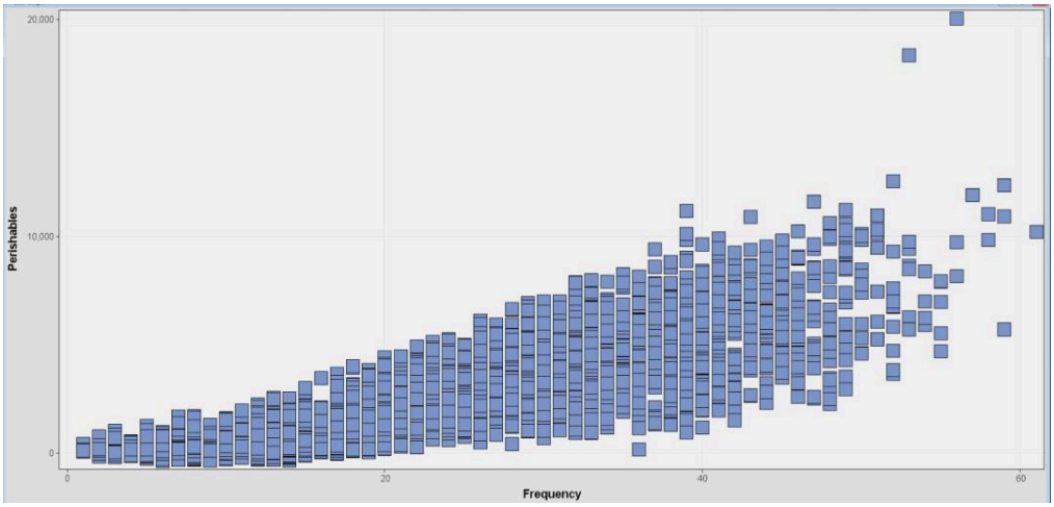


FIGURE 12 : BEVERAGES AND FREQUENCY

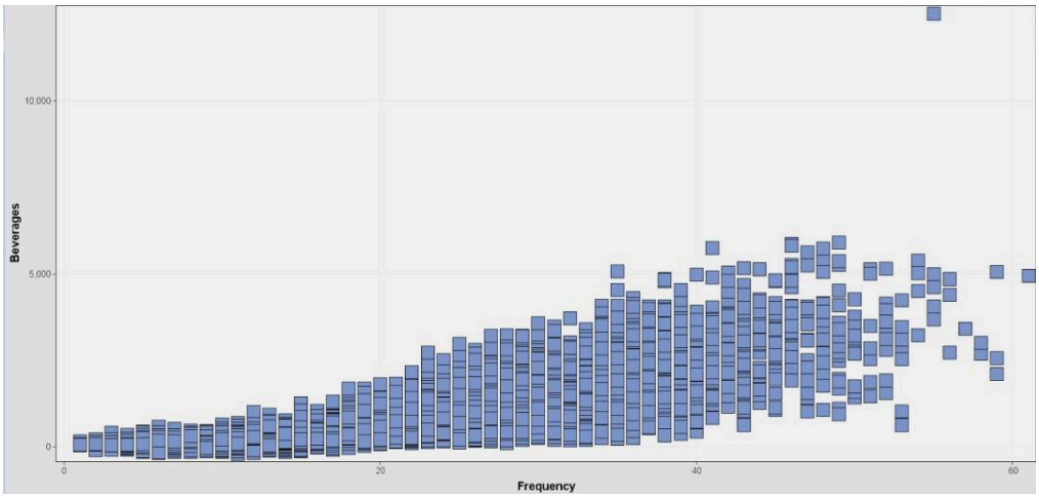
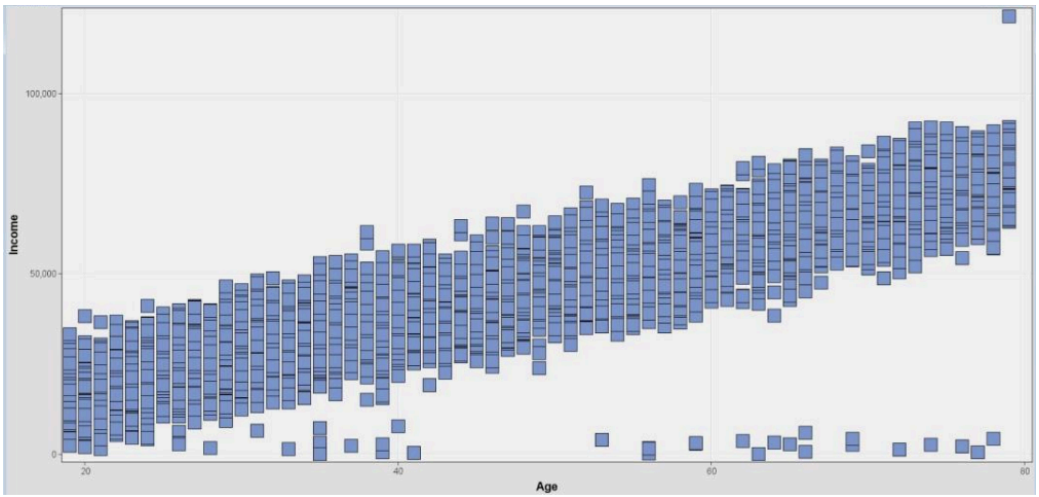
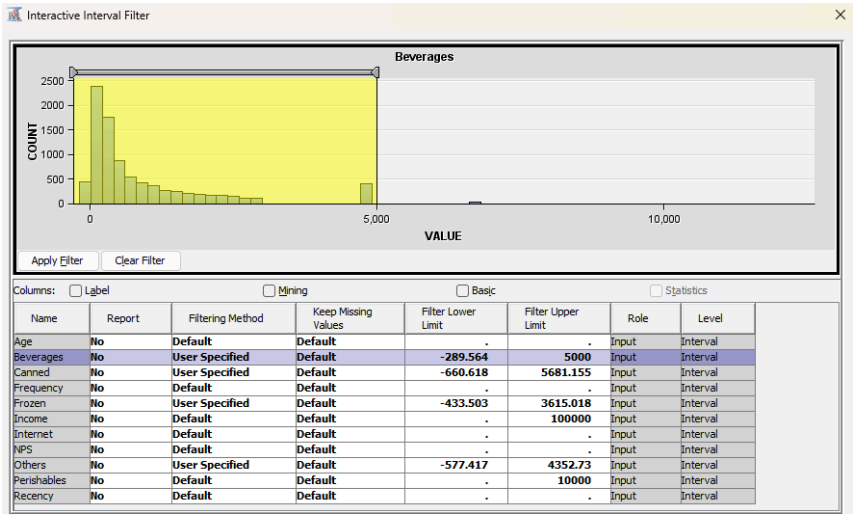


FIGURE 13 : INCOME AND AGE



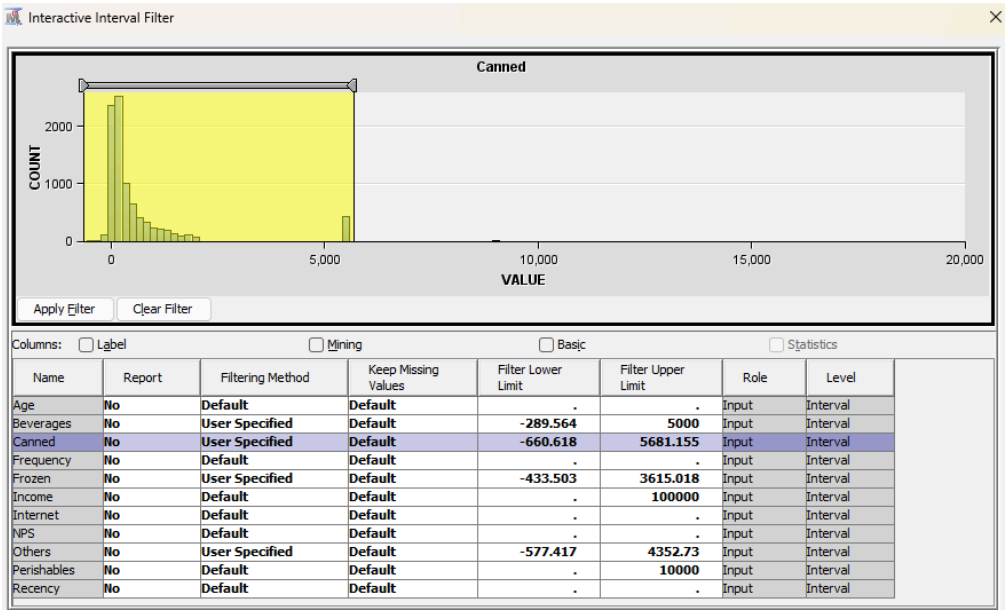
7.APPENDIX

FIGURE 14 : BEVERAGES - FILTER OUTLIER



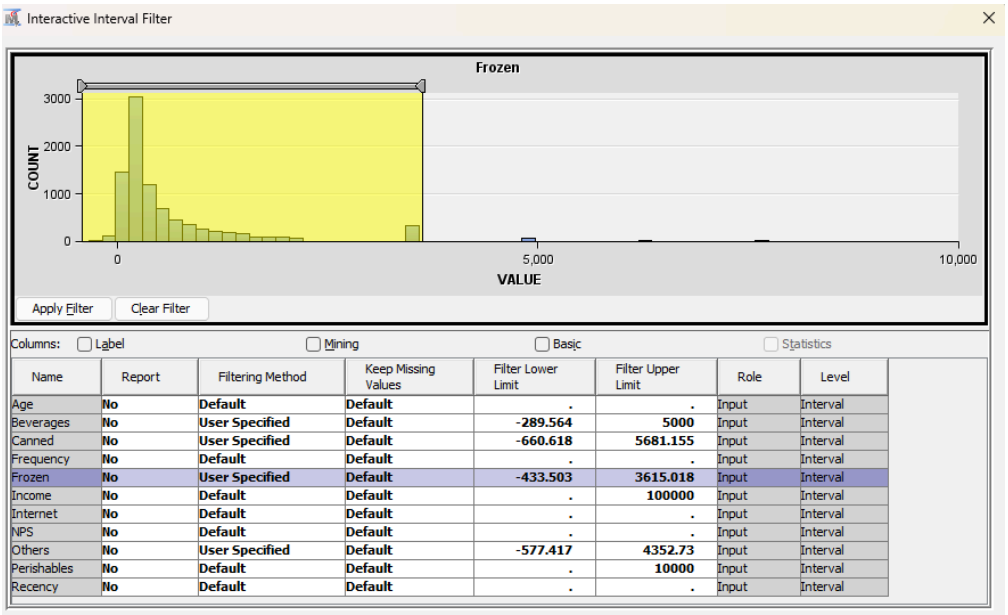
Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Role	Level
Age	No	Default	Default	.	.	Input	Interval
Beverages	No	User Specified	Default	-289.564	5000	Input	Interval
Canned	No	User Specified	Default	-660.618	5681.155	Input	Interval
Frequency	No	Default	Default	.	.	Input	Interval
Frozen	No	User Specified	Default	-433.503	3615.018	Input	Interval
Income	No	Default	Default	.	100000	Input	Interval
Internet	No	Default	Default	.	.	Input	Interval
NPS	No	Default	Default	.	.	Input	Interval
Others	No	User Specified	Default	-577.417	4352.73	Input	Interval
Perishables	No	Default	Default	.	10000	Input	Interval
Recency	No	Default	Default	.	.	Input	Interval

FIGURE 15 : CANNED - FILTER OUTLIER



Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Role	Level
Age	No	Default	Default	.	.	Input	Interval
Beverages	No	User Specified	Default	-289.564	5000	Input	Interval
Canned	No	User Specified	Default	-660.618	5681.155	Input	Interval
Frequency	No	Default	Default	.	.	Input	Interval
Frozen	No	User Specified	Default	-433.503	3615.018	Input	Interval
Income	No	Default	Default	.	100000	Input	Interval
Internet	No	Default	Default	.	.	Input	Interval
NPS	No	Default	Default	.	.	Input	Interval
Others	No	User Specified	Default	-577.417	4352.73	Input	Interval
Perishables	No	Default	Default	.	10000	Input	Interval
Recency	No	Default	Default	.	.	Input	Interval

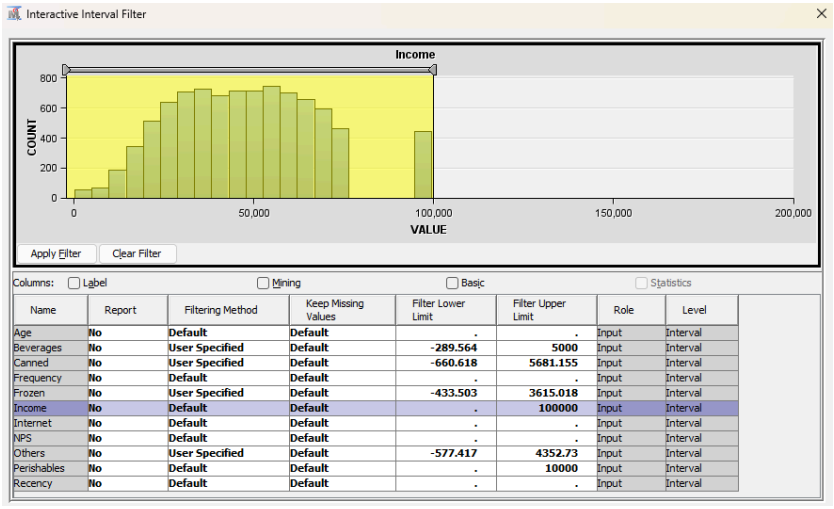
FIGURE 16: FROZEN - FILTER OUTLIER



Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Role	Level
Age	No	Default	Default	.	.	Input	Interval
Beverages	No	User Specified	Default	-289.564	5000	Input	Interval
Canned	No	User Specified	Default	-660.618	5681.155	Input	Interval
Frequency	No	Default	Default	.	.	Input	Interval
Frozen	No	User Specified	Default	-433.503	3615.018	Input	Interval
Income	No	Default	Default	.	100000	Input	Interval
Internet	No	Default	Default	.	.	Input	Interval
NPS	No	Default	Default	.	.	Input	Interval
Others	No	User Specified	Default	-577.417	4352.73	Input	Interval
Perishables	No	Default	Default	.	10000	Input	Interval
Recency	No	Default	Default	.	.	Input	Interval

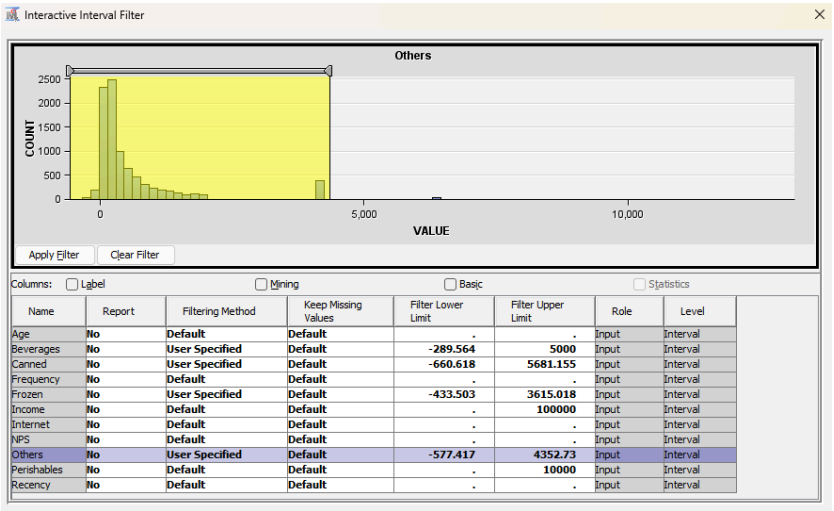
7.APPENDIX

FIGURE 17: INCOME - FILTER OUTLIER



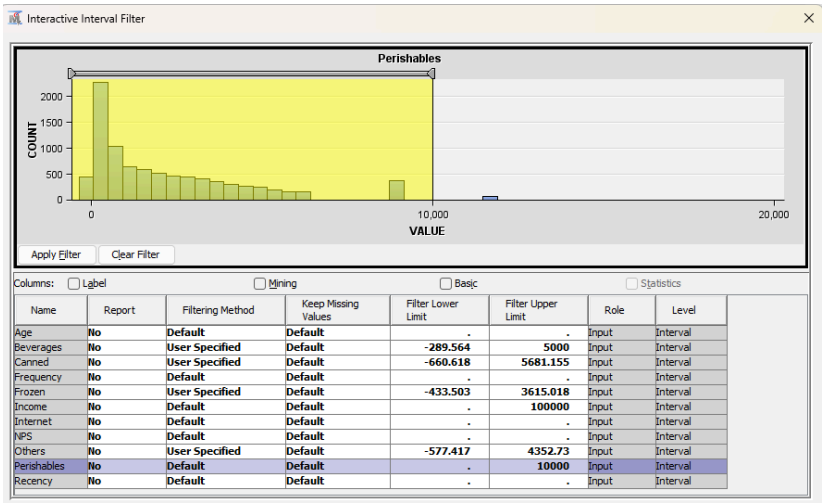
Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Role	Level
Age	No	Default	Default	.	.	Input	Interval
Beverages	No	User Specified	Default	-289.564	5000	Input	Interval
Canned	No	User Specified	Default	-660.618	5681.155	Input	Interval
Frequency	No	Default	Default	.	.	Input	Interval
Frozen	No	User Specified	Default	-433.503	3615.018	Input	Interval
Income	No	Default	Default	.	100000	Input	Interval
Internet	No	Default	Default	.	.	Input	Interval
NPS	No	Default	Default	.	.	Input	Interval
Others	No	User Specified	Default	-577.417	4352.73	Input	Interval
Perishables	No	Default	Default	.	10000	Input	Interval
Recency	No	Default	Default	.	.	Input	Interval

FIGURE 18 : OTHERS - FILTER OUTLIER



Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Role	Level
Age	No	Default	Default	.	.	Input	Interval
Beverages	No	User Specified	Default	-289.564	5000	Input	Interval
Canned	No	User Specified	Default	-660.618	5681.155	Input	Interval
Frequency	No	Default	Default	.	.	Input	Interval
Frozen	No	User Specified	Default	-433.503	3615.018	Input	Interval
Income	No	Default	Default	.	100000	Input	Interval
Internet	No	Default	Default	.	.	Input	Interval
NPS	No	Default	Default	.	.	Input	Interval
Others	No	User Specified	Default	-577.417	4352.73	Input	Interval
Perishables	No	Default	Default	.	10000	Input	Interval
Recency	No	Default	Default	.	.	Input	Interval

FIGURE 19: PERISHABLES - FILTER OUTLIER



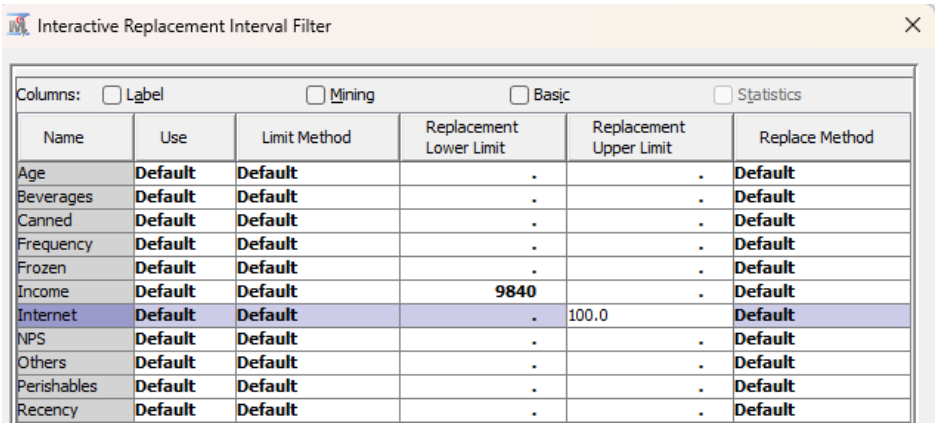
Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Role	Level
Age	No	Default	Default	.	.	Input	Interval
Beverages	No	User Specified	Default	-289.564	5000	Input	Interval
Canned	No	User Specified	Default	-660.618	5681.155	Input	Interval
Frequency	No	Default	Default	.	.	Input	Interval
Frozen	No	User Specified	Default	-433.503	3615.018	Input	Interval
Income	No	Default	Default	.	100000	Input	Interval
Internet	No	Default	Default	.	.	Input	Interval
NPS	No	Default	Default	.	.	Input	Interval
Others	No	User Specified	Default	-577.417	4352.73	Input	Interval
Perishables	No	Default	Default	.	10000	Input	Interval
Recency	No	Default	Default	.	.	Input	Interval

FIGURE 20 : NUMBER OF EXCLUDED OBSERVATIONS

```
41
42
43    Number Of Observations
44
45    Data
46    Role     Filtered    Excluded    DATA
47
48    TRAIN       8767         233       9000
49
50
```

7.APPENDIX

FIGURE 21 : REPLACEMENT OF VARIABLES INTERNET AND INCOME



Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit	Replace Method
Age	Default	Default	.	.	Default
Beverages	Default	Default	.	.	Default
Canned	Default	Default	.	.	Default
Frequency	Default	Default	.	.	Default
Frozen	Default	Default	.	.	Default
Income	Default	Default	9840	.	Default
Internet	Default	Default	.	100.0	Default
NPS	Default	Default	.	.	Default
Others	Default	Default	.	.	Default
Perishables	Default	Default	.	.	Default
Recency	Default	Default	.	.	Default

FIGURE 22 : CREATING A NEW VARIABLE: NEW_FREQUENCY

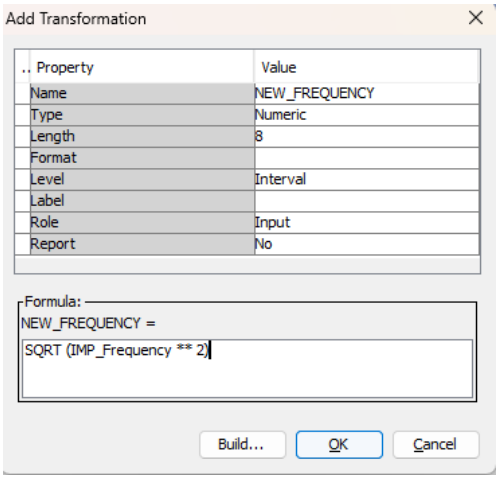
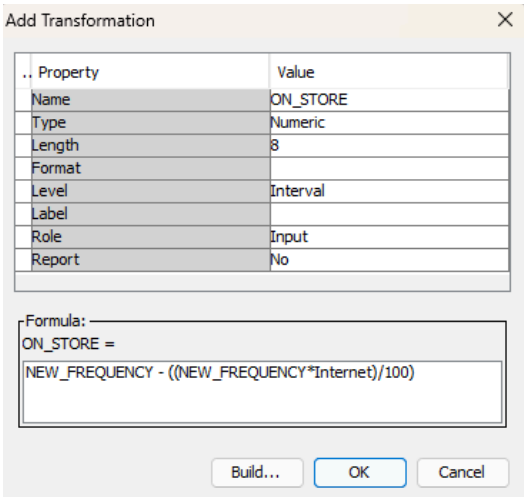


FIGURE 23 : CREATING A NEW VARIABLE: ON_STORE



7.APPENDIX

FIGURE 24 : INPUT MEAN'S PLOT

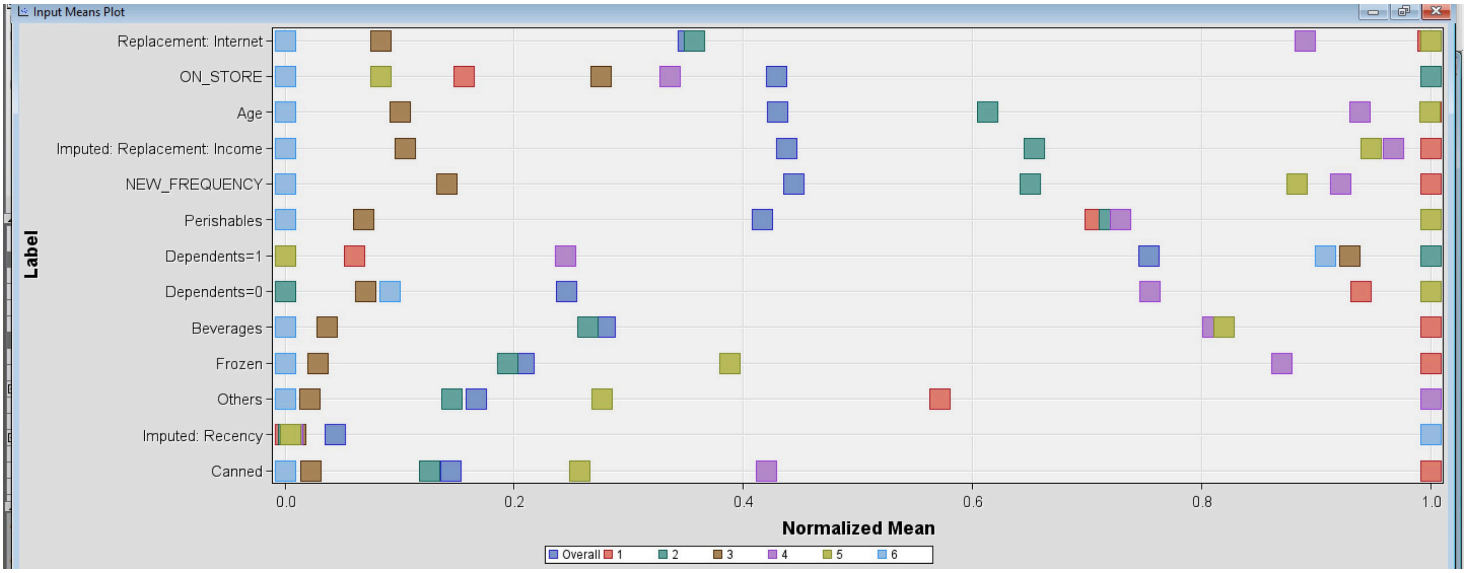
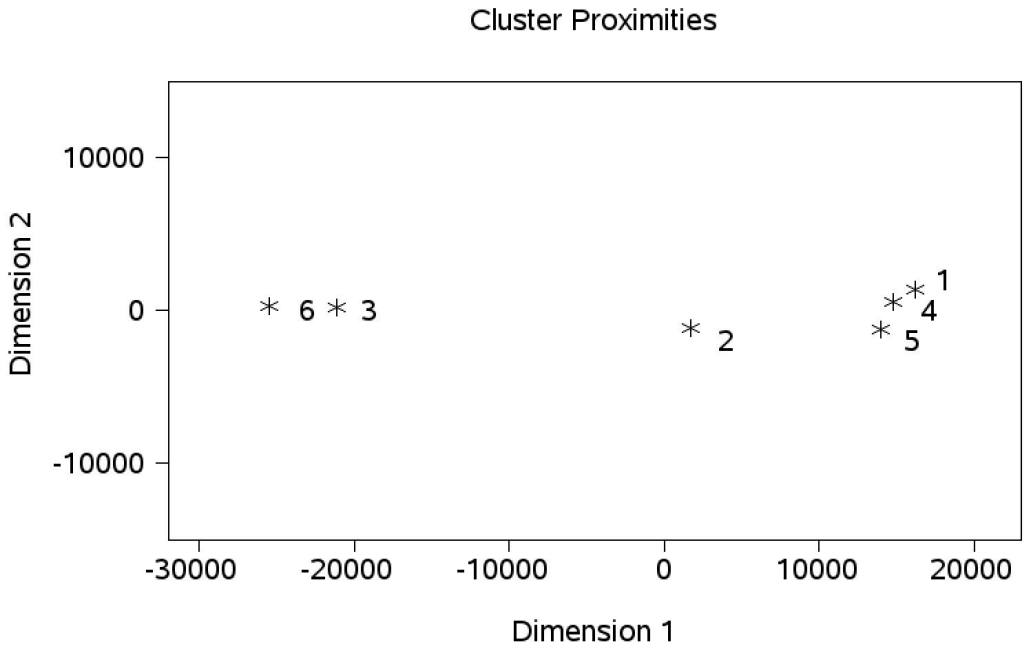


FIGURE 25 : CLUSTER PROXIMITIES



7.APPENDIX

FIGURE 26 : AGE BY TARGET

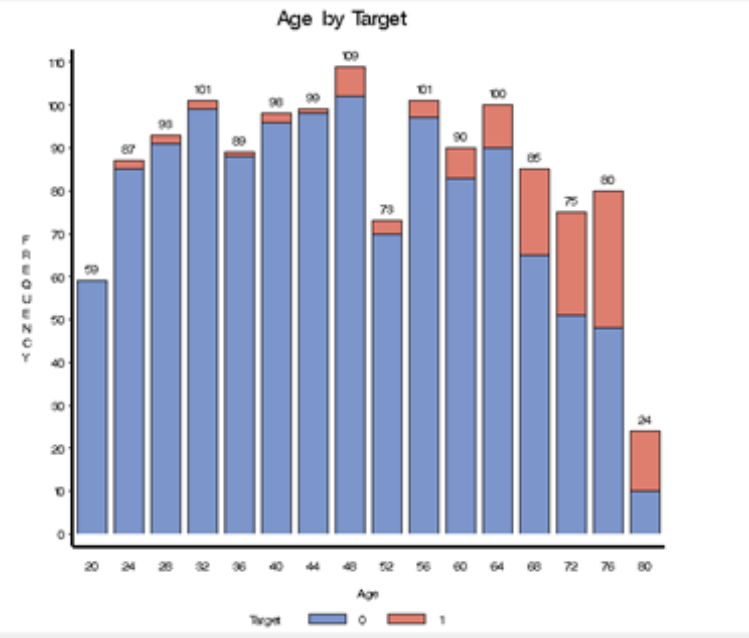


FIGURE 27 : BEVERAGES BY TARGET

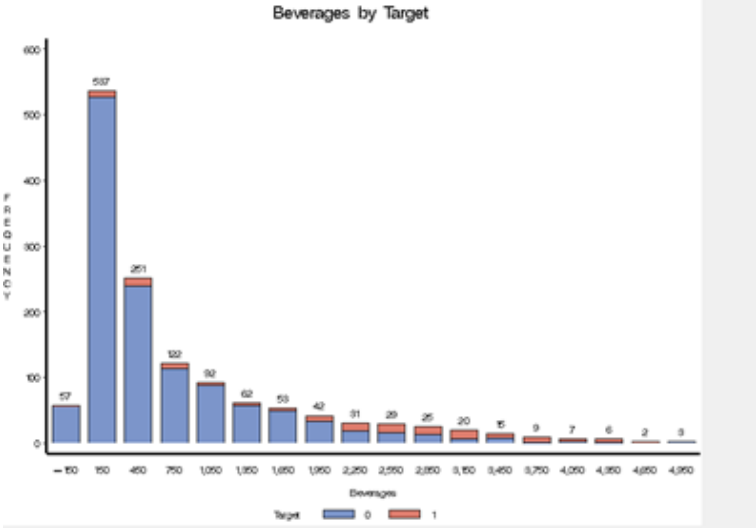
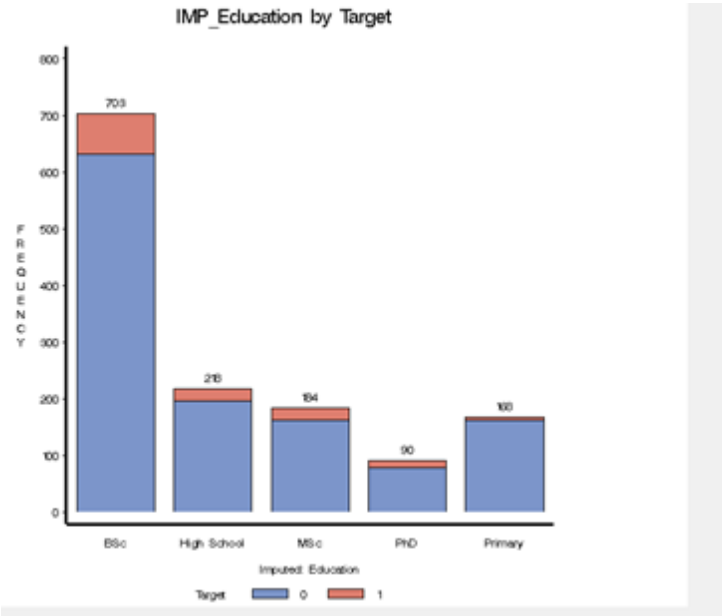


FIGURE 28 : EDUCATİON BY TARGET



7.APPENDIX

FIGURE 29 : MOST INFLUENCIAL VARIABLES

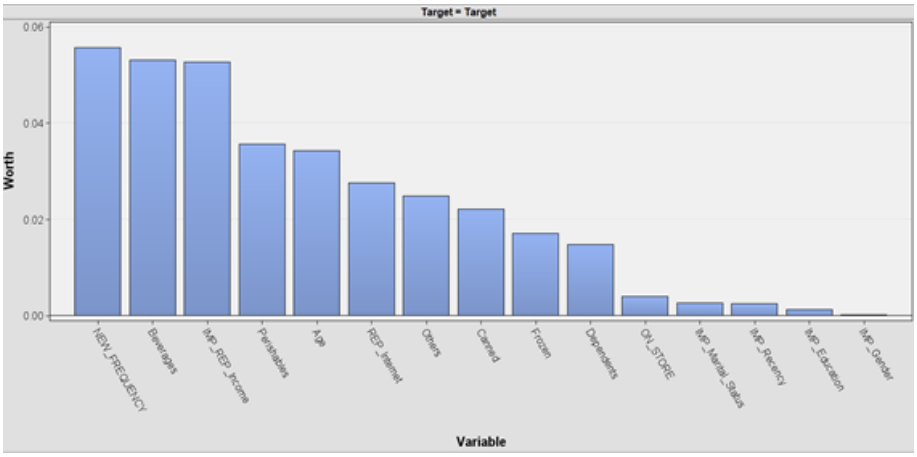


FIGURE 30 : DECISION TREE: ROC CHART

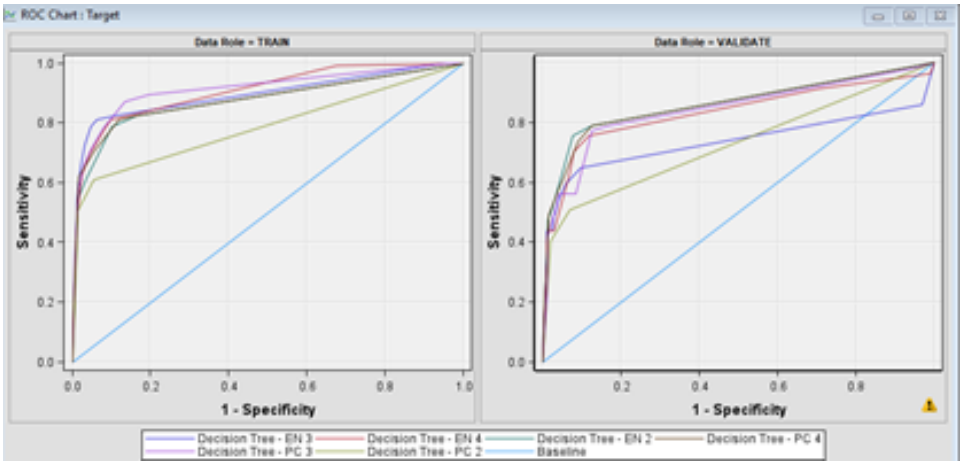
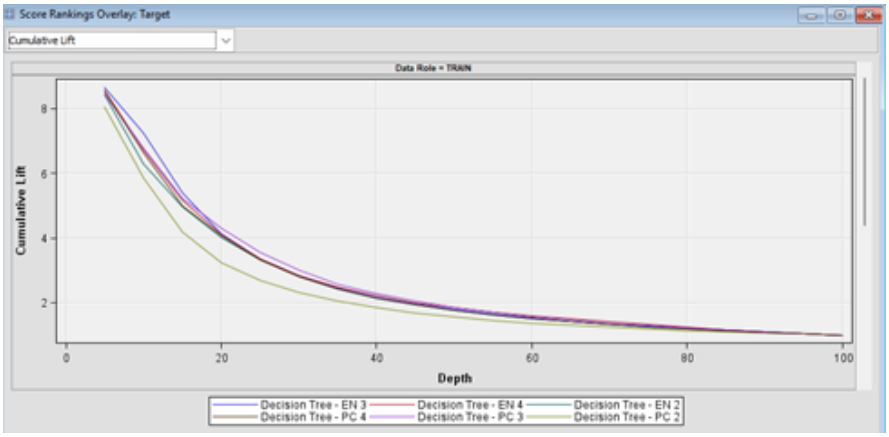


FIGURE 31: CUMULATIVE LIFT



7.APPENDIX

FIGURE 32 : CUMULATIVE % RESPONSE

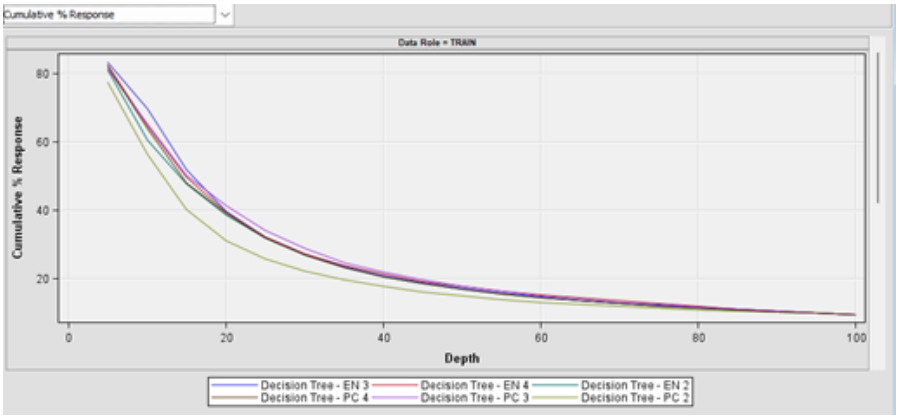


FIGURE 33 : CUMULATIVE % CAPTURED RESPONSE

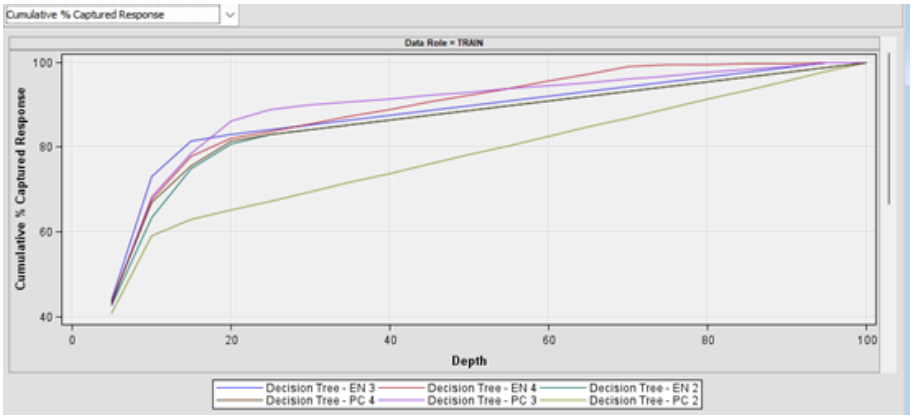
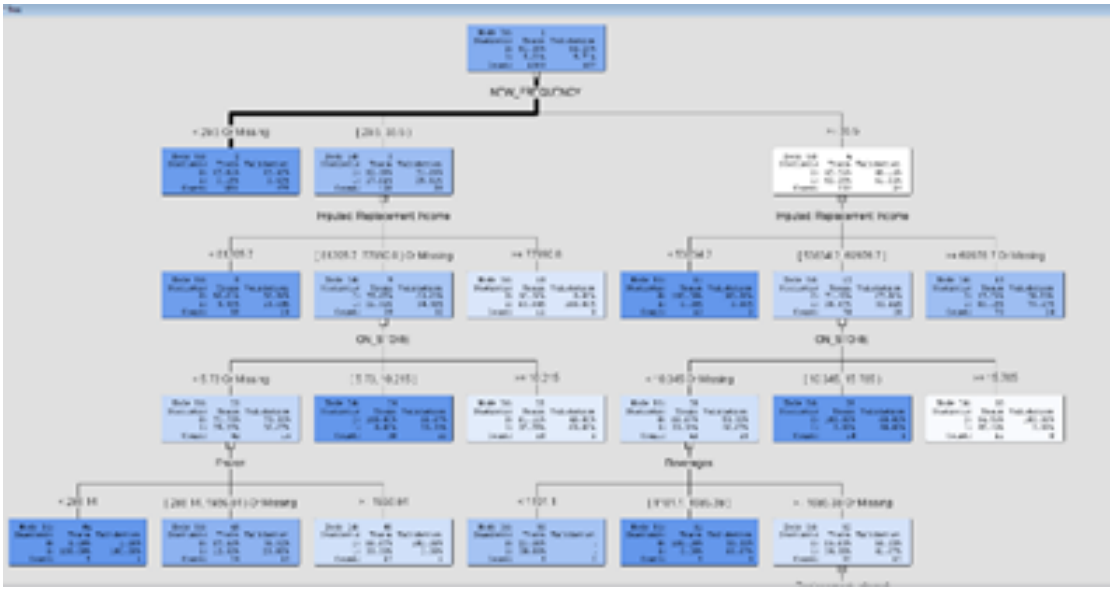


FIGURE 34: DECISION TREE- EN3



7.APPENDIX

FIGURE 35 : NEURAL NETWORK- CUMULATIVE LIFT

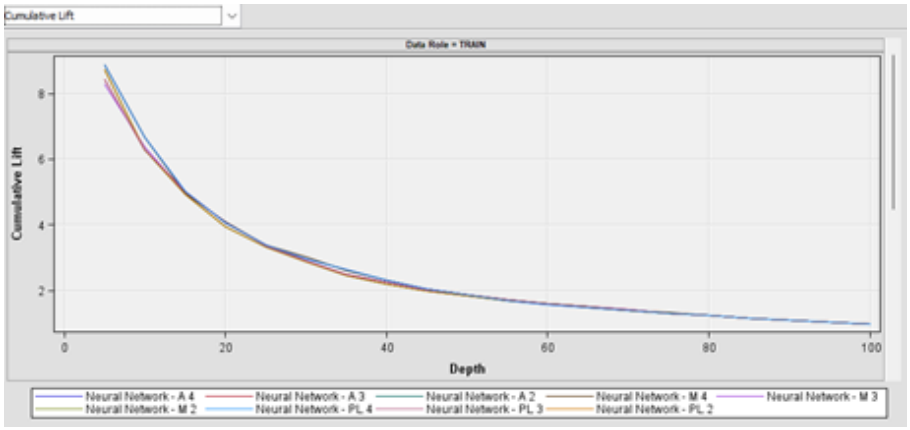


FIGURE 36 : CUMULATIVE % RESPONSE

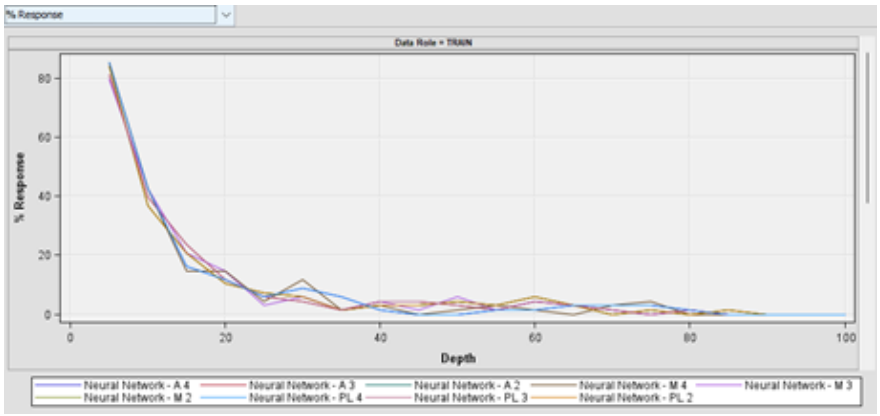


FIGURE 37 : CUMULATIVE % CAPTURED RESPONSE

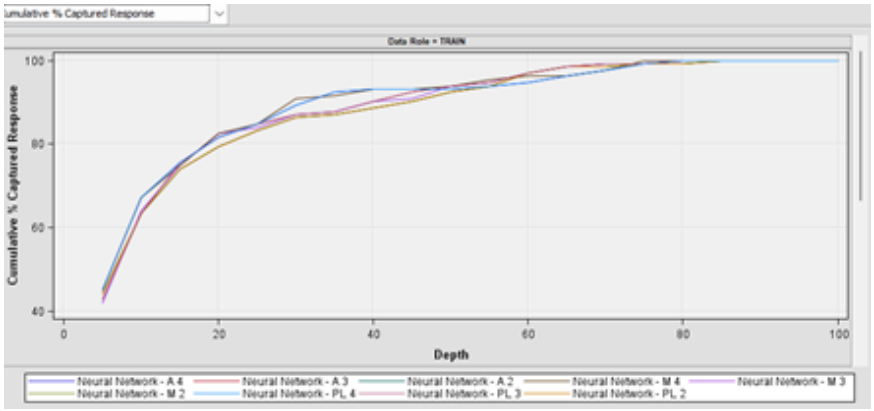


FIGURE 38 : MODEL COMPARISON

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Number of Estimated Weights
Y	MdlComp	Tree4	Decision ...	Target	Target	0.061329	1363	0.056493	0.978116	130.6694	0.047934	0.218939	2726	1363	587	0.061329	1	60.16397	0.051247	0.226378	1174			
	MdlComp2	Neural5	Neural N...	Target	Target	0.083475	1363	0.066031	0.988647	133.9547	0.04914	0.221675	2726	1363	587	0.083475	0.999952	74.71423	0.063641	0.252271	1174	1320	43	43