NOVA Information Management School (IMS), Universidade NOVA de Lisboa
Bachelor's Degree in Data Science – Foundational Aspects of Data Science
Professor Frederico Jesus, Teacher Ana Afonso

Health and Beauty Pharmacy

Group Project Report

Francisco Gomes | 20221810

João Capitão | 20221863

Miguel Nascimento | 20221876

Margarida Cruz | 20221929

Luis Mendes | 20221949

18th December 2022

ABSTRACT:

The introduction of a new nutritional service into the *Health and Beauty Pharmacy* (HBP) must be smartly conducted in order to maximize the profit. Given the necessity of a marketing approach, this document uses predictive and descriptive models to trace a profile for the customers that are more likely to purchase the service once advertised. Data containing details of HBP clients is strictly analysed. Moreover, how certain variables advance when compared to the frequency of scheduling a nutrition appointment is addressed.

KEYWORDS: Pharmacy; Data; Client Profile; Nutrition Appointment.

1. Introduction

HBP aims at offering a new weekly service to its clients – the opportunity to make an appointment with a nutritionist. Therefore, understanding which customers should be contacted and notified about the emerging service is the course to be pursued. This marketing approach has a cost of investment associated that, to the best interest of the company, must be minimized without compromising a profitable outcome. Hence, data science techniques must be used to predict demand.

The work here presented relies on the extensive dataset provided by the company. It encompasses twelve months of client information and past spending that is essential to identify a client profile that will be more likely to use the service once it is promoted. Once these are analysed, HBP will be better equipped to perform a data-driven decision and, subsequently, maximize the cost benefits of the marketing strategy.

The present report is expected to significantly contribute to that end. In particular, it provides answers for the following questions:

- What factors have the biggest impact in influencing a client to purchase the Nutrition Appointment?
- To which extent do they impact the purchase of the new service?
- Which practical implications can be drawn from that?

In answering these questions, this document is organized as follows: section two details the methods employed, whereas section three the results and main conclusions. Finally, section four presents limitations and suggestions.

2. Methodology

This section of the report presents the procedures taken on SAS Enterprise Miner and exposes some insights based on the data analysis in the form of graphics, tables and plots.

2.1. Data Pre-Processing

2.1.1. Variable definition

In the dataset provided by HBP, each observation contains data associated with a client, and the target is a binary variable indicating whether the customer purchased the new Nutrition appointment service. The role and data type definitions were set, as shown in **Fig. 1**[1], to configure how the variables should be used thenceforward.

Name	Role	Level	Report	Order	Drop
Acess	Input	Interval	No		No
Age	Input	Interval	No		No
Beauty	Input	Interval	No		No
Custid	ID	Nominal	No		No
Dayswus	Input	Interval	No		No
Educ	Input	Nominal	No		No
Freq	Input	Interval	No		No
Hair	Input	Interval	No		No
Income	Input	Interval	No		No
Kidhome	Input	Binary	No		No
Medicines	Input	Interval	No		No
Monetary	Input	Interval	No		No
NPS	Input	Ordinal	No		No
Nutriappoint	Target	Binary	No		No
Perdeal	Input	Interval	No		No
Recency	Input	Interval	No		No
Skin	Input	Interval	No		No
Teenhome	Input	Binary	No		No
Totcheck	Rejected	Binary	No		Yes
WebPurchase	Input	Interval	No		No
WebVisit	Input	Interval	No		No

***Figure 1** - Initial Variable Configuration*

[1] **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).

2.1.2. Descriptive Statistics

From the *StatExplore* node, it is possible to infer *Freq*, *Monetary*, *Age*, *Perdeal* and *Income* as the most valuable variables in the demand prediction, and *NPS*, *Teenhome* and *Educ* as the most irrelevant ones. Additionally, the median and mean values expressed in **Fig. 2** indicate that older people with higher incomes, and who spend money more frequently on the pharmacy tend to purchase the new service. Each variable's worth to predict the target is reported in **Appendix Fig. 3**.

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean
TRAIN	Nutriappoint	0	Monetary	260	1	3502	8	100000	574.3709
TRAIN	Nutriappoint	1	Monetary	1346	0	498	19	25000	1393.98
TRAIN	Nutriappoint	0	Freq	9	12	3491	1	46	12.81753
TRAIN	Nutriappoint	1	Freq	28	3	495	1	54	27.83434
TRAIN	Nutriappoint	0	Perdeal	31	1	3502	0	96	35.32267
TRAIN	Nutriappoint	1	Perdeal	3	0	498	0	82	11.43173
TRAIN	Nutriappoint	0	Skin	5	1	3502	0	75	7.358081
TRAIN	Nutriappoint	1	Skin	2	0	498	0	39	3.777108
TRAIN	Nutriappoint	0	Beauty	4	1	3502	0	61	7.227584
TRAIN	Nutriappoint	1	Beauty	2	0	498	0	42	3.885542
TRAIN	Nutriappoint	0	Acess	4	1	3502	0	77	7.173044
TRAIN	Nutriappoint	1	Acess	2	0	498	0	26	4.018072
TRAIN	Nutriappoint	0	Income	66029	1	3502	5500	8500000	69769.19
TRAIN	Nutriappoint	1	Income	98238	0	498	28648	250000	96204.83
TRAIN	Nutriappoint	0	Age	45	11	3492	18	120	46.14433
TRAIN	Nutriappoint	1	Age	65	2	496	24	78	62.79234

Figure 2 *– Interval Variable Summary Statistics (extract)*

The behaviour of the *Nutriappoint* variable in function of 3 of the most valuable variables is presented in **Figures 3-5,** extracted from the *Multiplot* node. Right away, the presence of outliers in many of the variables due to the graphs compression is detected. Though some early insights, as older people and frequent customers tend to buy the target, can be reached. Other histograms can be consulted in the *Appendices* section.

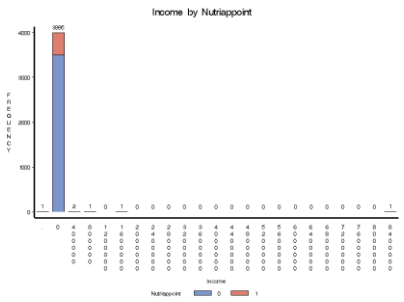


Figure 3 *– Income by Nutriappoint*

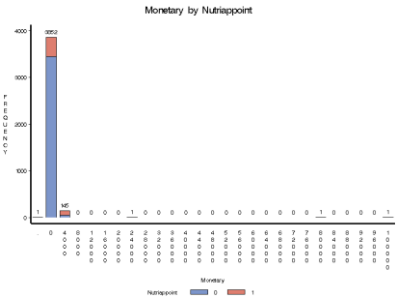


Figure 4 *– Monetary by Nutriappoint*

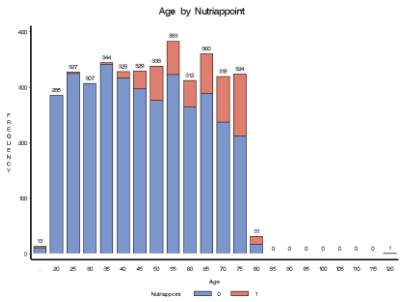


Figure 5 *– Age by Nutriappoint*

The various variable correlations of the dataset are expressed in **Fig. 6**. *WebVisits* and *WebPurchase* are highly related between themselves and with the *Perdeal* variable, and are inversely proportional to *Age* - which can be explained by the existing technology barrier. Moreover, older people are frequent customers of this pharmacy, contributing in a small percentage to purchases made on discount. It is also possible to identify a cluster composed of people who usually buy products from the Accessories, Beauty, Hair and Skin categories.

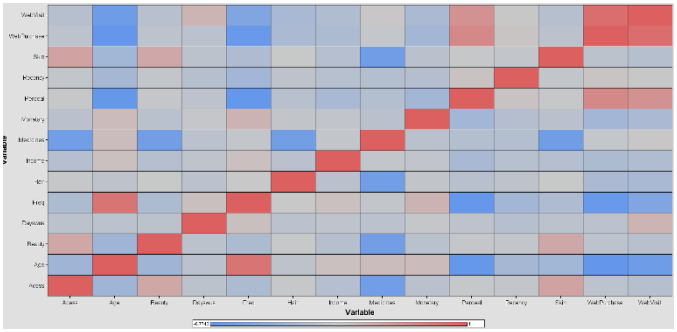
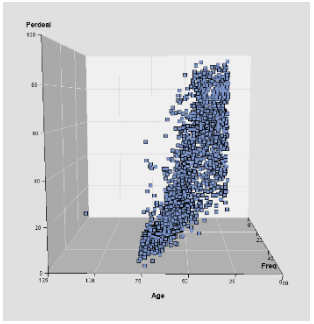
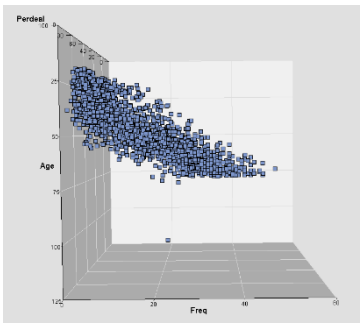


Figure 6 *– Variable Correlation (Variable Clustering node)*

Fig. 7 - 8 (extracted from the *GraphExplore* node) illustrates a different perspective on the relation between the variables *Age*, *Freq* and *Perdeal* that supports the previous insights. The most frequent buyers tend to be older and, usually, are not very keen on buying products at discounts. These insights are spoiled by an *Age* outlier.





Figures 7,8 *– 3D Scatter Plot (Age, Freq, Perdeal)*

2.1.3. *MetaData* Node

The variable *NPS* was rejected due to the high number of missing values (3280 according to **Appendix Fig. 2**) and its low worth as a variable to determine the target.

2.1.4. *Filter* Node

The boundaries stetted to eliminate outliers are stated in **Fig. 9**. These removed 15 observations (**Appendix Fig. 6**), which corresponds to about 0.37% of the total number of observations.

Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Role	Level
Acess	No	User Specified	Default	0	.	Input	Interval
Age	No	User Specified	Default	16	100	Input	Interval
Beauty	No	User Specified	Default	0	.	Input	Interval
Dayswus	No	User Specified	Default	0	1270.844	Input	Interval
Freq	No	Default	Default	0	.	Input	Interval
Hair	No	Default	Default	0	.	Input	Interval
Income	No	User Specified	Default	0	2000000	Input	Interval
Medicines	No	Default	Default	0	.	Input	Interval
Monetary	No	User Specified	Default	0	25000	Input	Interval
Perdeal	No	Default	Default	0	.	Input	Interval
Recency	No	Default	Default	0	.	Input	Interval
Skin	No	Default	Default	0	.	Input	Interval
WebPurchase	No	Default	Default	0	.	Input	Interval
WebVisit	No	Default	Default	0	.	Input	Interval

Figure 9 *– Filter Boundaries Configuration*

2.1.5. *Replacement* Node

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. From this point onwards "High School" corresponds to 12 years, "BSc" to 15 years, "MSc" to 17 years and "PhD" to 20 years. The total number of changes can be accessed on **Appendix Fig. 7**.

Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit
Acess	Default	User Specified	0	31
Age	Default	Default	0	.
Beauty	Default	User Specified	0	29
Dayswus	Default	Default	0	.
Freq	Default	User Specified	0	48
Hair	Default	Default	0	60
Income	Default	User Specified	0	122000
Medicines	Default	Default	0	.
Monetary	Default	User Specified	0	2085
Perdeal	Default	Default	0	.
Recency	Default	User Specified	0	180.0074
Skin	Default	User Specified	0	42
WebPurchase	Default	Default	0	.
WebVisit	Default	Default	0	.

Figure 10 *– Replacement Boundaries Configuration*

2.1.6. *Impute* Node

To compute the values for the missing values of the dataset, a *Tree* input method was used. Consequently, a total of 65 missing values were imputed (**Appendix Fig. 8**).

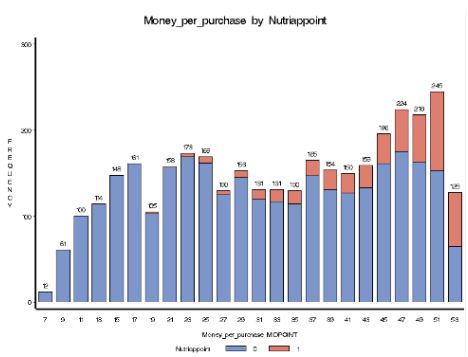
2.1.7. *Transform Variables* Node

Three new variables were created: ***Rate_of_Income*** provides the percentage of the customer's income spent on the pharmacy (*IMP_REP_Monetary / IMP_REP_Income * 100*); ***Years_Educ*** sets the variable values as numbers, multiplying the old variable *IMP_REP_Educ* by *1*; **Money_per_purchase** returns the average amount of money spent by the client in the pharmacy per purchase (*IMP_REP_Monetary / IMP_REP_Freq*).

After cleaning the dataset with filters, replacements, imputing missing values and creating new variables, it was once again analysed to identify changes or new insights.

Firstly, on *StatExplore*, the variable worth had been updated with the top led by the *Monetary*, *Freq*, and the new variables *Rate_of_Income* and *Money_per_purchase* (consult **Appendix Fig. 9**). The new variable *Years_Educ* did not add any new useful information to the dataset.

By analysing the histograms from *Multiplot*, it is possible to conclude that the outlier's problem had been solved, except for the new variable *Money_per_purchase*. Hence, another filter was applied to exclude those observations. **Fig. 11-12** represent the distribution of the target variable in function of the variables created. Clients who schedule more nutrition appointments, are the ones whose average amount of money per purchase is higher. On the other hand, the exact opposite happens on the percentage of the costumer´s income spent on the pharmacy. Other histograms are available in the *Appendices* section.



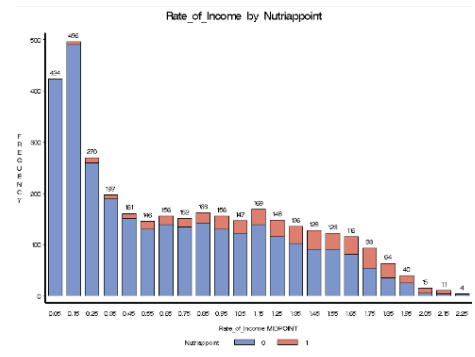


Figure 11 *- Money_per_purchase by Nutriappoint;* ***Figure 12*** *- Rate_of_Income by Nutriappoint*

2.2. Clustering

In order to determine the optimal number of clusters that allow a good heterogeneity among clusters without overcomplicating the model, the division from 2 to 8 clusters was tested. The RSQ values for each possibility were collected and registered in **Fig. 13**. The best value to be used according to the reasons stated above is 3 clusters.

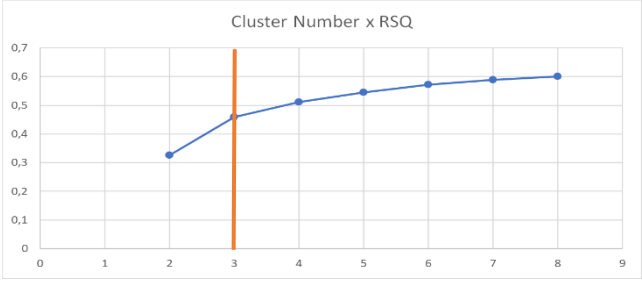


Figure 13 *- Cluster Number x RSQ Graph*

Fig. 14 displays the average values for the most valuable variables for each cluster.

Firstly, it is possible to identify cluster number 1 ("Youngsters") as a representation of the young part of the dataset - with the lowest average age among clusters. Another factor that supports this is the high percentage of purchases made at discounts and online, which points to a lower financial capacity and being comfortable using the Internet. Moreover, the fact that this cluster leaders in the categories related to self-care might indicate that it is composed in a larger part by female clients.

The second cluster ("Average People") is composed of normal buyers that do not stand out in the majority of the variables (except Medicines), which points to people that go (physically or not) to the pharmacy to buy almost exclusively medicines.

Lastly, the third cluster ("Rich/Seniors") is composed by people with higher incomes and/or advanced ages, that are frequent customers of the pharmacy, accordingly to the insights developed early in this report. More detailed information obtained from the *Segment Profile* node can be found in **Appendix Fig. 16 - 17**.

	Cluster nº	Age	Income	MoneyperPurchase	Monetary	Perdeal	Medicines	WebPurchases	Hair	Skin	Acces	Beauty
Youngsters	1	28.6301	39749.2	17.48198	78.9286	57.3328	30.0235	56.70470	34.1908	11.7832	11.3765	11.3418
Average People	2	46.6464	67723.4	30.30049	399.39	36.762	72.3115	49.16852	19.9626	2.54689	2.64066	2.50098
Rich/Seniors	3	67.0704	99698.1	46.98233	1364.45	5.44477	44.8021	21.96173	33.1572	7.58885	7.15221	7.13677

Figure 14 *- Clusters*

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.

2.3.1. Decision Trees

Three different methods of *Nominal Target Criterion* (*Probchisq, Entropy and Gini*) with three values for the maximum number of branches (2,3 and 4) were tested. To verify if these translate good predictive power or not, there is a need to compare them within themselves and with other models. The results are elaborated on the following section.

2.3.2. Neural Networks

For the purpose of finding the most efficient model for this project, Neural Networks (NN) using *Misclassification* and *Average Error* as the *Model Selection Criteria* with 2, 3 and 4 hidden units were also tested. The results of the models assessing are explained in the next section.

3. Model Comparison Results

Regarding Decision Trees, the ROC Chart in **Fig. 15** extracted from the *Model Comparison* Node shows the probability of buying the new service - true positive fraction - on the vertical axis, and the probability of not buying the new service - false positive fraction - on the horizontal axis. Therefore, the larger the area between the baseline (straight diagonal line) and the line representing the decision tree, the better the model is at correctly predicting the clients to contact. It allows to verify that the *Gini* approach with 3 branches (red line) is the best predictive model for this HBP specific dataset.

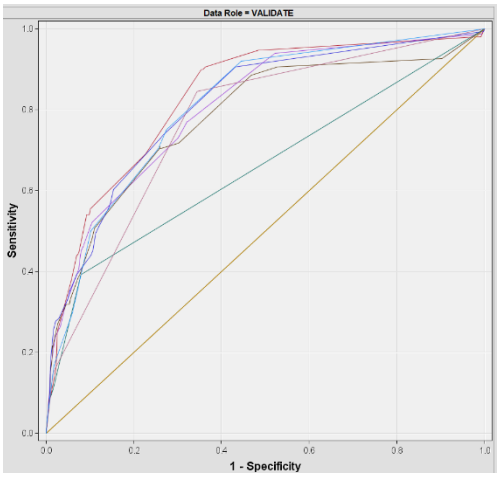


Figure 15 *– ROC Chart, Model Comparison Node (Decision Trees)*

According to the Cumulative Lift Chart in **Fig. 16** – also extracted from the *Model Comparison* Node -, if the company wishes to reach out to 10.0% of clients in the database that this model indicates, the results will be about 3.85 times better than contacting 10.0% of the clients indicated by a random baseline. The *Gini* approach with 3 branches (red line) will also perform better when contacting 15.0% and 20.0% of the clients. However, for a selection of 5% of the total customers, the *Gini* approach with 4 branches (purple line) can be more successful as it can perform around 5.32 times better than a random contact baseline (consult **Appendix Fig. 18 - 21** for more details).

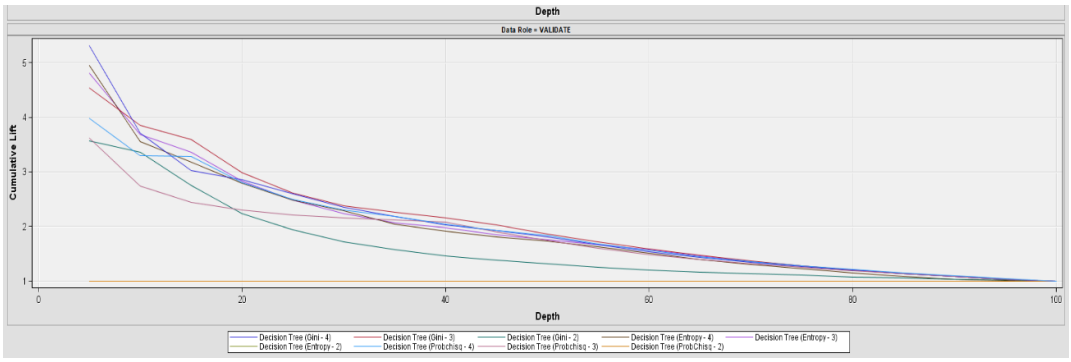


Figure 16 *– Cumulative Lift Chart, Model Comparison Node (Decision Trees)*

Regarding Neural Networks, the ROC Chart in **Fig. 17** extracted from the *Model Comparison (2)* Node allows to conclude that there is no big difference between the performance of the two methods (*Misclassification* and *Average Error*).

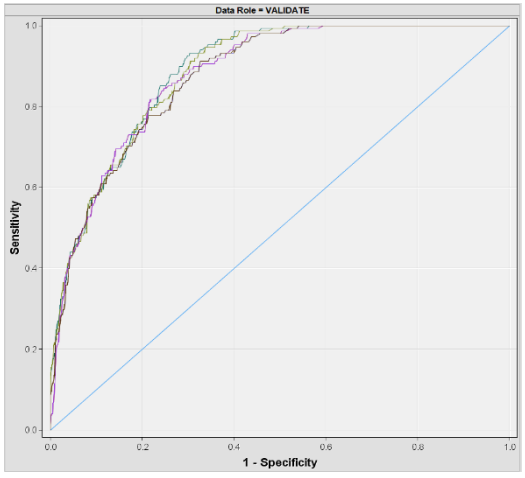


Figure 17 *– ROC Chart, Model Comparison (2) Node (Neural Networks)*

Nevertheless, the Cumulative Lift Chart in **Fig. 18** shows that to contact less than 15.0% of the clients in the database, the best predictive model to be used is one with 2 hidden units (green line); while for a bigger contact the best neural network is with 3 hidden units (consult **Appendix Fig. 22 - 25** for more details).

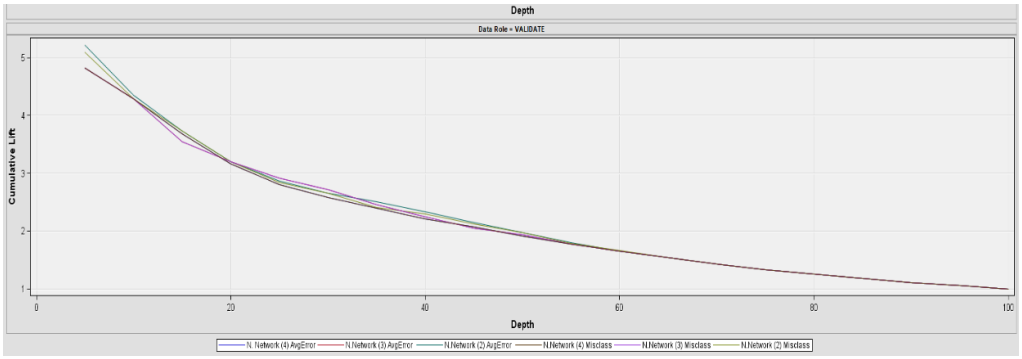


Figure 18 *– Cumulative Lift Chart, Model Comparison (2) Node (Neural Networks)*

Identifying the two best predictive models from both decision trees and neural networks and comparing them once again in the *Model Comparison* Node **- Fig. 19 -**, is possible to infer that there is a lot of variation regarding what is the best predictive model in relation with the percentage of people to be reached.

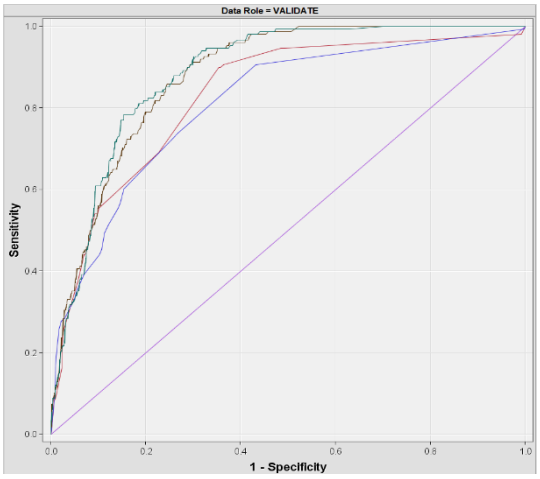


Figure 19 *– ROC Chart, Model Comparison (3) Node (final comparison)*

In accordance with **Fig. 20,** to contact 5.0% of the clients the decision tree with 4 branches and the *Gini* approach (blue line) is better, whereas from there onwards neural networks with 2 (brown line) and 3 (green line) hidden units dominate.

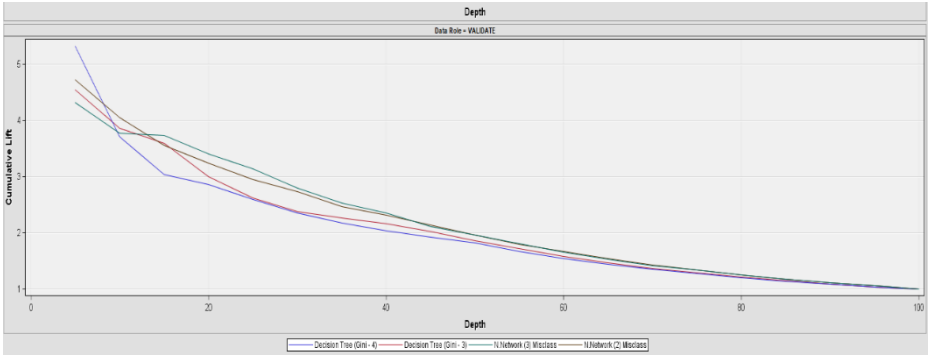


Figure 20 *– Cumulative Lift Chart, Model Comparison (3) Node (final comparison)*

4. Conclusions

Posterior to this analysis, it is possible to identify the main factors that determine whether a client might purchase the nutritional appointment or not (*Freq*, *Monetary*, *Income*, and others provenient from this values) and offer ways to HBP select clients to be informed of it. Using cluster analysis, 3 main customer profiles were established - which can be used if it is HBP wish to contact one of those groups of people.

Regarding predictive models, 3 good solutions in function of the percentage of people HBP wants to reach out to, that would predict with a good rate of success the clients to be contacted in comparison to a random selection, were identified. To contact 5.0% of the clients in the database, the Decision Tree with the *Gini* approach and 3 branches is the optimal solution; for 10.0% the Neural Network with 2 hidden units and for 15.0% forward the Neural Network with 3 hidden units.

5. Limitations and Suggestions

5.1.1. Limitations

It is important to mention that the sample studied is not representative of all HBP's customers. The data can be generalized; however, different customers can perform different actions than the corresponding to the data made available.

5.1.2. Suggestions for HBP

Depending on the company's budget for the marketing campaign, more or less clients can be contacted. Hence, there are different approaches HBP can follow.

If the company intends to contact 5% of the total clients, the *Gini4* is the best solution; for 10% it is recommended the Neural Network with 2 hidden units and for 20% the Neural Network with 3 hidden units (regarding that this one is not that different from the previous).

When choosing the strategy to follow, it is recommended to consider the simplicity of the models.

5.1.3. Suggestions for future work

As to future work, some recommendations are the implementation of a higher number of variables and an early detection of possible missing values, outliers and so forth.

As further work, it is suggest using seasonal data to figure out if there are differences on the target during different times of the year (i.e., summer, winter, etc).

APPENDICES

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	Nutriappo...	0	Monetary	260	1	3502	8	100000	574.3709	2226.79	38.42298	1597.29	INPUT	Monetary	-0.15086	1.060844	1
TRAIN	Nutriappo...	1	Monetary	1346	0	498	19	25000	1393.98	1251.668	13.52261	254.8033	INPUT	Monetary	1.060844	1.060844	2
TRAIN	Nutriappo...	0	Freq	9	12	3491	1	46	12.81753	10.89692	0.79727	-0.47364	INPUT	Freq	-0.12701	0.895764	1
TRAIN	Nutriappo...	1	Freq	28	3	495	1	54	27.83434	11.02926	-0.09181	-0.602	INPUT	Freq	0.895764	0.895764	2
TRAIN	Nutriappo...	0	Perdeal	31	1	3502	0	96	35.32267	27.86152	0.382139	-1.15497	INPUT	Perdeal	0.09195	0.646604	1
TRAIN	Nutriappo...	1	Perdeal	3	0	498	0	82	11.43173	16.01937	2.015729	3.9654	INPUT	Perdeal	-0.6466	0.646604	2
TRAIN	Nutriappo...	0	Skin	5	1	3502	0	75	7.358081	7.985508	2.035191	5.801583	INPUT	Skin	0.064499	0.453563	1
TRAIN	Nutriappo...	1	Skin	2	0	498	0	39	3.777108	4.351798	2.467665	10.71629	INPUT	Skin	-0.45356	0.453563	2
TRAIN	Nutriappo...	0	Beauty	4	1	3502	0	61	7.227584	8.17017	2.148705	5.852537	INPUT	Beauty	0.061086	0.429561	1
TRAIN	Nutriappo...	1	Beauty	2	0	498	0	42	3.885542	4.522796	2.758843	13.6778	INPUT	Beauty	-0.42956	0.429561	2
TRAIN	Nutriappo...	0	Acess	4	1	3502	0	77	7.173044	7.944055	2.147324	6.574115	INPUT	Acess	0.057932	0.407386	1
TRAIN	Nutriappo...	1	Acess	2	0	498	0	26	4.018072	4.446606	1.706186	3.375194	INPUT	Acess	-0.40739	0.407386	2
TRAIN	Nutriappo...	0	Income	66029	1	3502	5500	8500000	69769.19	147655.6	53.48909	3039.965	INPUT	Income	-0.04505	0.316784	1
TRAIN	Nutriappo...	1	Income	98238	0	498	28648	250000	96204.83	20592.44	0.509311	5.535544	INPUT	Income	0.316784	0.316784	2
TRAIN	Nutriappo...	0	Age	45	11	3492	18	120	46.14433	16.87706	0.12586	-1.06163	INPUT	Age	-0.04294	0.302343	1
TRAIN	Nutriappo...	1	Age	65	2	496	24	78	62.79234	11.00465	-0.50579	-0.5792	INPUT	Age	0.302343	0.302343	2
TRAIN	Nutriappo...	0	WebPurc...	47	1	3502	5	83	44.00428	18.0359	-0.36334	-0.87911	INPUT	WebPurc...	0.038884	0.27344	1
TRAIN	Nutriappo...	1	WebPurc...	28	0	498	6	69	30.7751	17.40122	0.31098	-1.31106	INPUT	WebPurc...	-0.27344	0.27344	2
TRAIN	Nutriappo...	0	Medicines	49	1	3502	1	97	49.36893	23.31974	-0.0583	-0.94442	INPUT	Medicines	-0.0328	0.230668	1
TRAIN	Nutriappo...	1	Medicines	63	0	498	13	98	62.81727	21.34852	-0.20069	-1.07919	INPUT	Medicines	0.230668	0.230668	2
TRAIN	Nutriappo...	0	Recency	52	1	3502	-1	549	64.02599	73.41799	3.968861	18.50979	INPUT	Recency	0.028195	0.198274	1
TRAIN	Nutriappo...	1	Recency	50	0	498	-12	100	49.92369	28.92525	-0.07017	-1.13133	INPUT	Recency	-0.19827	0.198274	2
TRAIN	Nutriappo...	0	WebVisit	6	1	3502	-3	10	5.324957	2.295021	-0.34308	-0.85154	INPUT	WebVisit	0.023096	0.162412	1
TRAIN	Nutriappo...	1	WebVisit	4	0	498	0	9	4.359438	2.347421	0.151608	-1.18827	INPUT	WebVisit	-0.16241	0.162412	2
TRAIN	Nutriappo...	0	Hair	28	1	3502	3	74	28.85523	12.52833	0.338444	-0.37954	INPUT	Hair	0.014707	0.103422	1
TRAIN	Nutriappo...	1	Hair	24	0	498	2	64	25.49598	14.44761	0.413269	-0.73923	INPUT	Hair	-0.10342	0.103422	2
TRAIN	Nutriappo...	0	Dayswus	892	8	3495	-15	1250	896.5931	207.1819	-0.08235	-0.79431	INPUT	Dayswus	-0.00131	0.009294	1
TRAIN	Nutriappo...	1	Dayswus	900	5	493	550	1249	906.1136	205.237	-0.00728	-1.21832	INPUT	Dayswus	0.009294	0.009294	2

Figure 1 - Interval Variables Summary Statistics

Data Role	Target	Target Level	Variable Name	Level	CODE	Frequency Count	Type	Percent Within	Level Index	Role	Label	Percent	Plot
TRAIN	Nutriappoint	0	Educ			4	10C	0.28547	1	INPUT	Educ	0.002499	1
TRAIN	Nutriappoint	1	Educ			4	1C	0.200803	1	INPUT	Educ	0002499	1
TRAIN	Nutriappoint	0	Educ	BSc		2	1207C	34.45618	2	INPUT	Educ	0.301675	1
TRAIN	Nutriappoint	1	Educ	BSc		0	172C	34.53815	2	INPUT	Educ	0.042989	1
TRAIN	Nutriappoint	0	Educ	High School		1	439C	12.53212	3	INPUT	Educ	0.109723	1
TRAIN	Nutriappoint	1	Educ	High School		3	48C	9.638554	3	INPUT	Educ	0.011997	1
TRAIN	Nutriappoint	0	Educ	MSc		0	1610C	45.96061	4	INPUT	Educ	0.402399	1
TRAIN	Nutriappoint	1	Educ	MSc		1	244C	48.99598	4	INPUT	Educ	0.060985	1
TRAIN	Nutriappoint	0	Educ	PhD		3	237C	6.765629	5	INPUT	Educ	0.059235	1
TRAIN	Nutriappoint	1	Educ	PhD		2	33C	6.626506	5	INPUT	Educ	0.008248	1
TRAIN	Nutriappoint	0	Kidhome			2	1N	0.028547	1	INPUT	Kidhome	0002499	1
TRAIN	Nutriappoint	0	Kidhome	0		0	1892N	54.01085	2	INPUT	Kidhome	0.472882	1
TRAIN	Nutriappoint	1	Kidhome	0		0	435N	87.3494	2	INPUT	Kidhome	0.108723	1
TRAIN	Nutriappoint	0	Kidhome	1		1	1610N	45.96061	3	INPUT	Kidhome	0.402399	1
TRAIN	Nutriappoint	1	Kidhome	1		1	63N	12.6506	3	INPUT	Kidhome	0.015746	1
TRAIN	Nutriappoint	0	NPS			0	2870N	81.92977	1	INPUT	NPS	0.717321	1
TRAIN	Nutriappoint	1	NPS			0	410N	82.32932	1	INPUT	NPS	0.102474	1
TRAIN	Nutriappoint	0	NPS	0		4	51N	1.455895	2	INPUT	NPS	0.012747	1
TRAIN	Nutriappoint	1	NPS	0		8	10N	2.008032	2	INPUT	NPS	0.002499	1
TRAIN	Nutriappoint	0	NPS	1		9	63N	1.798458	3	INPUT	NPS	0.015746	1
TRAIN	Nutriappoint	1	NPS	1		8	10N	2.008032	3	INPUT	NPS	0.002499	1
TRAIN	Nutriappoint	0	NPS	2		7	43N	1.227519	4	INPUT	NPS	0.010747	0
TRAIN	Nutriappoint	1	NPS	2		9	7N	1.405622	4	INPUT	NPS	0.00175	0
TRAIN	Nutriappoint	0	NPS	3		10	55N	1.570083	5	INPUT	NPS	0.013747	0
TRAIN	Nutriappoint	1	NPS	3		10	6N	1.204819	5	INPUT	NPS	0.0015	0
TRAIN	Nutriappoint	0	NPS	4		5	58N	1.655724	6	INPUT	NPS	0.014496	0
TRAIN	Nutriappoint	1	NPS	4		4	8N	1.606426	6	INPUT	NPS	0.002	0
TRAIN	Nutriappoint	0	NPS	5		1	66N	1.884099	7	INPUT	NPS	0.016496	0
TRAIN	Nutriappoint	1	NPS	5		2	6N	1.204819	7	INPUT	NPS	0.0015	0
TRAIN	Nutriappoint	0	NPS	6		2	56N	1.59863	8	INPUT	NPS	0.013997	0
TRAIN	Nutriappoint	1	NPS	6		3	7N	1.405622	8	INPUT	NPS	0.00175	0
TRAIN	Nutriappoint	0	NPS	7		3	55N	1.570083	9	INPUT	NPS	0.013747	0
TRAIN	Nutriappoint	1	NPS	7		11	4N	0.803213	9	INPUT	NPS	0009998	0
TRAIN	Nutriappoint	0	NPS	8		8	60N	1.712818	10	INPUT	NPS	0.014996	0
TRAIN	Nutriappoint	1	NPS	8		5	8N	1.606426	10	INPUT	NPS	0.002	0
TRAIN	Nutriappoint	0	NPS	9		6	64N	1.827005	11	INPUT	NPS	0.015996	0
TRAIN	Nutriappoint	1	NPS	9		1	17N	3.413655	11	INPUT	NPS	0.004249	0
TRAIN	Nutriappoint	0	NPS	10		11	62N	1.769912	12	INPUT	NPS	0.015496	0
TRAIN	Nutriappoint	1	NPS	10		7	5N	1.004016	12	INPUT	NPS	0.00125	0
TRAIN	Nutriappoint	0	Teenhome			2	1N	0.028547	1	INPUT	Teenhome	0002499	0
TRAIN	Nutriappoint	0	Teenhome	0		0	1784N	50.92778	2	INPUT	Teenhome	0.445889	0
TRAIN	Nutriappoint	1	Teenhome	0		1	291N	58.43373	2	INPUT	Teenhome	0.072732	0
TRAIN	Nutriappoint	0	Teenhome	1		1	1718N	49.04368	3	INPUT	Teenhome	0.429393	0
TRAIN	Nutriappoint	1	Teenhome	1		0	207N	41.56627	3	INPUT	Teenhome	0.051737	0

Figure 2 - Class Variables Summary Statistics

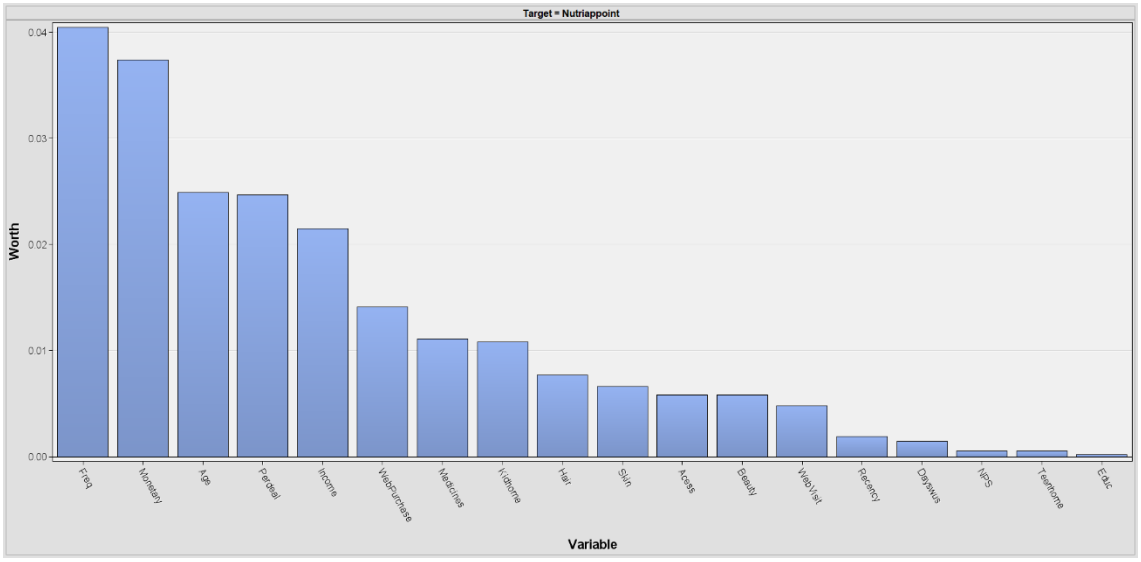


Figure 3 - Variables Worth, SatExplore

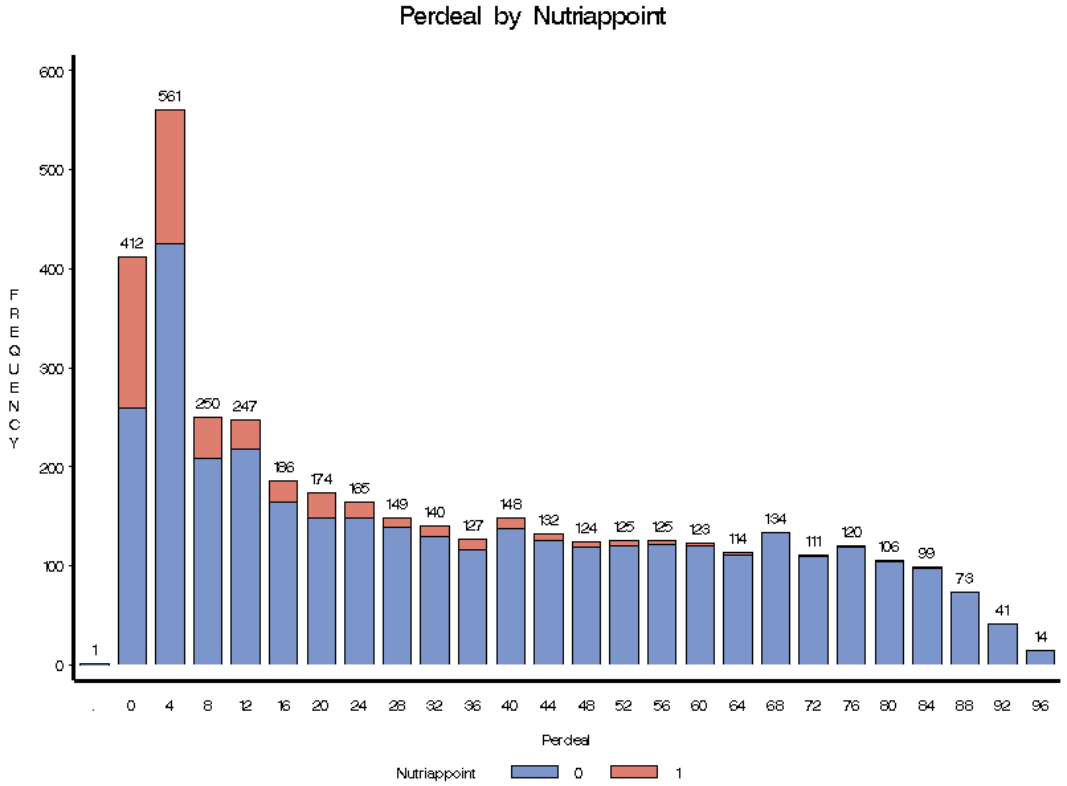


Figure 4 - Histogram Perdeal by Nutriappoint

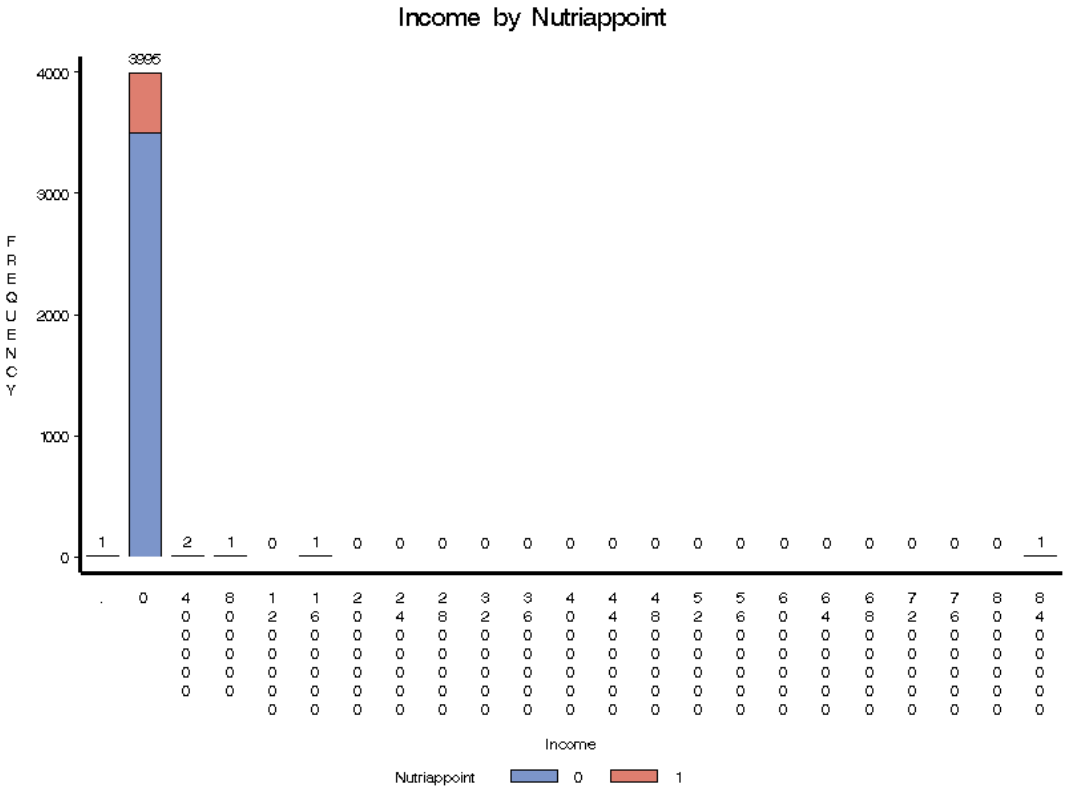


Figure 5 - Histogram Income by Nutriappoint

	Custid	Dayswus	Age	Educ	Income	Kidhome	Teenhome	Freq	Recency	Monetary	Perdeal	Medicines	Skin	Hair	Beauty	Acess	WebPurchase	WebVisit	NPS	Totcheck	Nutriappoint
1	1520.0	1229.0	45.0	BSc	8500000.0	0.0	1.0	20.0	63.0	839.0	25.0	81.0	0.0	15.0	2.0	2.0	57.0	9.0	.	1.0	0.0
2	1796.0	-10.0	67.0	BSc	104481.0	0.0	0.0	23.0	44.0	1023.0	2.0	31.0	0.0	26.0	41.0	2.0	21.0	2.0	6.0	0.0	0.0
3	2615.0	1038.0	78.0	BSc	122719.0	0.0	0.0	46.0	-10.0	2470.0	1.0	64.0	2.0	26.0	4.0	4.0	10.0	2.0	.	1.0	1.0
4	10114.0	1118.0	29.0	BSc	46165.0	1.0	1.0	3.0	21.0	100000.0	70.0	28.0	18.0	25.0	13.0	17.0	42.0	5.0	.	0.0	0.0
5	5427.0	914.0	31.0	BSc	41188.0	1.0	0.0	9.0	18.0	80000.0	41.0	27.0	12.0	41.0	11.0	9.0	62.0	8.0	.	0.0	0.0
6	7225.0	1102.0	60.0	High School	97464.0	0.0	1.0	22.0	-12.0	956.0	10.0	65.0	4.0	22.0	2.0	7.0	45.0	6.0	.	0.0	1.0
7	9047.0	1011.0	72.0	MSc	109732.0	0.0	0.0	39.0	-12.0	2046.0	1.0	48.0	3.0	40.0	3.0	5.0	14.0	2.0	.	0.0	1.0
8	4378.0	575.0	20.0	High School	34363.0	1.0	0.0	3.0	-1.0	21.0	80.0	12.0	6.0	17.0	22.0	44.0	71.0	5.0	8.0	0.0	0.0
9	3437.0	1019.0	120.0	BSc	92836.0	0.0	1.0	19.0	35.0	771.0	10.0	30.0	21.0	28.0	17.0	4.0	42.0	5.0	.	0.0	0.0
10	4724.0	745.0	67.0	MSc	117264.0	0.0	0.0	36.0	65.0	1853.0	0.0	35.0	8.0	34.0	12.0	11.0	6.0	-3.0	.	1.0	0.0
11	1231.0	-15.0	64.0	PhD	99829.0	0.0	0.0	26.0	30.0	1210.0	2.0	76.0	0.0	22.0	1.0	1.0	36.0	5.0	.	1.0	0.0
12	7848.0	-15.0	36.0	MSc	51993.0	1.0	0.0	4.0	68.0	57.0	36.0	52.0	0.0	39.0	4.0	4.0	51.0	5.0	.	0.0	0.0
13	9400.0	-15.0	55.0	MSc	95010.0	0.0	1.0	21.0	70.0	895.0	12.0	46.0	8.0	34.0	5.0	7.0	38.0	4.0	.	0.0	0.0
14	2115.0	-15.0	47.0	MSc	74083.0	0.0	1.0	3.0	22.0	41.0	40.0	72.0	3.0	22.0	3.0	0.0	53.0	3.0	.	0.0	0.0
15	2957.0	1128.0	20.0	High School	26924.0	0.0	0.0	4.0	18.0	69.0	35.0	12.0	13.0	32.0	27.0	16.0	27.0	-1.0	.	0.0	0.0

Figure 6 -Excluded Observations

Replacement Counts

Obs	Variable	Label	Role	Train
1	Acess	Acess	INPUT	68
2	Age	Age	INPUT	0
3	Beauty	Beauty	INPUT	102
4	Dayswus	Dayswus	INPUT	0
5	Educ	Educ	INPUT	3975
6	Freq	Freq	INPUT	10
7	Hair	Hair	INPUT	32
8	Income	Income	INPUT	66
9	Medicines	Medicines	INPUT	0
10	Monetary	Monetary	INPUT	117
11	Perdeal	Perdeal	INPUT	0
12	Recency	Recency	INPUT	150
13	Skin	Skin	INPUT	8
14	WebPurchase	WebPurchase	INPUT	0
15	WebVisit	WebVisit	INPUT	0

Figure 7 - Replacement Counts

Variable Name	Number of Missing for TRAIN ▲	Impute Method	Imputed Variable
Kidhome	1	TREE	IMP_Kidhome
REP_Acess	1	TREE	IMP_REP_Acess
REP_Beauty	1	TREE	IMP_REP_Beauty
REP_Hair	1	TREE	IMP_REP_Hair
REP_Income	1	TREE	IMP_REP_Income
REP_Medicines	1	TREE	IMP_REP_Medicines
REP_Monetary	1	TREE	IMP_REP_Monetary
REP_Perdeal	1	TREE	IMP_REP_Perdeal
REP_Recency	1	TREE	IMP_REP_Recency
REP_Skin	1	TREE	IMP_REP_Skin
REP_WebPurchase	1	TREE	IMP_REP_WebPurchase
REP_WebVisit	1	TREE	IMP_REP_WebVisit
Teenhome	1	TREE	IMP_Teenhome
REP_Educ	11	TREE	IMP_REP_Educ
REP_Age	13	TREE	IMP_REP_Age
REP_Dayswus	13	TREE	IMP_REP_Dayswus
REP_Freq	15	TREE	IMP_REP_Freq

Figure 8 - Imputation Summary

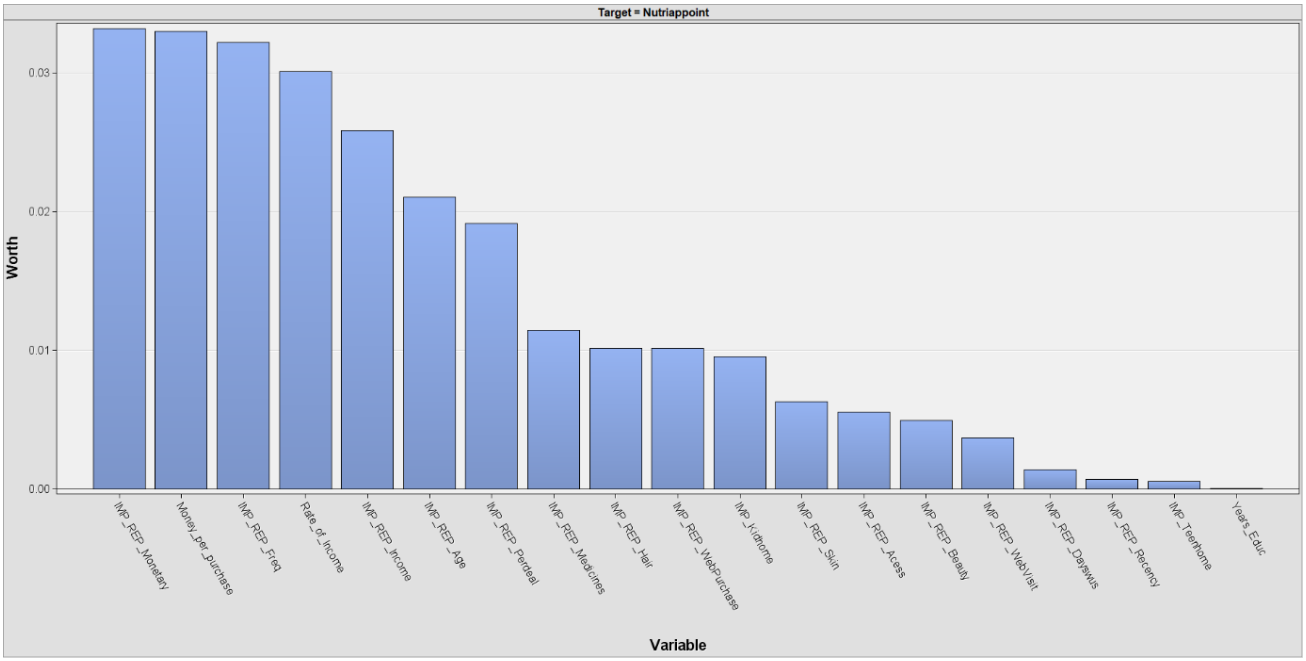


Figure 9 – New Variable Worth, StatExplore

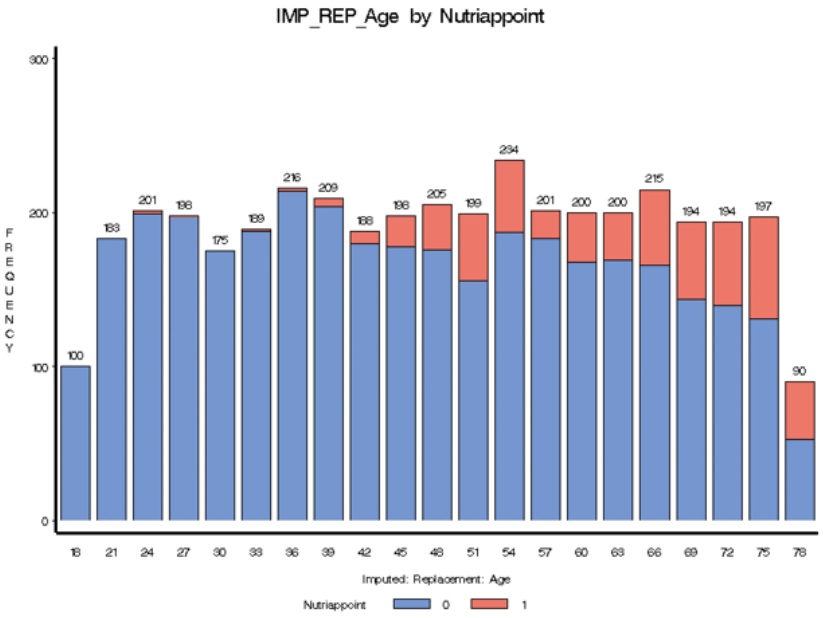


Figure 10 8 -Histogram IMP_REP_Age by Nutriappoint

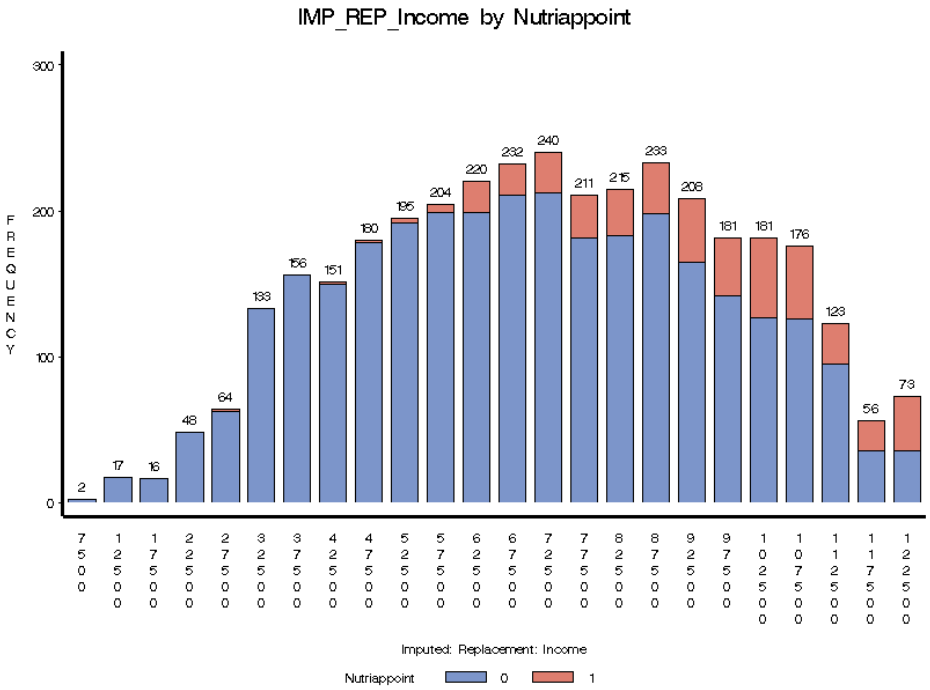


Figure 11 9 - Histogram IMP_REP_Income by Nutriappoint

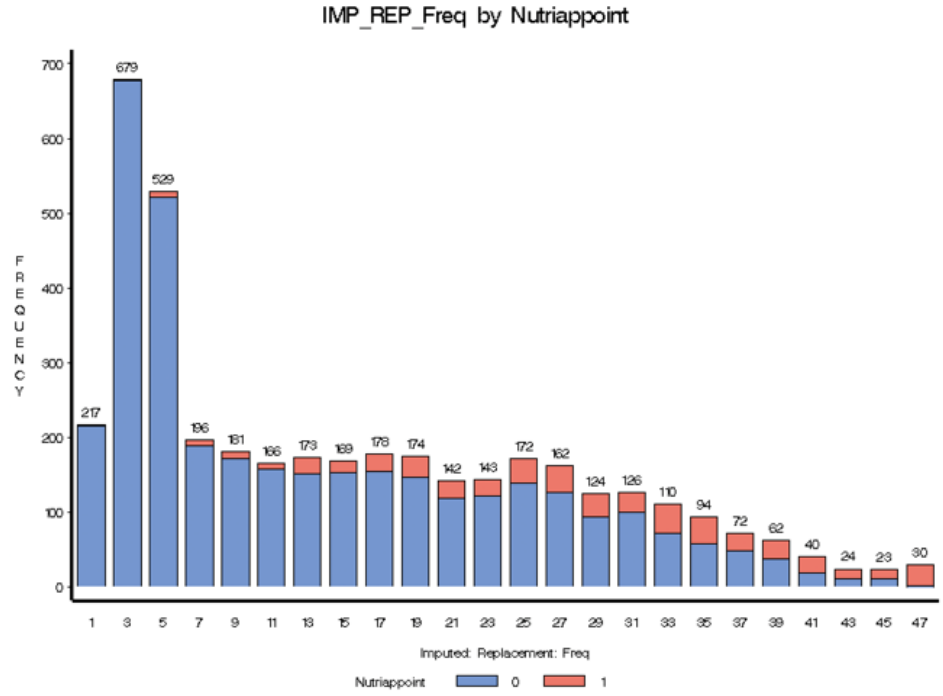


Figure 1210 - Histogram IMP_REP_Freq by Nutriappoint

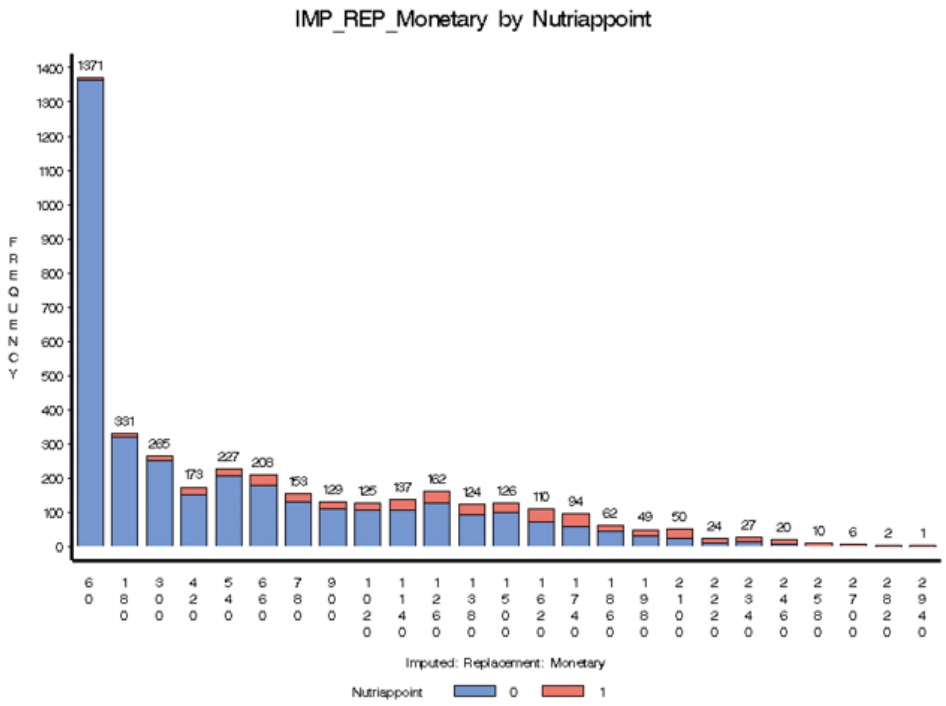


Figure 13 - Histogram IMP_REP_Monetary by Nutriappoint

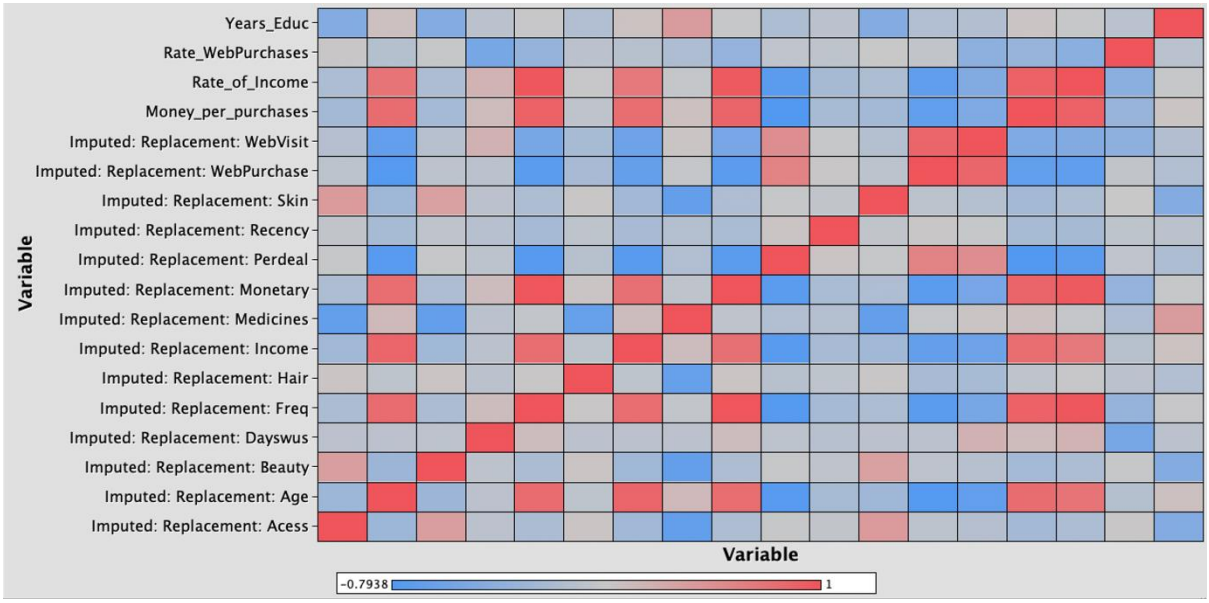


Figure 14 – New Variable Correlation

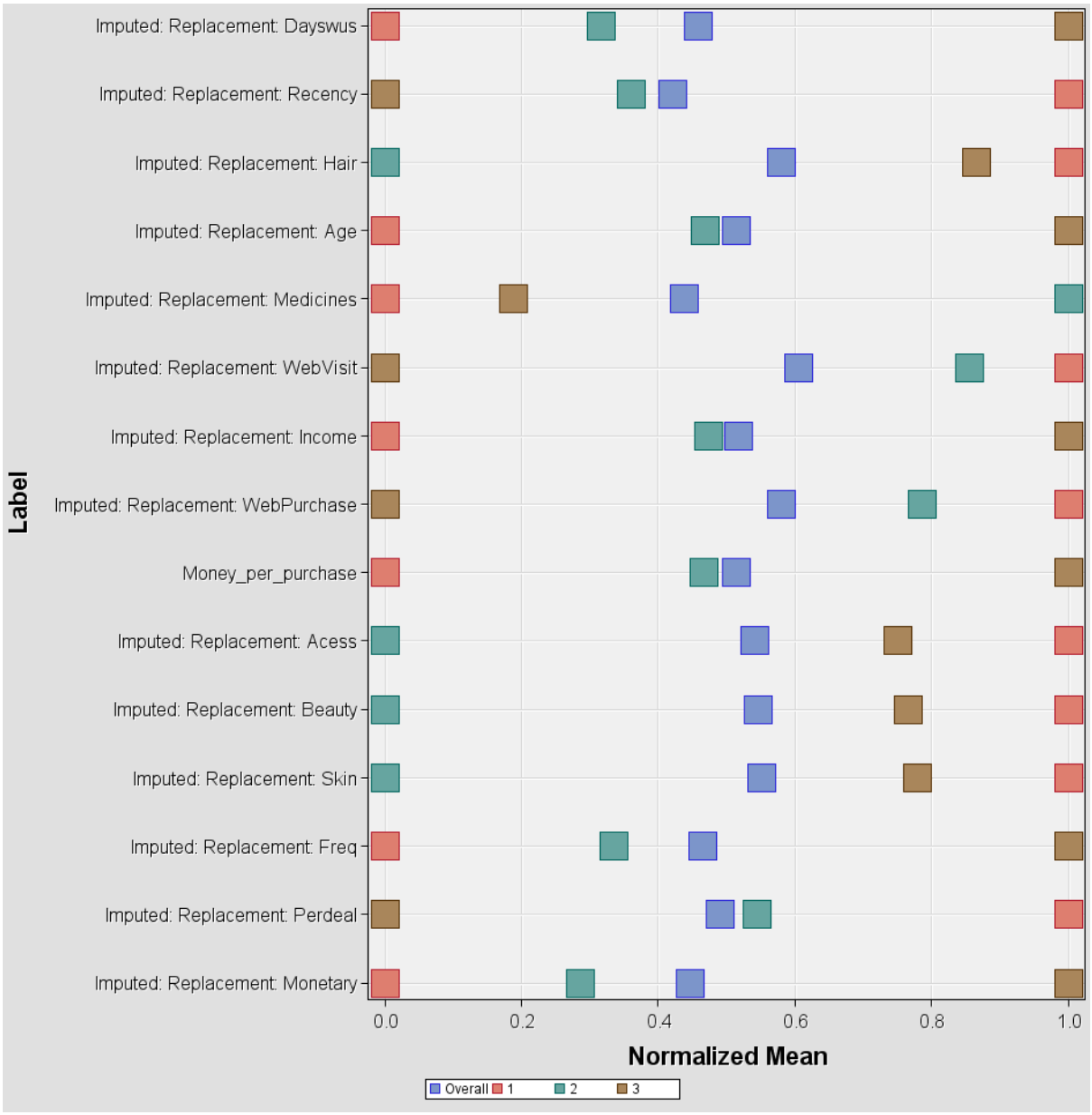


Figure 15 – Input Means Plot

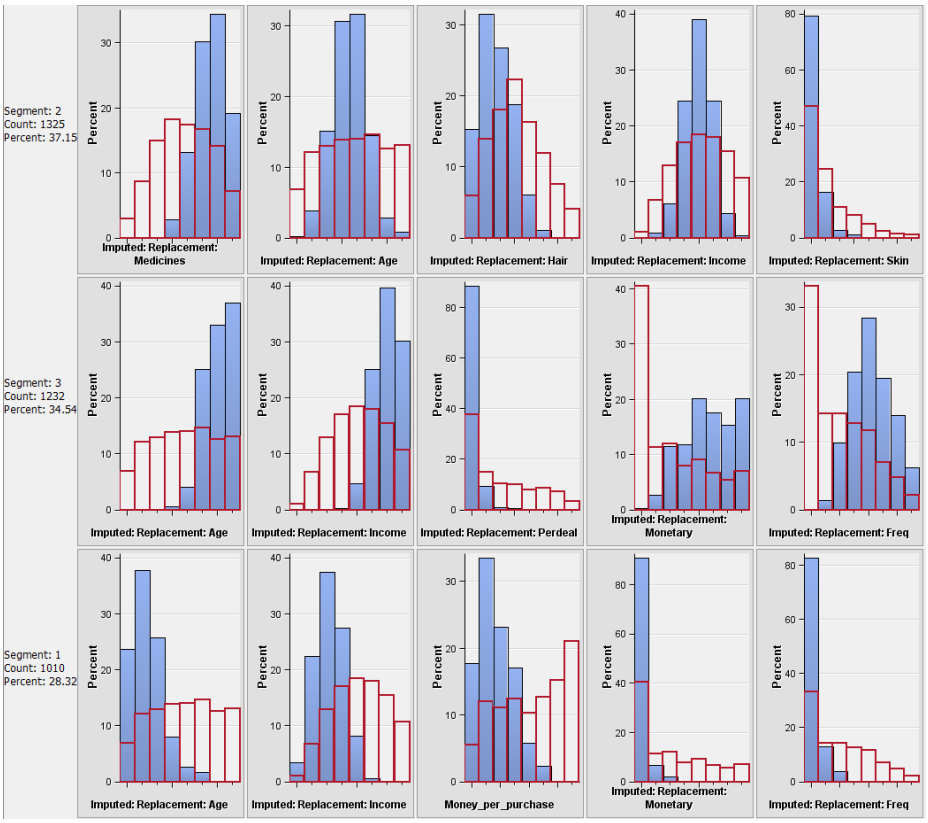


Figure 16 – 3 Clusters Segment Profile (1)

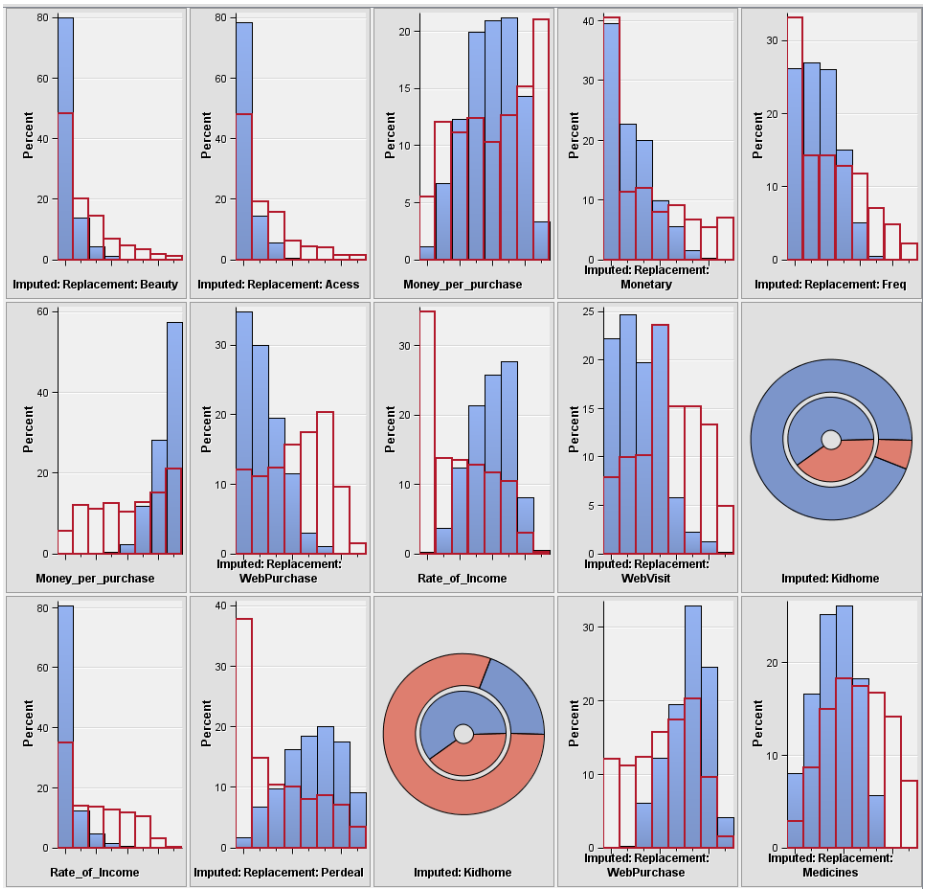


Figure 17 – 3 Clusters Segment Profile (2)

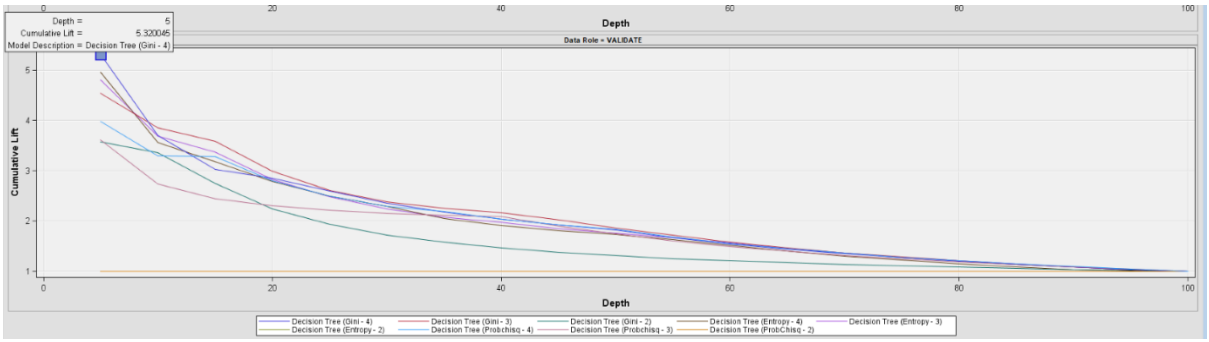


Figure 11 – DT Gini4 Cumulative Lift (5.3%, depth = 5)

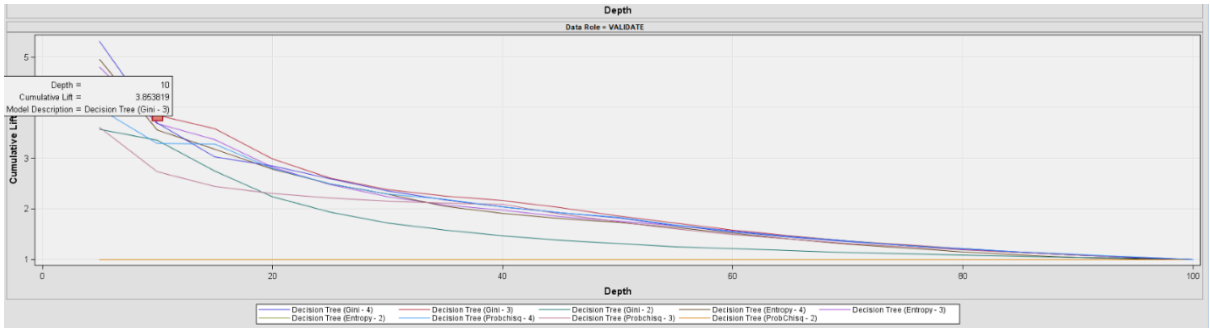


Figure 12 – DT Gini3 Cumulative Lift (3.85%, depth = 10)

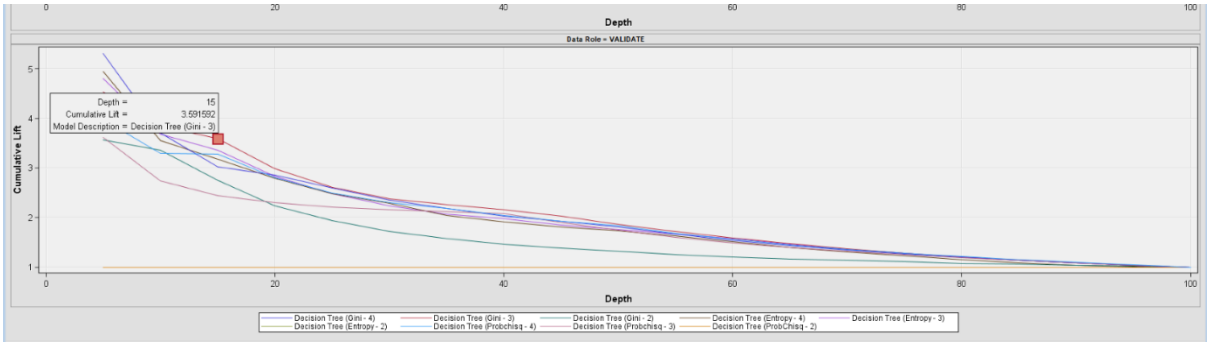


Figure 20 - DT Gini3 Cumulative Lift (3.59%, depth = 15)

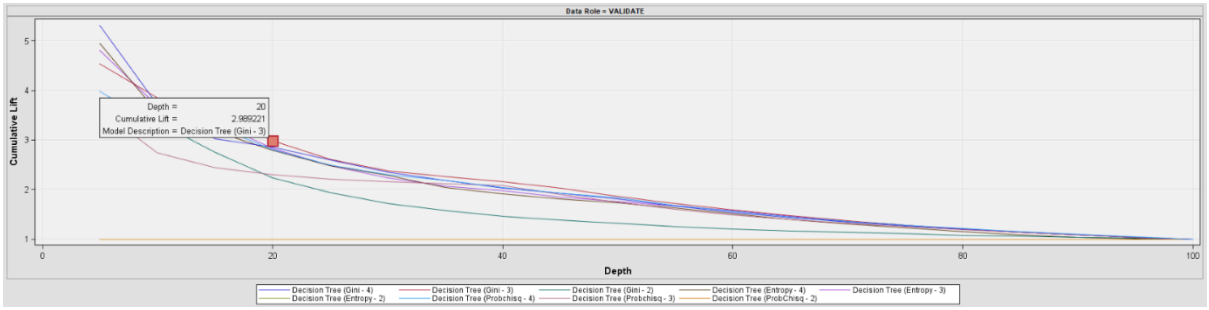


Figure 21 - DT Gini3 Cumulative Lift (2.98%, depth = 20)

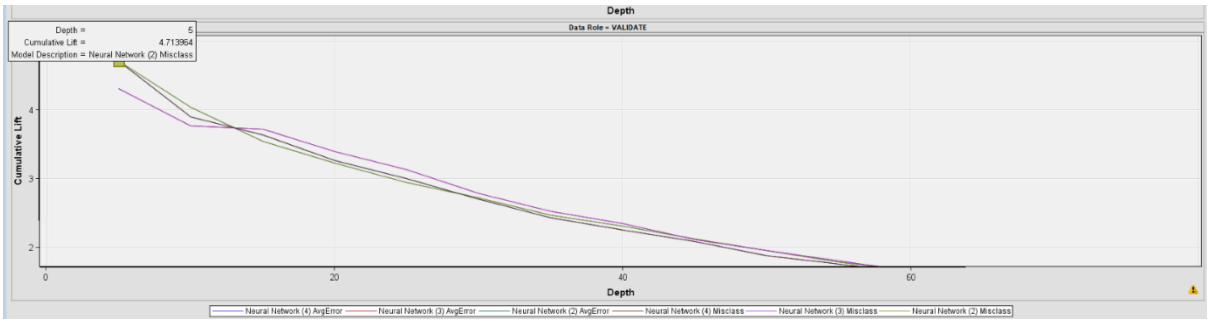


Figure 22 - NN (2) Misclass Cumulative Lift (4.71%, depth = 5)

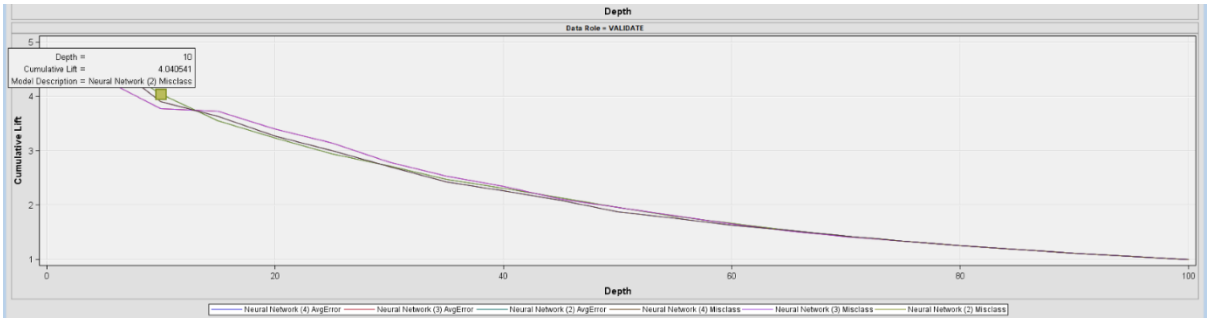


Figure 23 - NN (2) Misclass Cumulative Lift (4.0%, depth = 10)

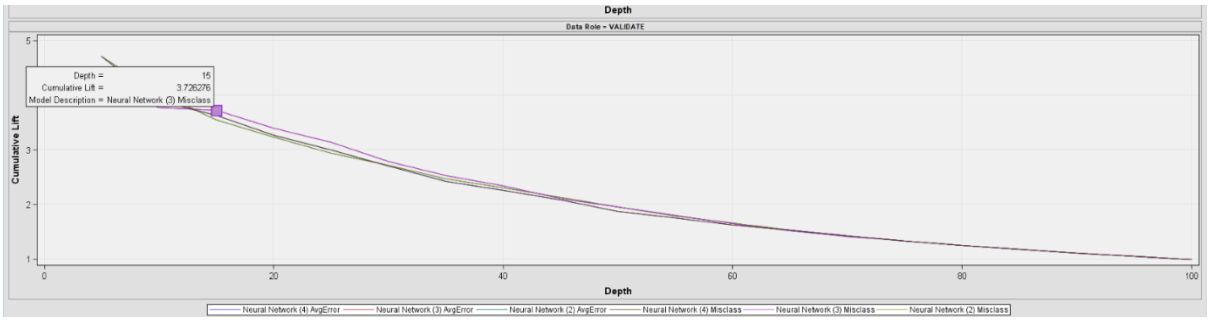


Figure 24 - NN (3) Misclass Cumulative Lift (3.72%, depth = 15)

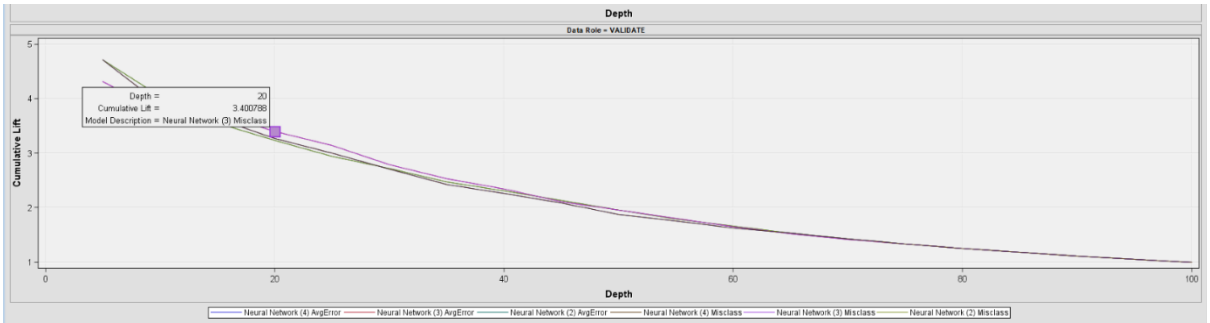


Figure 25 - NN (3) Misclass Cumulative Lift (3.4%, depth = 20)