An Analysis of a Company's Ideal Customers

Cueva Bravo, Sánchez Galindo, Soto Lazarte, Vega Feriz

Centennial College

Business Analytics Capstone: SAS Enterprise Miner Project

David Parent

July 15, 2022

Contents

Summary

The present report communicates the results of a statistical analysis constructed with the program SAS Enterprise Miner. A public domain database was chosen to study a supermarket's customers' characteristics such as income, education, household members, purchasing patterns, and preferred types of purchasing channels. This project has been aimed to identify the best potential customers for a company by determining whether they accept the offer to the last marketing campaign conducted by the company. Different statistical modeling tools, including decision trees, logistic regressions, and neural networks were applied. A Model Comparison node used returned a Neural Network as the best model.

Data

The Data

The present study is an analysis of the customers' traits and behaviors that have the greatest impact on the likelihood that a purchase will be made. With this in mind, the "Customer Personality Analysis: Analysis of Company's ideal Customers" dataset, created by Akash Patel in 2021, was ideal since it was constructed to conduct analyses of this kind. The dataset was downloaded from the platform Kaggle.

According to the problem statement provided by Patel (2021):

"Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers."

Data Dictionary. The dataset contains attributes organized in the following groups: People, Products, Promotion, Place. The People group contains variables pertaining to the socio-

economic characteristics of customers, including *income*, *education*, *marital_status* among others. Refer to Figure 1.

Figure 1. Group 1: People

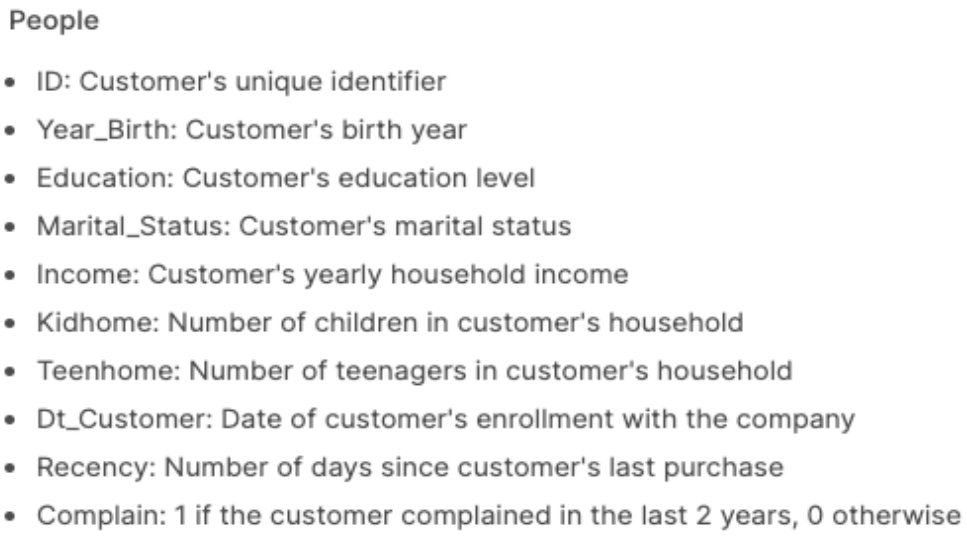


Figure 1. From: Customer Personality Analysis, by Akash Patel, 2021. Retrieved from https://www.kaggle.com/imakash3011/customer-personality-analysis

The Products group describes the purchasing patterns of customers by indicating the amount of dollars that customers spent on various products' categories during the last two years. Examples of these categories are *MntWines*, *MntFruits*, etc. Refer to Figure 2.

Figure 2. Group 2: Products

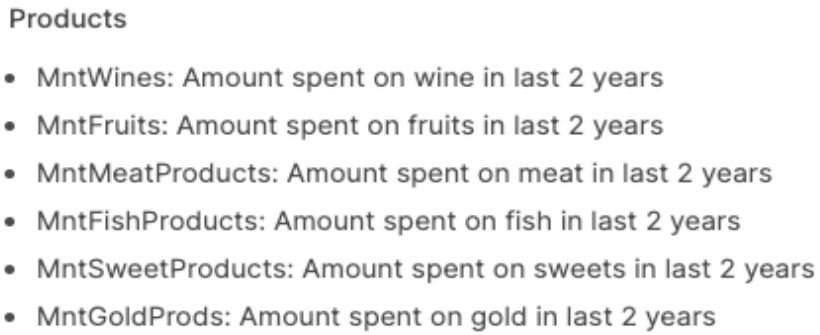


Figure 2. From: Customer Personality Analysis, by Akash Patel, 2021. Retrieved from https://www.kaggle.com/imakash3011/customer-personality-analysis

The Promotion group can be used to understand how well a customer responds to different marketing efforts conducted by the company. Variables in this group include *AcceptedCmp1*, *NumDealsPurchases*, *Response,* etc. Refer to Figure 3.

Figure 3. Group 3: Promotion

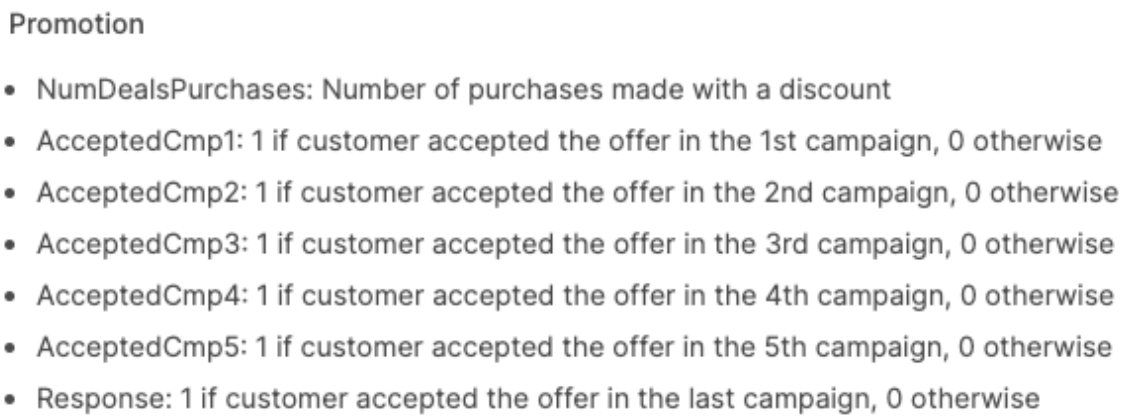


Figure 3. From: Customer Personality Analysis, by Akash Patel, 2021. Retrieved from https://www.kaggle.com/imakash3011/customer-personality-analysis

The group Place describes which purchasing channels are preferred by customers. Variables in this group include *NumWebPurchases, NumCatalogPurchases, and NumbWebVisitsMonth*, among others. Refer to Figure 4.

Figure 4. Group 4: Place

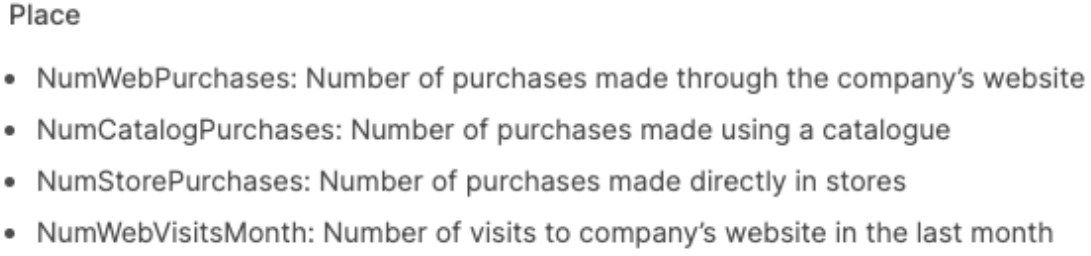


Figure 4. From: Customer Personality Analysis, by Akash Patel, 2021. Retrieved from https://www.kaggle.com/imakash3011/customer-personality-analysis

Dataset Statistics

The dataset has 2,240 observations and 29 variables. *income* is the only variable that presents missing observations (24 missing values or 1.07 percent of the complete dataset).

There is a case of correlation and multicollinearity between the different *AcceptedCmp#* variables. These variables are thought to follow a time-series pattern given that campaigns might have been implemented in a consecutive fashion, one after the other. These variables are also highly correlated with *Response*. As it can be found in the Data Dictionary section of this report, *Response* is 1 when *AcceptedCmp5* is 1.

In addition to multicollinearity and missing observations, it was found that the following variables present severe right-tailed skewness and outliers: *income, MntFishProducts, MntFruits, MntGoldProducts, MntMeatProducts, MntSweetProducts, MntWines.* Refer to Figure 5.

Figure 5. Dataset Statistics

Name	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
AcceptedCmp1	2	0	.	.	.	.	.	.
AcceptedCmp2	2	0	.	.	.	.	.	.
AcceptedCmp3	2	0	.	.	.	.	.	.
AcceptedCmp4	2	0	.	.	.	.	.	.
AcceptedCmp5	2	0	.	.	.	.	.	.
Complain	2	0	.	.	.	.	.	.
Dt_Customer	21	0	.	.	.	.	.	.
Education	5	0	.	.	.	.	.	.
ID	.	0	0	11191	5592.16	3246.662	0.039832	-1.19003
Income	.	1.071429	1730	666666	52247.25	25173.08	6.763487	159.6367
Kidhome	3	0	.	.	.	.	.	.
Marital_Status	8	0	.	.	.	.	.	.
MntFishProducts	.	0	0	259	37.52545	54.62898	1.919769	3.096461
MntFruits	.	0	0	199	26.30223	39.77343	2.102063	4.050976
MntGoldProds	.	0	0	362	44.02188	52.16744	1.886106	3.551709
MntMeatProducts	.	0	0	1725	166.95	225.7154	2.083233	5.516724
MntSweetProducts	.	0	0	263	27.06295	41.2805	2.136081	4.376548
MntWines	.	0	0	1493	303.9357	336.5974	1.175771	0.598744
NumCatalogPurchases	14	0	.	.	.	.	.	.
NumDealsPurchases	15	0	.	.	.	.	.	.
NumStorePurchases	14	0	.	.	.	.	.	.
NumWebPurchases	15	0	.	.	.	.	.	.
NumWebVisitsMonth	16	0	.	.	.	.	.	.
Recency	.	0	0	99	49.10938	28.96245	-0.00199	-1.2019
Response	2	0	.	.	.	.	.	.
Teenhome	3	0	.	.	.	.	.	.
Year_Birth	.	0	1893	1996	1968.806	11.98407	-0.34994	0.717447
Z_CostContact	1	0	.	.	.	.	.	.
Z_Revenue	1	0	.	.	.	.	.	.

Figure 5. Output from dataset statistics

Variables

The binary variable *response* is the best option to explain customers' purchasing decision because:

1. the objective of this project is to analyze customer behavior and its impact on purchasing decision.

2. the shortfalls of the dataset; especially as it relates to the multicollinearity explained before.

It was decided that the five *AcceptedCmp#* variables should be rejected on the grounds of high multicollinearity, and more specifically, due to the negative impact that they might bring as they are highly related to the selected target variable *response. Dt_Customer, Z-CostContact,* and *Z_Revenue* were automatically rejected by SAS Enterprise Miner.

After rejecting the variables mentioned in the paragraph above, a total of 19 input variables remained and were used in the model, refer to Figure 6. The remaining variables' data types are as follow:

- Binary: one variable was declared as binary.
- Nominal: four variables were declared as nominal.
- Interval: fourteen variables were declared as interval.

Data Wrangling

As it has been described in various sections of this report, there are variables related to purchasing patterns such as *MntWines*. There are also variables that describe the consumers' preferred purchasing channels such as *NumWebPurchases*, and one variable explaining how often customers buy their products (in *recency*). There are also additional socioeconomic variables such as *income* and *marital_status*.

SAS Enterprise Miner correctly assessed the variable levels. There was no need to switch variables to interval in the data importation process. However, there is suspicion that the variables that describe the amount of dollars spent on the various class-items might add to the curse of dimensions. Due to the potential of limiting the predictive capabilities of our models by rejecting an excess of variables, it was decided to accept the aforementioned variables.

Figure 6. Variable Role

Name	Role	Level
AcceptedCmp1	Rejected	Binary
AcceptedCmp2	Rejected	Binary
AcceptedCmp3	Rejected	Binary
AcceptedCmp4	Rejected	Binary
AcceptedCmp5	Rejected	Binary
Complain	Input	Binary
Dt_Customer	Rejected	Nominal
Education	Input	Nominal
ID	ID	Interval
Income	Input	Interval
Kidhome	Input	Nominal
Marital_Status	Input	Nominal
MntFishProducts	Input	Interval
MntFruits	Input	Interval
MntGoldProds	Input	Interval
MntMeatProducts	Input	Interval
MntSweetProducts	Input	Interval
MntWines	Input	Interval
NumCatalogPurchases	Input	Interval
NumDealsPurchases	Input	Interval
NumStorePurchases	Input	Interval
NumWebPurchases	Input	Interval
NumWebVisitsMonth	Input	Interval
Recency	Input	Interval
Response	Target	Binary
Teenhome	Input	Nominal
Year_Birth	Input	Interval
Z_CostContact	Rejected	Unary
Z_Revenue	Rejected	Unary

Figure 6. Variable definition

After exploring the dataset and checking the results of a few models, it was found that the model selected by the Model Comparison node was a full regression. This model included *all* the variables available. Many of these variables had an excessive and irrelevant number of classes that had a very low number of observations; in some cases, the classes were ambiguous or impossible to decipher. All these problems took a toll on the predictive capabilities of the model. As a result, it was concluded that it was best to collapse the classes for the variables *education* and *marital_status*.

Looking to improve the sensitivity of the predictive model and to address the poor class design problem, an Interactive Binning node was used. For *education,* the observations that fall in the Master and PhD classes were binned in a single group labeled “2”. The observations that fall in 2nd Cycle and Basic were binned in a single group labeled “3”. The observations found in

the class Graduation were left unchanged and assigned to the class grouped in “1”. Following the rationale of being together or alone, regardless of legal status, *marital_status* was collapsed as follows: the classes Married and Together were binned in group “1”. The remaining variables, Divorced, Single, Widow, Absurd, Alone, and YOLO were binned in group “2”. With the addition of the new *Grouped: Marital_Status* and *Grouped: Education* binned variables, it was decided to drop the original *education* and *marital_status* variables to avoid the potential of accidentally introducing multicollinearity to the model. For more details about this binning methodology, refer to the Appendix section of this paper, Appendix 1 and Appendix 2.

A Replacement node was used as an alternative to collapse the classes found in *education* and *marital_status*. The replacement classes used for *education* are GRAD for the Graduation class, PG for PhD and Master, and UG for 2nd Cycle, and Basic. The replacement values used for *marital_status* are TO for Married and Together, SI for Single, SI/M for Divorced and Widow, and OT for Absurd, Alone, and YOLO. It is important to mention that the classes created in the replacement node were designed looking to maintain the original classes as much as possible and introduce the least amount of bias. Refer to Appendix 3.

The data exported from the Replacement node was used to construct additional models that were compared to the Interactive Binning data models. As it was later found, the fit statistics of some Replacement node data models are superior to those resulting from Interactive Binning data.

Decision Trees

The first modeling technique used in the project was Decision Tree. This is one of the easier to use and more forgiving modeling tools available; nonetheless, it is still important to feed the model the most complete and readable dataset possible. Given that the 1.07% of missing

values present in income were correctly categorized as missing, there was no need to run a Replacement node before running these models. Only File Import and Data Partition nodes were used prior to running the first Decision Tree. The dataset was equally partitioned, with 50% of the observations going to the training dataset and the remaining 50% of the observations going to the validation dataset. Refer to Figure 7 for a visualization of the nodes run up to this point.

Figure 7. Decision Tree Diagram

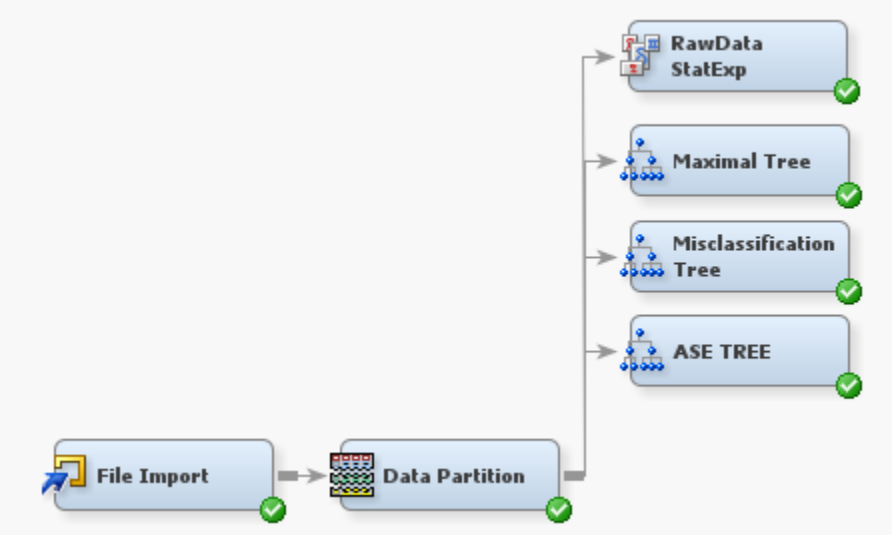


Figure 7. Diagram 1

Maximal Tree

The first Decision Tree run was a Maximal Tree. This was accomplished using the interactive training tool found in the Decision Tree node and training the root node. Refer to Figure 8 to see the Maximal Tree.

After training the root node to obtain the Maximal Tree, the setting Use Frozen Tree was changed to *yes* and the node was run. The resulting Maximal Tree has a total of 30 leaves, the Validation Misclassification Rate stands at 15.42% and the Validation Average Squared Error stands at 11.75%. The following variables are found to have the highest logworth when splitting

the root node: *MntWines*(16.433), *NumCatalogPurchases*(12.709), *Income* (12.348). Refer to Figure 9 to consult the Maximal Tree results window.

Figure 8. Maximal Tree

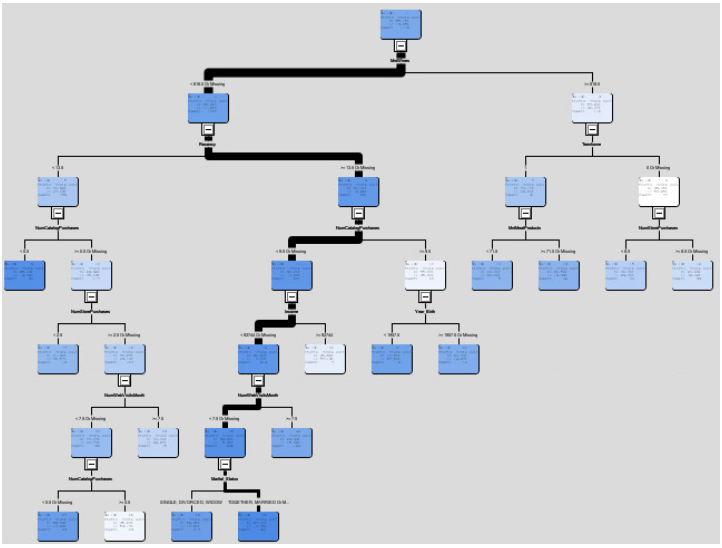


Figure 8. Maximal Tree output

Figure 9. Maximal Tree Results Window

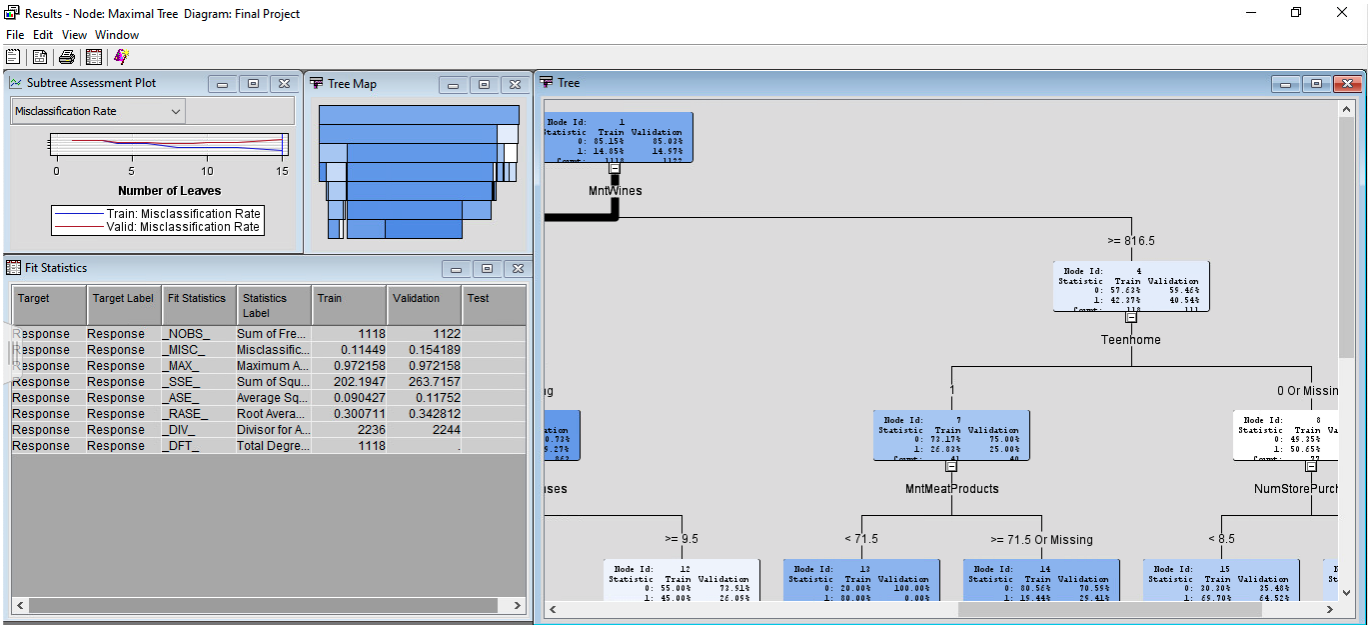


Figure 9. Results Window

Misclassification Tree

The second Decision Tree was a Misclassification. The following settings were applied before running this model: Method – Assessment, Assessment Measure – Misclassification. The resulting Tree has a total of 17 leaves, but only 7 are selected in the leaf statistics. The Validation Misclassification Rate stands at 13.90%, and the Validation Average Squared Error stands at 11.11%. Refer to Figure 10 to consult the Misclassification Tree results window.

Figure 10. Misclassification Tree Results Window

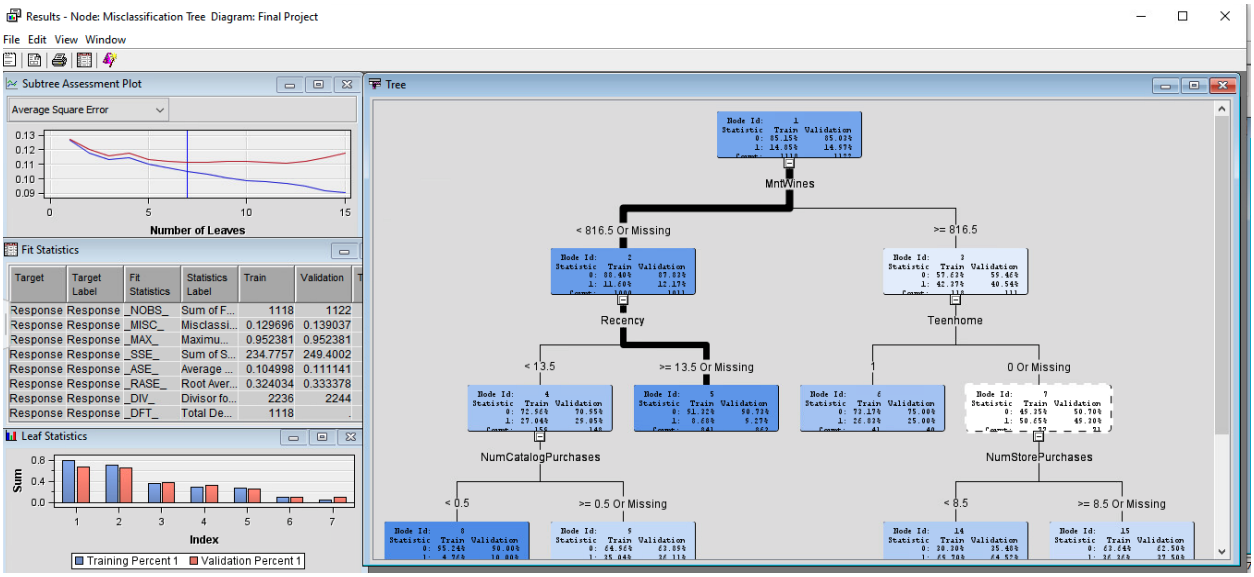


Figure 10. Results of misclassification tree

Average Squared Error (ASE) Tree

The third and final Decision Tree was an Average Squared Error. The following settings were applied before running this model: Method – Assessment, Assessment Measure – Average Squared Error. The resulting Tree has a total of 29 leaves, but only 11 are selected in the leaf statistics. The Validation Misclassification Rate stands at 14.08%; the Validation Average Squared Error stands at 11.02%. The following variables are found to be competing splits out of

the root node in the ASE Tree: *MntWines*, *Recency*, *TeenHome*. Refer to Figure 11 to consult the ASE Tree results window.

Figure 11. Average Squared Error Tree Results Window

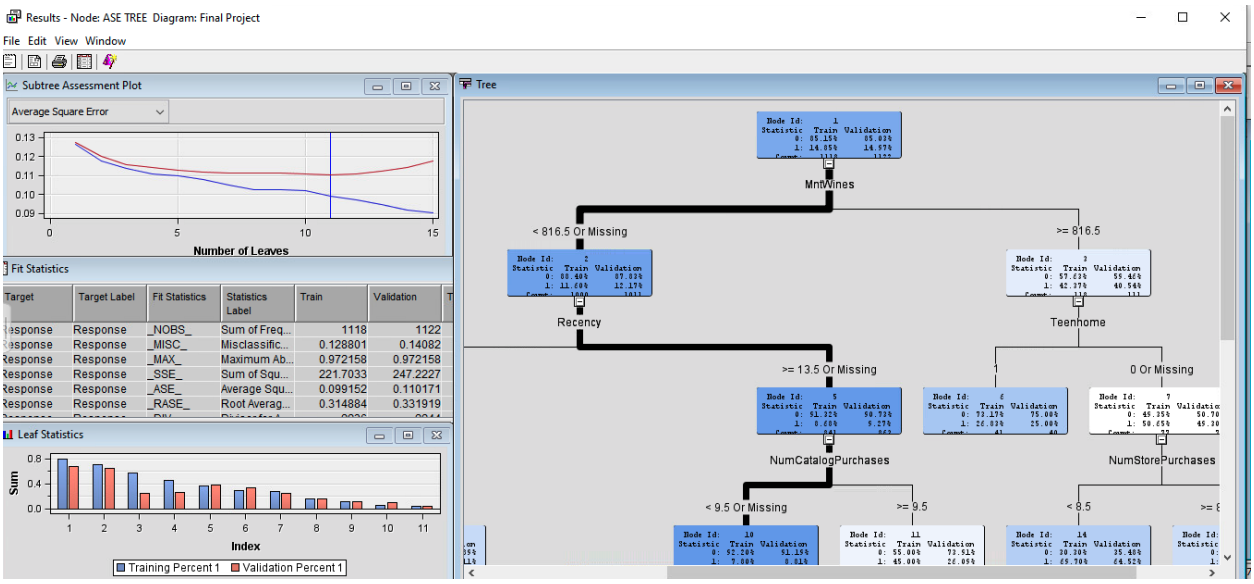


Figure 11. Results of ASE tree

Misclassification Tree – C.C. Misc. Tree

As it will be shown in later sections of this report, a fourth Decision Tree was run using a clean dataset. This node was called Class Collapse Misclassification Tree and was used for completeness and comparison purposes. Refer to Figure 19 for a visualization of the complete diagram.

Decision Trees' Comparison

According to the first three Decision Tree's validation data's misclassification rates and average squared errors, the Misclassification Tree was superior. This tree has the lowest misclassification rate (13.90%) and the lowest number of total leaves (17). The Average Squared Error Tree has the lowest average squared error (11.02%); however, it has a higher misclassification rate and a greater number of leaves. The Leaf Statistics of the three Decision

Trees, show that the Misclassification Tree is superior, with only seven leaves returned as optimal.

Logistic Regression

Data Cleaning

The second modeling technique used was Logistic Regression. As opposed to Decision Trees, Logistic Regressions are more demanding in terms of data quality. Even though incomplete data can be fed to the model, Logistic Regressions do not handle missing values well. The evaluation of a model's resulting function is impossible with missing values, which creates skewed results. Also, missing values create skewed estimates with limited predictive capabilities.

In regards to the 1.07% of missing observations in *income*, an Impute node was used to solve the issue. The Default Input Method: mean, was used to calculate the imputed values. An *Imputed:income* variable was created in addition to an *Imputation Indicator for Income* that shows 1 for imputed values and 0 otherwise. Refer to Figure 5 for a complete visualization of the most important statistics of the dataset.

Fourteen class levels included in the variables *income, NumDealsPurchases, NumWebPurchases, MntFishProducts, MntFruits, MntGoldProducts, MntMeatProducts, MntSweetProducts, and MntWines* returned high skewness measures, refer to Figure 12. Cap and Floor nodes with three Standard Deviations from the Mean were used to limit the skewness in the Interactive Binning and Class Collapse data. Although the Cap and Floor nodes did a decent job at reducing the skew, it was not completely eliminated; 13 class levels maintained a skewness between 1 and 1.96. Taking into consideration that applying logarithmic transformations make the interpretability of models more challenging, it was decided that transformations would be only applied to the variables that returned skewness measures higher than 1, but only for the Interactive Binning data. The Class Collapse data remained unchanged.

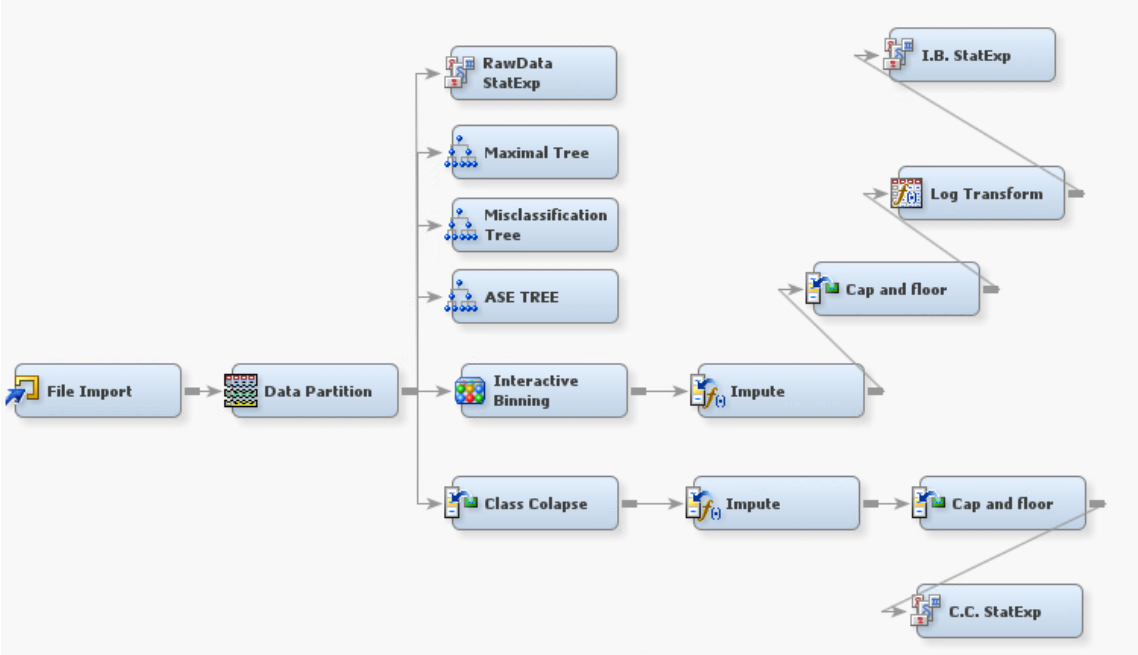
Refer to Figure 13 for a visualization of the data cleaning process that was implemented. Refer to Appendix 4 and Appendix 5 for visualizations of the Class Collapse and Interactive Binning data descriptive statistics.

Figure 12. Skewness Measures

Variable	Skewness	Target Level
NumDealsPurchases	2.5035930	
MntMeatProducts	2.4231660	
MntFruits	2.2374410	
MntSweetProducts	2.2310280	
MntGoldProds	2.2074760	
MntFishProducts	2.0760660	
NumWebPurchases	2.0260320	
NumCatalogPurchases	1.994350	
NumDealsPurchases	1.82491	
MntFruits	1.7546111	
MntSweetProducts	1.5427711	
MntFishProducts	1.3114321	
MntWines	1.2705630	
MntGoldProds	1.1278111	
MntMeatProducts	0.8807851	
NumStorePurchases	0.7475370	
IMP_Income	0.5503610	
Recency	0.5307171	
MntWines	0.4775151	
NumCatalogPurchases	0.4642081	
NumStorePurchases	0.3957971	
NumWebPurchases	0.384491	
NumWebVisitsMonth	0.3214840	
Year_Birth	-0.041640	
Recency	-0.057590	
Year_Birth	-0.169171	
IMP_Income	-0.253161	
NumWebVisitsMonth	-0.326341	

Figure 12. Skewness Measures

Figure 13. Data Cleaning



Logistic Regressions – Interactive Binning (I.B.)

The next step in the elaboration of this project was to run multiple logistic regressions with the following two purposes: 1) creating a model that could potentially explain in the best way possible the interaction between our target and independent variables, and 2) running a model that identifies the best variables to include in a different type of model, like Neural Network. With this in mind, the following logistic regressions were run on the Interactive Binning data: Full regression, forward inclusion, backward exclusion, and stepwise logistic regression. Refer to Figure 16 for a visual of the Interactive Binning Logistic Regressions' summary.

Full Regression (I.B. Full Regression). A logistic regression with the Selection Model set to *none* was run. The lowest Odds Ratio for this model was 0.361 for *GRP_Marital_Status* 1 vs 2, which suggests that being in a "together" *marital_status* as opposed to "alone" reduces the odds of a positive response to the target variable by 63.9%. The highest Odds Ratio stands at 5.931 for *LOG_REP_NumCatalogPurchases*. This ratio suggests that catalog purchases and the target variable have a strong positive relationship. No unusual Odds Ratios were found. Given that at this point this is one of various competing models, only the ratios found on the opposite side of the spectrum shall be discussed. The odds ratios of the best model will be explored in greater detail.

LOG_REP_NumCatalogPurchase had the greatest impact on *response*. This variable returned the highest Standardized Estimate at 0.4078; its Pr>ChiSq statistics was lower than 0.0001. The Average Squared Error and Misclassification Rate for the validation data in this model stand at 0.0940 and 0.1301 respectively.

Forward Inclusion (I.B. Forward Inclusion). A logistic regression with the Selection Model set to *forward*, Selection Criterion set to *validation error* was run. The lowest Odds Ratio of the model stands at 0.381 for *GRP_Marital_Status* 1 vs 2, which suggests that being in a "together" *marital_status* as opposed to "alone" reduces the odds of a positive response to the target variable by 61.9%. The highest Odds Ratio stands at 5.473 for *LOG_REP_NumCatalogPurchases*. This ratio suggests that there is a strong relationship between the logarithmic transformation of the capped value of the number of catalog purchases. No unusual Odds Ratios were found.

GRP_Education appears as the first variable listed after the Intercept in the effects model; *LOG_REP_NumCatalogPurchases* returns the highest Standardized Estimate again. Its Pr>ChiSq is lower than 0.0001. The Average Squared Error and Misclassification Rate for the validation data in this model stand at 0.0937 and 0.1310 respectively.

Backward Exclusion (I.B. Backward Exclusion). A logistic regression with the Selection Model set to *backward*, Selection Criterion set to *validation error* was run. The lowest Odds Ratio of the model stands at 0.366 for *GRP_Marital_Status* 1 vs 2, which suggests that being in a "together" *marital_status* as opposed to "alone" reduces the odds of a positive response to the target variable by 63.4%. The highest Odds Ratio stands at 5.816 for *LOG_REP_NumCatalogPurchases*. As it has been explained for previous models, this ratio suggests there is a strong positive relationship between the independent and target variables.

GRP_Education appears as the first variable listed after the Intercept in the effects model; Similar to previous models, the *REP_NumCatalogPurchases* returned the highest Standardized Estimate and a Pr>ChiSq 0.0001. The Average Squared Error and Misclassification Rate for the validation data in this model stand at 0.0930 and 0.1274 respectively.

Stepwise Regression (I.B. Stepwise). A logistic regression with the Selection Model set to *stepwise*, Selection Criterion set to *validation error* was run. We found that the relevant statistics for this model are identical to those described in the Forward Inclusion model.

Stepwise Regression (I.B. Best Sequence). A second logistic regression with the Selection Model set to *stepwise,* Selection Criterion set to *validation error* was run. However, these regression's selection options were tuned differently. Refer to Figure 14.

At 0.363, *GRP_Marital_Status* 1 vs 2 returned the lowest odds ratio. This suggests that being in a "together" *marital_status* as opposed to "alone" reduces the odds of a positive response to the target variable by 63.7%. The highest Odds Ratio stands at 5.872 for *LOG_REP_NumCatalogPurchases*. Like it has been previously stated, this ratio suggests a positive relation between the independent and target variables. No unusual Odds Ratios were found.

Figure 14. Selection Criteria – "Best Sequence"

Property	Value
Sequential Order	No
Entry Significance Level	1.0
Stay Significance Level	0.5
Start Variable Number	0
Stop Variable Number	0
Force Candidate Effects	0
Hierarchy Effects	Class
Moving Effect Rule	None
Maximum Number of Steps	30

Figure 14. Selection tuning

GRP_Marital_Status appears as the first variable listed after the Intercept in the effects model. Once again, *LOG_REP_NumCatalogPurchases* returns the highest Standardized Estimates and a Pr>ChiSq lower than 0.0001. The Average Squared Error and Misclassification Rate for the validation data in this model stand at 0.0938 and 0.1266 respectively. The Iteration Plot's Misclassification Rate indicator shows that the best model is found in the 16th iteration of

the regression. There are no signs of aggressive overfitting. Refer to Figure 15 for a visualization of the Misclassification Rate plot.

Refer to Figure 16 to get a visualization of the main fit statistics for the regression models discussed in this section.

Figure 15. Iteration Plot – (I.B.) Best Sequence Regression

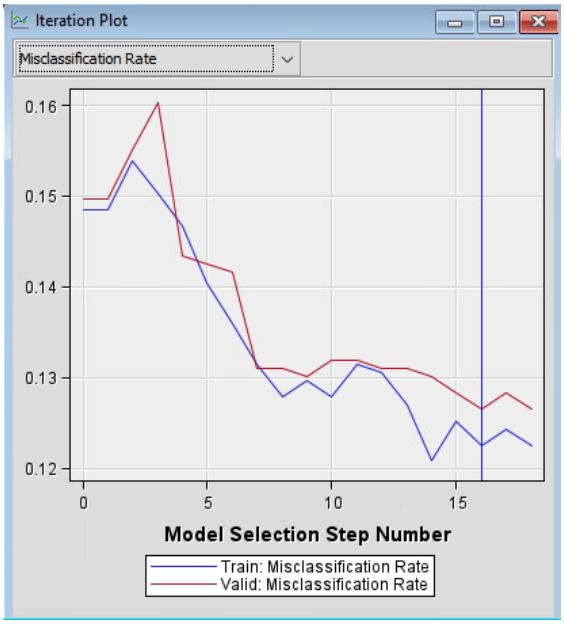


Figure 15. Misclassification rate

Figure 16. Interactive Binning Logistic Regressions' Summary

Model	Lowest Odds Ratio	Highest Odds Ratio	Highest Standardized Estimate	Validation Misclassification Rate
I.B. Full Regression	GRP_Marital_Status 1 vs 2; 0.361	LOG_REP_NumCatalogPurchases; 5.931	LOG_REP_NumCatalogPurchases; 0.4078	13.01%
I.B. Forward Inclusion	GRP_Marital_Status 1 vs 2; 0.381	LOG_REP_NumCatalogPurchases; 5.473	LOG_REP_NumCatalogPurchases; 0.7154	13.10%
I.B. Backward Exclusion	GRP_Marital_Status 1 vs 2; 0.366	LOG_REP_NumCatalogPurchases; 5.816	LOG_REP_NumCatalogPurchases; 0.7409	12.74%
I.B. Stepwise	GRP_Marital_Status 1 vs 2; 0.381	LOG_REP_NumCatalogPurchases; 5.473	LOG_REP_NumCatalogPurchases; 0.7154	13.10%
I.B. Best Sequence	GRP_Marital_Status 1 vs 2; 0.363	LOG_REP_NumCatalogPurchases; 5.872	LOG_REP_NumCatalogPurchases; 0.7450	12.66%

Figure 16. Summary of Interactive Binning Logistic Regressions

Logistic Regression – Replacement (Class Collapse)

Stepwise Regression (C.C. Best Sequence). As it has been mentioned before, the data created by the Class Collapse node was also used to create additional models that include a Misclassification Tree, Neural Networks, and a Logistic Regression. Since the Logistic Regression "I.B. Best Sequence" returned the best accuracy statistics for the models created in the previous section of this report, an identical model was constructed using the Class Collapse data. Refer to Figure 17 for a visualization of the modeling diagram.

Figure 17. Logistic Regression Models

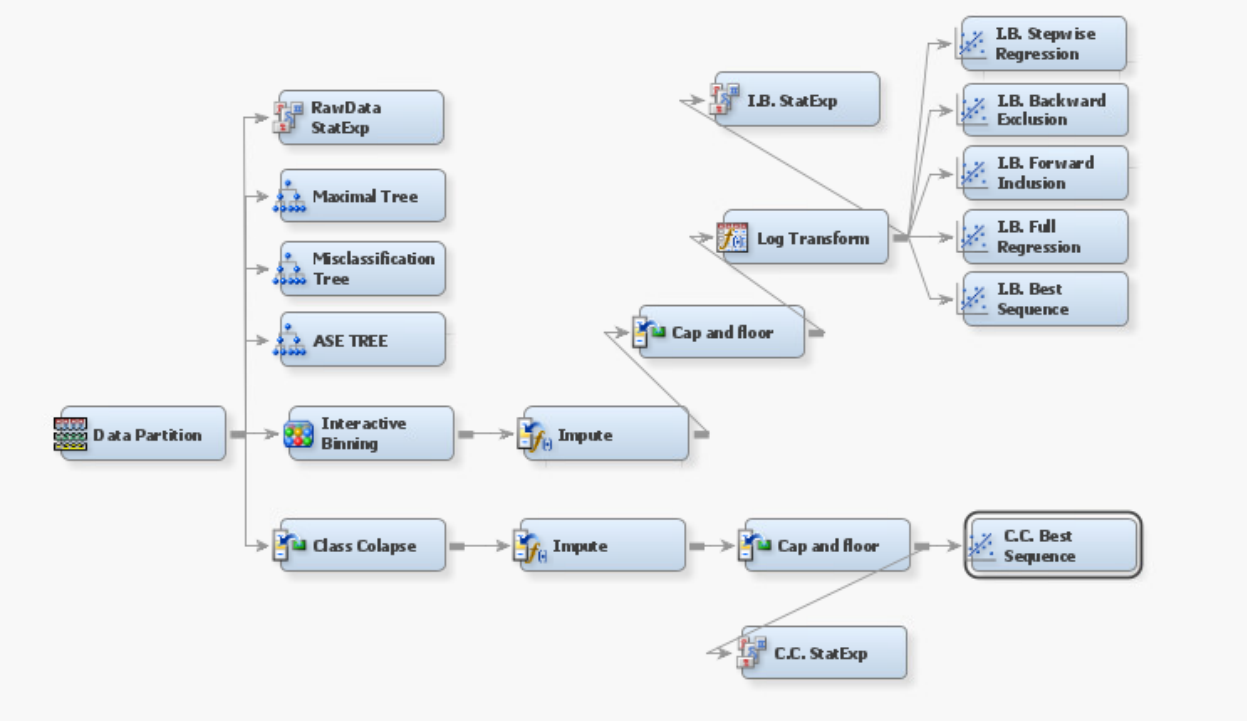


Figure 17. Diagram 3

The lowest Odds Ratio for the C.C. Best Sequence Logistic Regression stands at 0.823 for the *Teenhome* 1 vs 2, which suggests that having 1 teen at home as opposed to 2 reduces the odds of a positive response to the target variable by 17.7%. The highest Odds Ratio stands at 5.472 for the variable *REP_Marital_Status* OT vs TO. This ratio suggests that being in an

"Other" marital status as opposed to "Together" increases the odds of a positive response to the target variable by 447%. No unusual Odds Ratios were found.

REP_IMP_Income appears as the first variable listed after the Intercept in the effects model; *REP_NumWebVisitsMonth* and *REP_NumCatalogPurchases* return the highest Standardized Estimates. The Average Squared Error and Misclassification Rate for the validation data stand at 0.0943 and 0.1267 respectively. The Iteration Plot's Misclassification Rate indicator shows that the best model is found in the 9th iteration of the regression. There are no signs of overfitting. Refer to Figure 18.

Figure 18. Iteration Plot – Class Collapse Best Sequence Regression

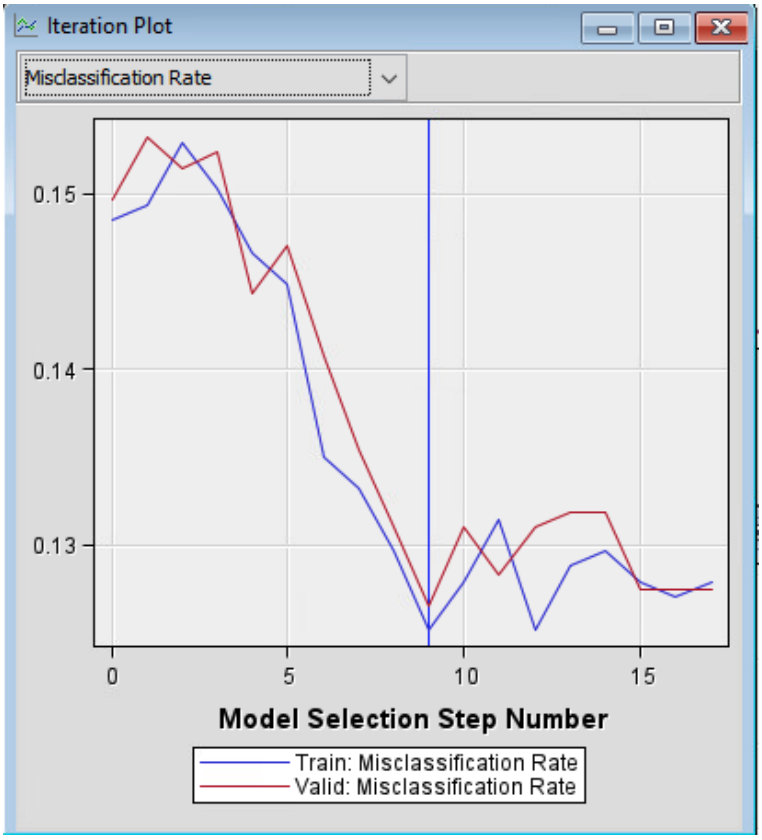


Figure 18. Misclassification rate

Neural Network

Characteristics of Neural Networks

Neural Network was the third modeling technique implemented in this project. Similar to regression models, Neural Networks require a complete dataset to work properly and fall short when dealing with the curse of dimensions. Looking to solve this issue, the I.B. Best Sequence and C.C. Best Sequence Logistic Regression models were used to identify the best variables to feed to the Neural Networks created in this project. As opposed to regression models, Neural Networks are well suited to handle outliers; therefore, a set of Neural Networks were run with data that was not processed by Cap and Floor nodes. Additional sets of Neural Networks were constructed with capped data for completeness reasons.

Neural Networks – Interactive Binning

The first set of Neural Network models was built using the Impute node data in the Interactive Binning "path" of this project. An exhaustive parameter optimization process was performed manually in order to obtain the best accuracy statistics. However, only the three most relevant Neural Networks were documented. After testing the first few models with different combinations of hidden units and iterations, it became fairly evident that during the first 10 to 20 iterations, most models became aggressively overfitted. Following this rationale, it was decided that the maximum number of iterations that would be tested for further models would be 50, unless new evidence suggested the opposite. Increasing the number of iterations seemed unnecessary and only a waste of computational resources. On the other hand, modifying computational power by changing the number of hidden units for different models proved to have a direct impact on model performance. The first three Neural Networks documented had 4, 5, and 6 hidden units. All models were built with 50 iterations.

I.B. 4HU 50 IT. The twentieth iteration of this model was found to have the best fit statistics. The Average Squared Error and Misclassification Rate for the validation data stand at 0.0941 and 0.1319 respectively. This model was not trained before running the node.

Figure 19. Diagram – NN Interactive Binning

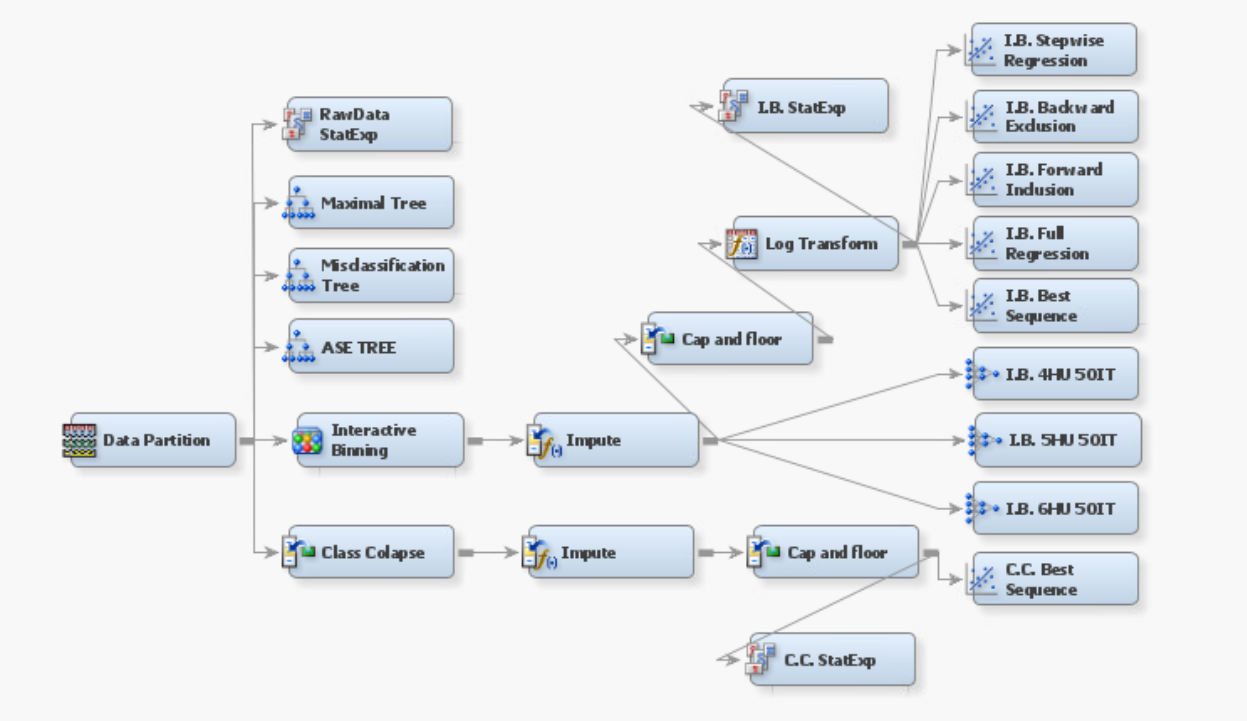


Figure 19. Diagram 4

I.B. 5HU 50 IT. The twenty-third iteration of this model was found to have the best fit statistics. The Average Squared Error and Misclassification Rate for the validation data stand at 0.0929 and 0.1274 respectively. This model was not trained before running the node.

I.B. 6HU 50 IT. Similar to the previous model, the twenty-third iteration the best fit statistics were found during the twenty-third iteration. The Average Squared Error and Misclassification Rate for the validation data stand at 0.0912 and 0.1310 respectively. This model was not trained before running the node.

The model that returned the best fit statistics for this set of Neural Networks was the I.B. 5HU 50IT. After reviewing the iteration plot for this model, it was found that there was a case of overfitting after the twenty-third iteration. Refer to Figure 20 for a visualization of the results panel for the best Neural Network model.

Figure 20. Results Panel I.B. Neural Network 5HU 50 IT

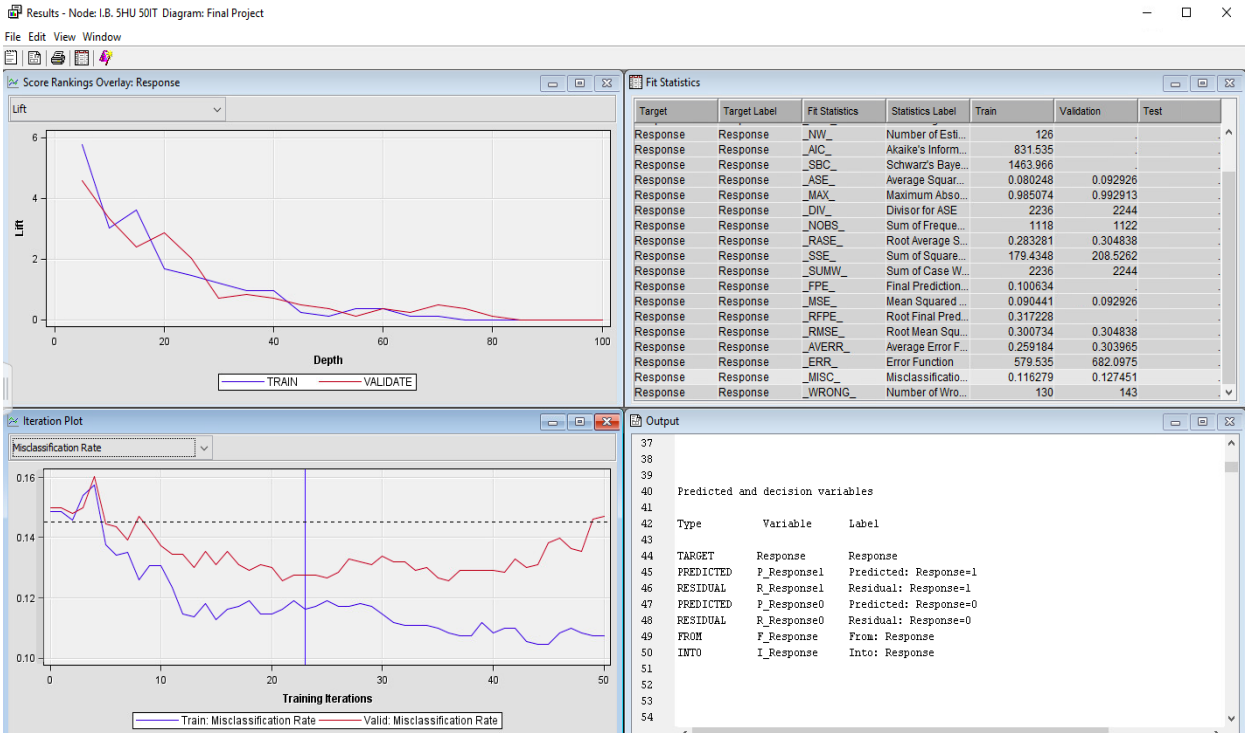


Figure 20. Results window

Neural Network – Class Collapse

A new set of Neural Network models was constructed using the capped data from the Class Collapse "path" of this model. Similar to the I.B. Neural Networks a number of different combinations of hidden units and iterations was tested, but only the most relevant models were documented. Refer to Figure 21 to find the results of these models.

The C.C. 5 HU 50IT returned the best fit statistics. Although the C.C. 3 HU 50IT and C.C. 5 HU 50IT have the same Misclassification Rate, it was found that the latter model was

considered superior since it returned the lowest Average Squared Error of the two. It must be noted that the two additional hidden units of the chosen model make it more complex; however, a superior performance may be expected from this model. These models were not trained before the nodes were run. Refer to Figure 22 for a visualization of the models created up to this point.

Figure 22. Complete Neural Network Nodes

Figure 21. C.C. Neural Network Results

Maximum Iterations	Hidden Units	Average Squared Error	Misclasssification Rate	Iterations
50	3	9.57%	12.57%	3
50	5	9.10%	12.57%	7
50	7	9.29%	13.63%	7
50	8	9.12%	12.30%	8

Figure 21. C.C. Neural Network Results' Summary

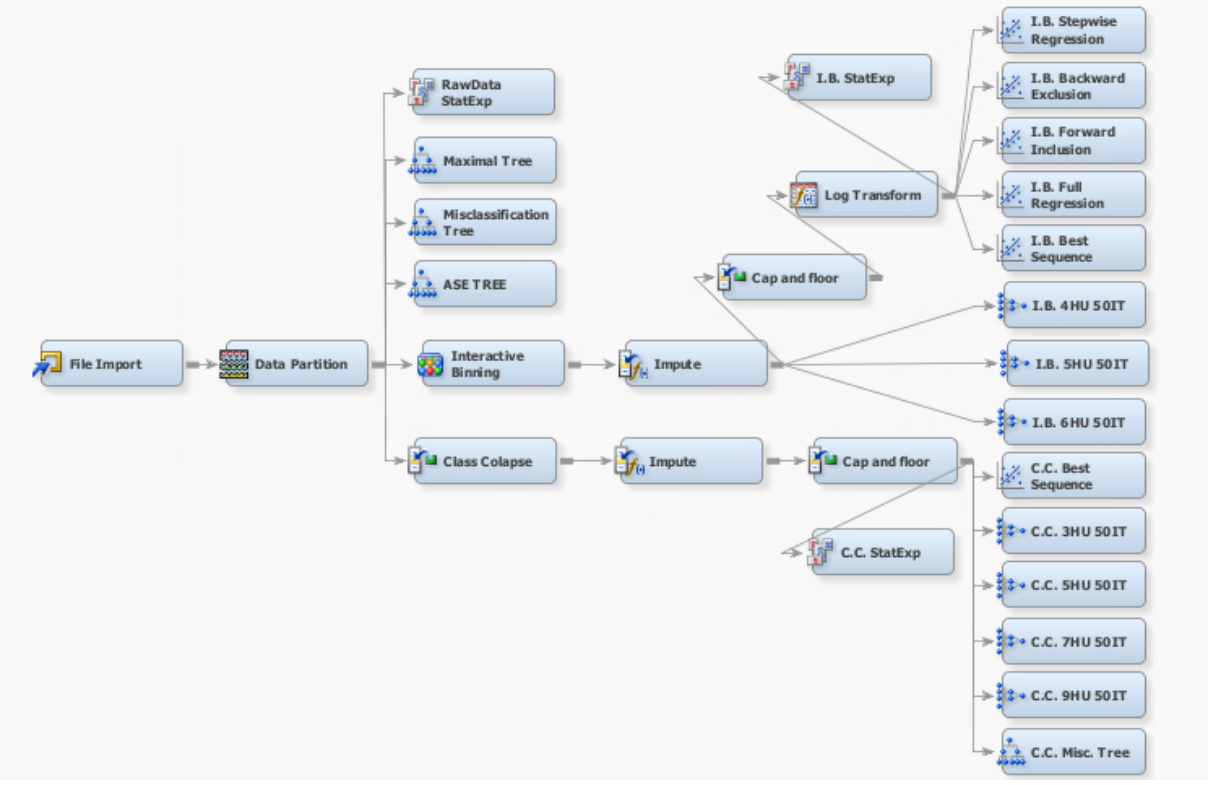


Figure 22. Diagram 5

Neural Network – I.B. Best Sequence / C.C. Best Sequence

As it has been explained before, implementing a variable selection technique has the potential of improving Neural Networks' performance. Therefore, a third and final set of Neural Network models was constructed based on the I.B. Best Sequence and C.C. Best Sequence Logistic Regressions as these models returned the accuracy statistics. Similar to previous sets of Neural Networks, different combinations of hidden units and iterations were tested, but only the most relevant models were documented. None of these models were trained before running. Refer to Figure 23 to find a summary of these models.

Figure 23. I.B. / C.C. Best Sequence – Neural Networks Summary

Data	Maximum Iterations	Hidden Units	Average Squared Error	Misclasssification Rate	Iterations
I.B. Best Sequence	50	5	9.36%	12.39%	28
I.B. Best Sequence	50	7	8.92%	13.10%	20
I.B. Best Sequence	50	9	9.29%	12.74%	12
I.B. Best Sequence	50	11	9.74%	11.76%	22
C.C. Best Sequence	50	3	9.29%	13.01%	5
C.C. Best Sequence	50	5	9.07%	12.21%	6
C.C. Best Sequence	50	6	8.73	11.76%	8
C.C. Best Sequence	50	7	9.33%	13.01%	13

Figure 23. I.B. / C.C. Neural Network Results

Model Comparison

A Model Comparison node was used to compare the models created during the different stages of this project. The node was run using all the default settings, including Assessment Report and Selection Criteria. Refer to Figure 24 for a visualization of the final diagram.

According to the results of the Model Comparison Node, the C.C. B.S. 6HU 50IT model returned the lowest validation Misclassification Rate, which was selected as the selection criteria at the beginning of this project. The validation Misclassification Rate, validation ROC Index, and validation Average Squared Error stand at 0.1176, 0.875, and 0.0874 respectively. Refer to Figure 25 to consult the fit statistics for all the models.

Figure 24. Final Diagram

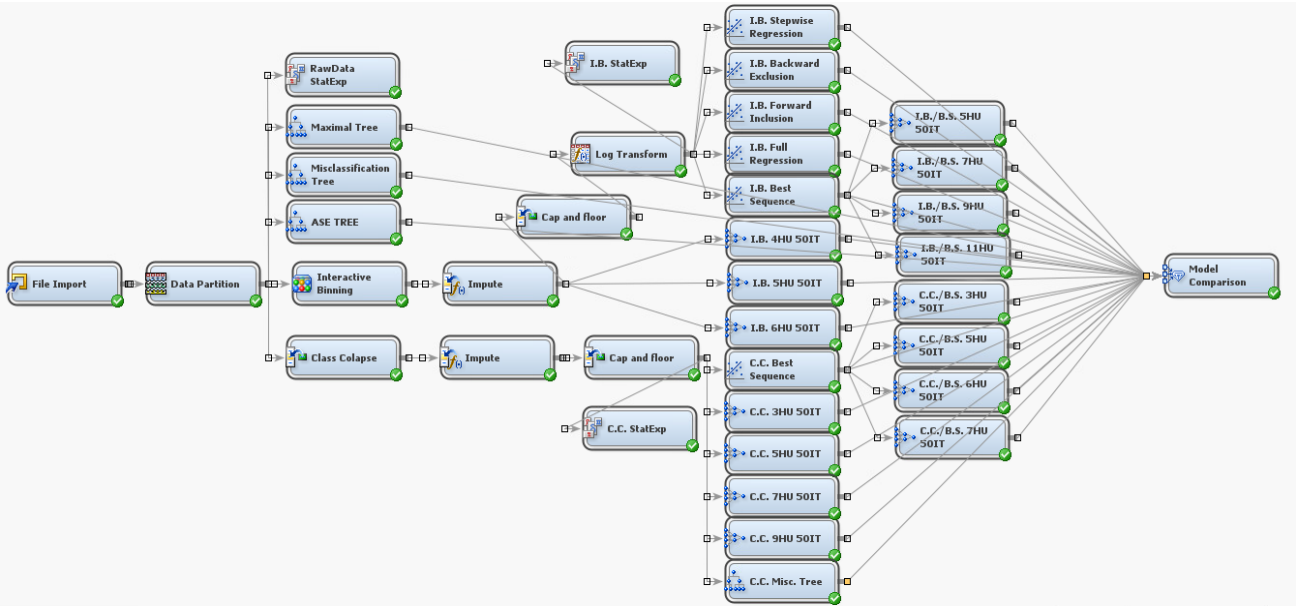


Figure 24. Diagram 6

Figure 25. Model Comparison Fit Statistics

Results - Node: Model Comparison Diagram: Final Project

File Edit View Window

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate ▲	Valid: Roc Index	Valid: Average Squared Error	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	Tra De Fre
Y	Neural24	Neural24	C.C./B.S. 6HU 50IT	Response	Response	0.117647	0.875	0.087372	1118	0.112701	0.992061	185.8633	0.083123	0.288311	2236	
	Neural23	Neural23	I.B./B.S. 11HU 50IT	Response	Response	0.117647	0.878	0.08741	1118	0.117174	0.99506	190.7185	0.085295	0.292052	2236	
	Neural21	Neural21	C.C./B.S. 5HU 50IT	Response	Response	0.122103	0.866	0.090724	1118	0.11449	0.991256	196.0749	0.08769	0.296125	2236	
	Neural27	Neural27	C.C. 9HU 50IT	Response	Response	0.122995	0.864	0.091239	1118	0.124329	0.990506	203.84	0.091163	0.301932	2236	
	Neural22	Neural22	I.B./B.S. 5HU 50IT	Response	Response	0.123886	0.863	0.093557	1118	0.111807	0.98613	183.3013	0.081977	0.286317	2236	
	Neural17	Neural17	C.C. 5HU 50IT	Response	Response	0.125668	0.871	0.091328	1118	0.134168	0.991961	207.3213	0.09272	0.304499	2236	
	Neural2	Neural2	C.C. 3HU 50IT	Response	Response	0.125668	0.85	0.095688	1118	0.127013	0.960535	216.5555	0.096849	0.311207	2236	
	Reg6	Reg6	I.B. Best Sequence	Response	Response	0.12656	0.872	0.093847	1118	0.12254	0.987964	210.5142	0.094148	0.306835	2236	
	Reg5	Reg5	C.C. Best Sequence	Response	Response	0.12656	0.858	0.094363	1118	0.125224	0.990632	219.1548	0.098012	0.313069	2236	
	Neural11	Neural11	I.B./B.S. 9HU 50IT	Response	Response	0.127451	0.872	0.092899	1118	0.121646	0.976677	195.5416	0.087452	0.295722	2236	
	Neural28	Neural28	I.B. 5HU 50IT	Response	Response	0.127451	0.864	0.092926	1118	0.116279	0.985074	179.4348	0.080248	0.283281	2236	
	Reg3	Reg3	I.B. Backward Exclusion	Response	Response	0.127451	0.872	0.093063	1118	0.121646	0.986418	211.6746	0.094667	0.307679	2236	
	Neural19	Neural19	C.C./B.S. 3HU 50IT	Response	Response	0.130125	0.86	0.092908	1118	0.125224	0.992532	209.3254	0.093616	0.305967	2236	
	Neural26	Neural26	C.C./B.S. 7HU 50IT	Response	Response	0.130125	0.869	0.0933	1118	0.121646	0.98696	196.5158	0.087887	0.296458	2236	
	Reg	Reg	I.B. Full Regression	Response	Response	0.130125	0.871	0.094036	1118	0.126118	0.98829	210.0981	0.093962	0.306532	2236	
	Neural10	Neural10	I.B./B.S. 7HU 50IT	Response	Response	0.131016	0.872	0.089292	1118	0.109123	0.995001	185.8146	0.083101	0.288273	2236	
	Neural29	Neural29	I.B. 6HU 50IT	Response	Response	0.131016	0.875	0.09112	1118	0.117174	0.971241	194.9014	0.087165	0.295238	2236	
	Reg2	Reg2	I.B. Forward Inclusion	Response	Response	0.131016	0.863	0.093727	1118	0.127907	0.98804	215.1877	0.096238	0.310222	2236	
	Reg4	Reg4	I.B. Stepwise Regression	Response	Response	0.131016	0.863	0.093727	1118	0.127907	0.98804	215.1877	0.096238	0.310222	2236	
	Neural14	Neural14	I.B. 4HU 50IT	Response	Response	0.131907	0.866	0.094131	1118	0.121646	0.996708	199.6602	0.089293	0.29882	2236	
	Neural30	Neural30	C.C. 7HU 50IT	Response	Response	0.136364	0.865	0.092867	1118	0.132379	0.985924	209.6119	0.093744	0.306177	2236	
	Tree6	Tree6	Misclassification Tree	Response	Response	0.139037	0.688	0.111141	1118	0.129696	0.952381	234.7757	0.104998	0.324034	2236	
	Tree8	Tree8	C.C. Misc. Tree	Response	Response	0.139037	0.688	0.111141	1118	0.129696	0.952381	234.7757	0.104998	0.324034	2236	
	Tree7	Tree7	ASE TREE	Response	Response	0.14082	0.755	0.110171	1118	0.128801	0.972158	221.7033	0.099152	0.314884	2236	
	Tree5	Tree5	Maximal Tree	Response	Response	0.154189	0.742	0.11752	1118	0.11449	0.972158	202.1947	0.090427	0.300711	2236	

Figure 25. Results window

Conclusion

The. C.C. B.S. 6HU 50IT was the best model at describing the relationship between *response* and the various explanatory variables tested. However, a few variables stood out in the

vast majority of the models built in this project. For example, *REP_NumWebVisitsMonth*, *REP_NumCatalogPurchases*, and *LOG_REP_NumCatalogPurchases* appeared to have the highest impact over the target variable when assessed in the different regression models. *GRP_Marital_Status* 1 vs 2 and *LOG_REP_NumCatalogPurchases* returned the highest and lowest odds ratio for all the Interactive Binning data models. *Teenhome* 0 vs 1 and *REP_Marital_Status* OT vs TO returned relevant odds ratios for the Class Collapse data. *MntWines* and *Teenhome* also appeared as variables of interest in every Decision Tree model. Other variables of interest are *REP_IMP_Income* and *REP_Education*.

The C.C. B.S. 6HU 50IT model predicts that 10 observations have a probability of 85% or higher of returning a positive response to the target variable, refer to Figure 21. Taking into account only the variables that consistently appeared to be relevant, the 10 individuals with the highest positive response prediction probabilities show the following:

- *REP_NumCatalogPurchases*: 90% of observations had a number of catalog purchases above the mean. For people who had a positive response to the selected target variable, the mean number of catalog purchases almost doubled in comparison to those who had a negative response.
- *REP_NumWebVisitsMonth*: 50% of the observations had a number of monthly web visits equal or higher than the mean.
- *REP_Marital_Status*: 70% of the observations fall in the SI and SI/M classes even when 63.5% of the total observations fall in the mode class TO.
- *REP_Teenhome*: 100% of the observations have zero teens at home.
- *REP_MntWines*: 70% of the observations have an expenditure on wine above the mean.
- *REP_IMP_ Income*: 100% of the observations fall above the mean.

- *REP_Education*: 90% of the observations are college graduates.

As an additional note, it was also found that 80% of the observations in this group are younger than average and have not kids at home.

Figure 26. Neural Network Prediction Statistics

EMWS1.Neural24_VALIDATE

	Predicted: Response=1	Replacement: NumCatalogPurchases	Replacement: NumWebVisitsMonth	Replacement: Marital_Status	Teenhome	Replacement: MntWines	Replacement: Imputed: Income	Replacement: Education
1	0.9624537449580156	10.0	5.0	TO	0	179.0	86836.0	PG
2	0.9606257664179574	11.071202075	1.0	TO	0	1.0	115897.16873	GRAD
3	0.953595669684951	11.0	3.0	SI/M	0	395.0	79946.0	GRAD
4	0.8992999283973858	2.0	5.0	SI	0	709.0	81698.0	UG
5	0.889581914341593	5.0	6.0	SI	0	676.0	76467.0	GRAD
6	0.8804995767418657	5.0	9.0	TO	0	163.0	88097.0	PG
7	0.8750807509064714	6.0	2.0	SI/M	0	667.0	78789.0	PG
8	0.8669610677565089	3.0	1.0	SI	0	1285.0	95169.0	PG
9	0.8616924322568357	3.0	4.0	SI/M	0	464.0	76982.0	GRAD
10	0.8568921652777681	4.0	9.0	SI	0	1311.6863204	68126.0	PG

Figure 26. Neural Network assessment

Recommendation

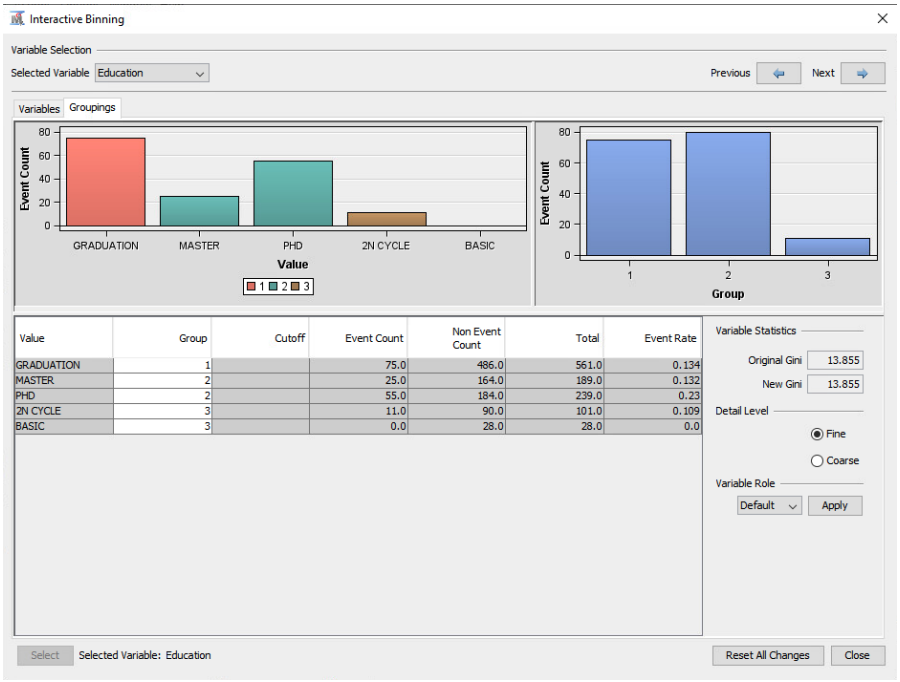
As it has been noted, having an income higher than average, a college degree, not being in a relationship (or together), not having teens or kids at home, being younger than average, and displaying a preference for catalog purchases and wine products are traits that a have positive impact on customers' response to the last marketing campaign and the selected target variable.

Based on the above findings, the following recommendation would be delivered to a potential enterprise customer:

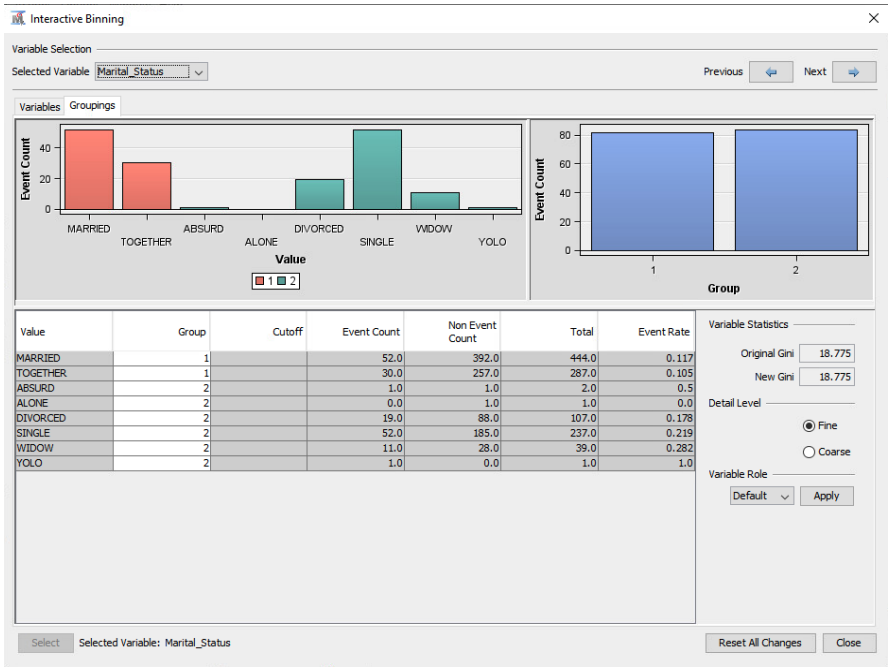
Focus on young, college educated, customers who are not in relationships. The preferred customer should have an income higher-than-average and no teens or kids at home. Lastly, the preferred customer should have a preference for wines and catalog purchases.

Appendix

Appendix 1. Interactive Binning – Education



Appendix 2. Interactive Binning – Marital Status



Appendix 3. Class Collapse – Education & Marital Status

Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type
Complain	0		1107	N
Complain	1		11	N
Complain	_UNKNOWN_	**_DEFAULT_**	.	N
Education	Graduation	**GRAD**	561	C
Education	PhD	**PG**	239	C
Education	Master	**PG**	189	C
Education	2n Cycle	**UG**	101	C
Education	Basic	**UG**	28	C
Education	_UNKNOWN_	**_DEFAULT_**	.	C
Kidhome	0		644	N
Kidhome	1		450	N
Kidhome	2		24	N
Kidhome	_UNKNOWN_	**_DEFAULT_**	.	N
Marital_Status	Married	**TO**	444	C
Marital_Status	Together	**TO**	287	C
Marital_Status	Single	**SI**	237	C
Marital_Status	Divorced	**SI/M**	107	C
Marital_Status	Widow	**SI/M**	39	C
Marital_Status	Absurd	**OT**	2	C
Marital_Status	Alone	**OT**	1	C
Marital_Status	YOLO	**OT**	1	C
Marital_Status	_UNKNOWN_	**_DEFAULT_**	.	C
Response	0		952	N
Response	1		166	N
Response	_UNKNOWN_	**_DEFAULT_**	.	N
Teenhome	0		555	N
Teenhome	1		543	N
Teenhome	2		20	N
Teenhome	_UNKNOWN_	**_DEFAULT_**	.	N

OK Cancel

Appendix 4. Descriptive Statistics – Class Collapse Final Data

Target	Target Level	Variable	Median	Skewness	Standard Deviation	Mean	Missing	Non Missing	Minimum	Maximum
Response	0	REP_MntMeatProducts	57	1.804984	181.5667	137.6626	0	952	0	790.9765
Response	1	REP_MntMeatProducts	145	0.744615	253.0235	267.1132	0	166	1	790.9765
Response	0	REP_MntWines	135	1.259372	303.8293	264.9577	0	952	0	1311.686
Response	1	REP_MntWines	452	0.393244	417.4153	505.1472	0	166	1	1311.686
Response	0	REP_NumCatalogPurchases	1	1.26183	2.541789	2.297344	0	952	0	11.0712
Response	1	REP_NumCatalogPurchases	4	0.464208	3.134609	4.277108	0	166	0	11
Response	0	REP_MntFishProducts	11	1.884025	49.03311	33.947	0	952	0	199.6925
Response	1	REP_MntFishProducts	25	1.169907	62.78353	53.10984	0	166	0	199.6925
Response	0	REP_MntGoldProds	22	1.788067	46.45216	39.87451	0	952	0	198.9765
Response	1	REP_MntGoldProds	39	1.042142	57.06195	62.14401	0	166	0	198.9765
Response	0	REP_MntSweetProducts	7	1.924897	39.18456	25.74281	0	952	0	155.3407
Response	1	REP_MntSweetProducts	19	1.369765	43.87958	36.77738	0	166	0	155.3407
Response	0	REP_MntFruits	7	1.964574	35.93849	23.8494	0	952	0	143.5626
Response	1	REP_MntFruits	19	1.482697	39.28251	33.61409	0	166	0	143.5626
Response	0	REP_NumWebPurchases	3	0.81533	2.610352	3.886732	0	952	0	12.72296
Response	1	REP_NumWebPurchases	5	0.38449	2.649571	5.26506	0	166	0	11
Response	0	REP_Recency	51	-0.05759	28.47586	50.49055	0	952	0	99
Response	1	REP_Recency	30	0.530717	29.02006	36.86747	0	166	0	99
Response	0	REP_IMP_Income	50898	0.123752	19982.68	50657.55	0	952	1730	115897.2
Response	1	REP_IMP_Income	62972	-0.25316	22258.22	60110.54	0	166	7500	105471
Response	0	REP_NumStorePurchases	5	0.747537	3.274002	5.715336	0	952	0	13
Response	1	REP_NumStorePurchases	6	0.395797	3.144825	6.23494	0	166	2	13
Response	0	REP_NumDealsPurchases	2	1.454776	1.697628	2.297498	0	952	0	8.324457
Response	1	REP_NumDealsPurchases	1	1.426589	2.022413	2.373173	0	166	0	8.324457
Response	0	REP_NumWebVisitsMonth	6	-0.17602	2.327333	5.269586	0	952	0	12.60767
Response	1	REP_NumWebVisitsMonth	6	-0.32634	2.529021	5.307229	0	166	1	9
Response	0	REP_Year_Birth	1970	-0.04164	11.47016	1968.935	0	952	1941	1996
Response	1	REP_Year_Birth	1970	-0.16917	12.97875	1969.157	0	166	1943	1995

Appendix 5. Descriptive Statistics – Binning Final Data

Target	Target Level	Variable	Median	Skewness	Standard Deviation	Mean	Missing	Non Missing	Minimum	Maximum
Response	0	LOG_REP_NumCatalogPurchases	0.693147	0.208394	0.74723	0.91484	0	952	0	2.490823
Response	1	LOG_REP_NumCatalogPurchases	1.609438	-0.40495	0.691461	1.453334	0	166	0	2.484907
Response	0	REP_NumWebPurchases	3	0.81533	2.610352	3.886732	0	952	0	12.72296
Response	1	REP_NumWebPurchases	5	0.38449	2.649571	5.26506	0	166	0	11
Response	0	REP_Recency	51	-0.05759	28.47586	50.49055	0	952	0	99
Response	1	REP_Recency	30	0.530717	29.02006	36.86747	0	166	0	99
Response	0	LOG_REP_MntFruits	2.079442	0.138834	1.545141	2.157799	0	952	0	4.973713
Response	1	LOG_REP_MntFruits	2.995732	-0.4585	1.51295	2.711969	0	166	0	4.973713
Response	0	LOG_REP_MntSweetProducts	2.079442	0.14836	1.590963	2.176597	0	952	0	5.052038
Response	1	LOG_REP_MntSweetProducts	2.995732	-0.36902	1.62801	2.685052	0	166	0	5.052038
Response	0	LOG_REP_MntMeatProducts	4.060443	-0.0921	1.528317	3.972641	0	952	0	6.674532
Response	1	LOG_REP_MntMeatProducts	4.983607	-0.62941	1.426268	4.882499	0	166	0.693147	6.674532
Response	0	LOG_REP_MntFishProducts	2.484907	-0.02563	1.626269	2.465363	0	952	0	5.301774
Response	1	LOG_REP_MntFishProducts	3.258097	-0.38283	1.698399	2.980165	0	166	0	5.301774
Response	0	LOG_REP_MntWines	4.912655	-0.47201	1.779298	4.52735	0	952	0	7.179831
Response	1	LOG_REP_MntWines	6.115892	-1.13787	1.661586	5.466526	0	166	0.693147	7.179831
Response	0	LOG_REP_MntGoldProds	3.135494	-0.33459	1.280842	3.035092	0	952	0	5.2982
Response	1	LOG_REP_MntGoldProds	3.688879	-1.08704	1.23649	3.612082	0	166	0	5.2982
Response	0	REP_IMP_Income	50898	0.123752	19982.68	50657.55	0	952	1730	115897.2
Response	1	REP_IMP_Income	62972	-0.25316	22258.22	60110.54	0	166	7500	105471
Response	0	REP_NumStorePurchases	5	0.747537	3.274002	5.715336	0	952	0	13
Response	1	REP_NumStorePurchases	6	0.395797	3.144825	6.23494	0	166	2	13
Response	0	LOG_REP_NumDealsPurchases	1.098612	0.420577	0.459538	1.081732	0	952	0	2.232641
Response	1	LOG_REP_NumDealsPurchases	0.693147	0.420371	0.532831	1.066427	0	166	0	2.232641
Response	0	REP_NumWebVisitsMonth	6	-0.17602	2.327333	5.269586	0	952	0	12.60767
Response	1	REP_NumWebVisitsMonth	6	-0.32634	2.529021	5.307229	0	166	1	9
Response	0	REP_Year_Birth	1970	-0.04164	11.47016	1968.935	0	952	1941	1996
Response	1	REP_Year_Birth	1970	-0.16917	12.97875	1969.157	0	166	1943	1995

References

Patel, A. (2021). Customer Personality Analysis (Version V1) [Data set]. Kaggle. https://www.kaggle.com/imakash3011/customer-personality-analysis