Final Project

Bridget Bremenour, Zeynep Cetin, Mustafa Mansour

5/2/2023

Superstore Data

Introduction

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Id	Numeric	5	0	Unique ID for e...	None	None	8	Right	Scale
2	Year_Birth	Numeric	7	2	Year of birth	None	None	8	Right	Scale
3	Education	String	10	0	Education level	None	None	10	Left	Nominal
4	Marital_Stat...	String	8	0	Marital status	None	None	8	Left	Nominal
5	Income	Numeric	8	2	Yearly income	None	None	8	Right	Scale
6	Kidhome	Numeric	4	2	Number of kids...	None	None	8	Right	Ordinal
7	Teenhome	Numeric	4	2	Number of tee...	None	None	8	Right	Ordinal
8	Dt_Customer	Date	10	0	Date of custom...	None	None	13	Right	Scale
9	Month	Numeric	8	2	Months since c...	None	None	8	Right	Scale
10	Recency	Numeric	5	2	Number of day...	None	None	8	Right	Scale
11	MntWines	Numeric	7	2	Amount spent ...	None	None	8	Right	Scale
12	MntFruits	Numeric	6	2	Amount spent ...	None	None	8	Right	Scale
13	MntMeatPro...	Numeric	7	2	Amount spent ...	None	None	8	Right	Scale
14	MntFishPro...	Numeric	6	3	Amount spent ...	None	None	8	Right	Scale
15	MntSweetPr...	Numeric	6	2	Amount spent ...	None	None	8	Right	Scale
16	MntGoldPro...	Numeric	6	2	Amount spent ...	None	None	8	Right	Scale
17	NumDeals...	Numeric	5	2	Number of pur...	None	None	8	Right	Scale
18	NumWebP...	Numeric	5	2	Number of pur...	None	None	8	Right	Scale
19	NumCatalo...	Numeric	5	2	Number of pur...	None	None	8	Right	Scale
20	NumStoreP...	Numeric	5	2	Number of pur...	None	None	8	Right	Scale
21	NumWebVi...	Numeric	5	2	Number of web...	None	None	8	Right	Scale
22	Response	Numeric	4	0	Accepted the of...	{0, No}...	None	8	Right	Nominal
23	Complain	Numeric	4	2	Complain	None	None	8	Right	Nominal

This data set comes from a big superstore that wants to run an end-of-the-year promotion discounting a gold membership. This promotion will only be sent to existing customers, and we want to build the model that best predicts which customers are most likely to response to send out less promotion items and still get a high rate of response.

There are 19 possible predictors, these are:

- Year of Birth (Numeric)
- Education (Nominal): High-school, Associate, Bachelor, Masters, PhD.
- Marital_Status (Nominal): Divorced, Single, Married, Together, Widowed.
- Income (Numeric): Yearly income ($)
- Kidhome (Numeric): Number of kids at home
- Teenhome (Numeric): Number of teenagers at home
- Dt_Customer (Date): Date that customer has joined
- Months (Numeric): Months since the customer has joined.
- Recency (Numeric): Number of days since last purchase
- MntWines (Numeric): amount of money spent on wine ($)
- MntFruits (Numeric): amount of money spent on fruits ($)
- MntMeatProd (Numeric): amount of money spent on meat products ($)
- MntFishProd (Numeric): amount of money spent on fish products ($)
- MntSweetProd (Numeric): amount of money spent on sweet products ($)
- MntGoldProd (Numeric): amount of money spent on gold products ($)
- NumDealsPurchase (Numeric): Number of purchases made with discount
- NumWebPurchase (Numeric): Number of purchases made on the website

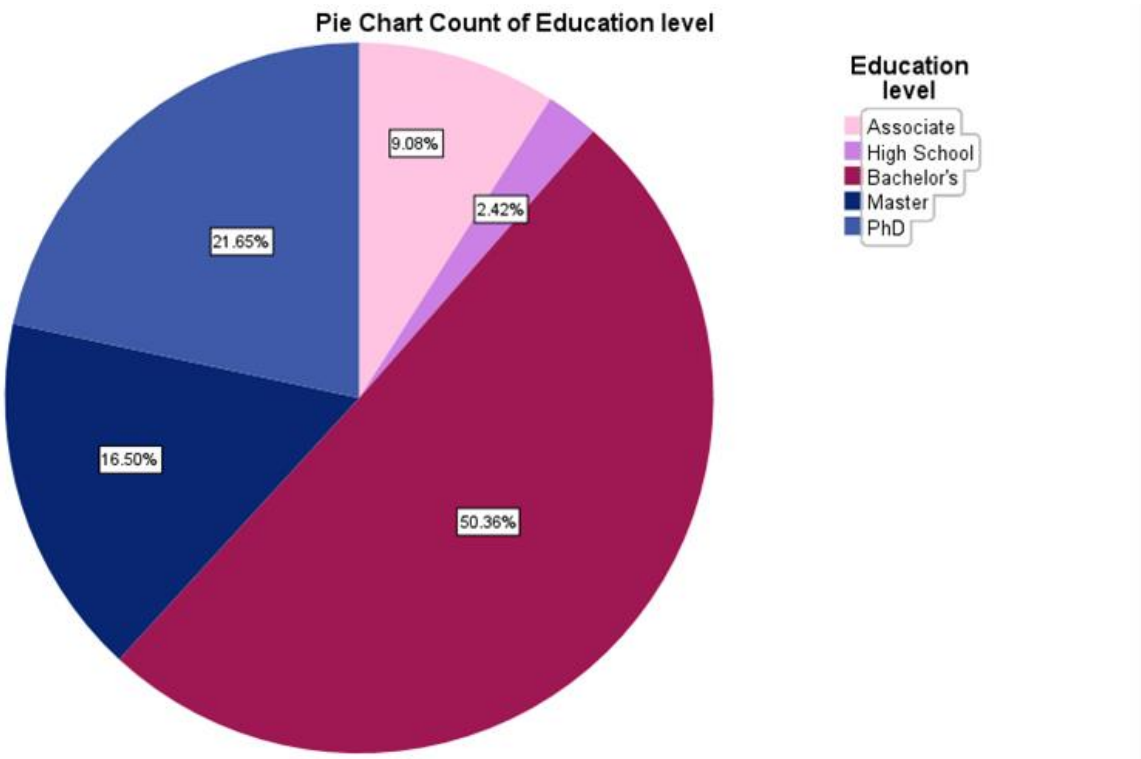
- NumCatalogPurchase (Numeric): Number of purchases made by mail
- NumStorePurchase (Numeric): Number of purchases made in store
- NumWebVisit (Numeric): Number of visits made to the website
- Response (Binary): Responded to the promotion (Yes = 1, No = 0)

Descriptive Statistics

The majority of the variables in this dataset are continuous. The following table shows summary statistics of the continuous variables. There are no apparent outliers or missing values. The highest average amount spent on a product was $303.94 spent on wine. The maximum value spent on product was $1,725.00 spent on meat.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Year of birth	2240	1893.00	1996.00	1968.8058	11.98407
Yearly income	2216	1730.00	666666.00	52247.2514	25173.07666
Number of kids at home	2240	.00	2.00	.4442	.53840
Number of teens at home	2240	.00	2.00	.5062	.54454
Months since customer joined	2240	99.00	134.00	115.8728	7.68590
Number of days since last purchase	2240	.00	99.00	49.1094	28.96245
Amount spent on wine purchases last 2 years	2240	.00	1493.00	303.9357	336.59739
Amount spent on fruit purchases last 2 years	2240	.00	199.00	26.3022	39.77343
Amount spent on meat product purchases last 2 years	2240	.00	1725.00	166.9500	225.71537
Amount spent on fish product purchases last 2 years	2240	.000	259.000	37.52545	54.628979
Amount spent on sweet product purchases last 2 years	2240	.00	263.00	27.0629	41.28050
Amount spent on gold product purchases last 2 years	2240	.00	362.00	44.0219	52.16744
Number of purchases made with discount	2240	.00	15.00	2.3250	1.93224
Number of purchases using web	2240	.00	27.00	4.0848	2.77871
Number of purchases using catalog (Mail)	2240	.00	28.00	2.6621	2.92310
Number of purchases in store	2240	.00	13.00	5.7902	3.25096
Number of web visits last month	2240	.00	20.00	5.3165	2.42665
Accepted the offer in the last campaign	2240	0	1	.15	.356
Valid N (listwise)	2216				



The pie chart was built based on the education levels of the existing customers. We can see that more than half (50.36%) of the existing customers have at least a bachelor's degree, and the least amount of the customers have a high school degree 2.42%.

Marital status * Accepted the offer in the last campaign Crosstabulation

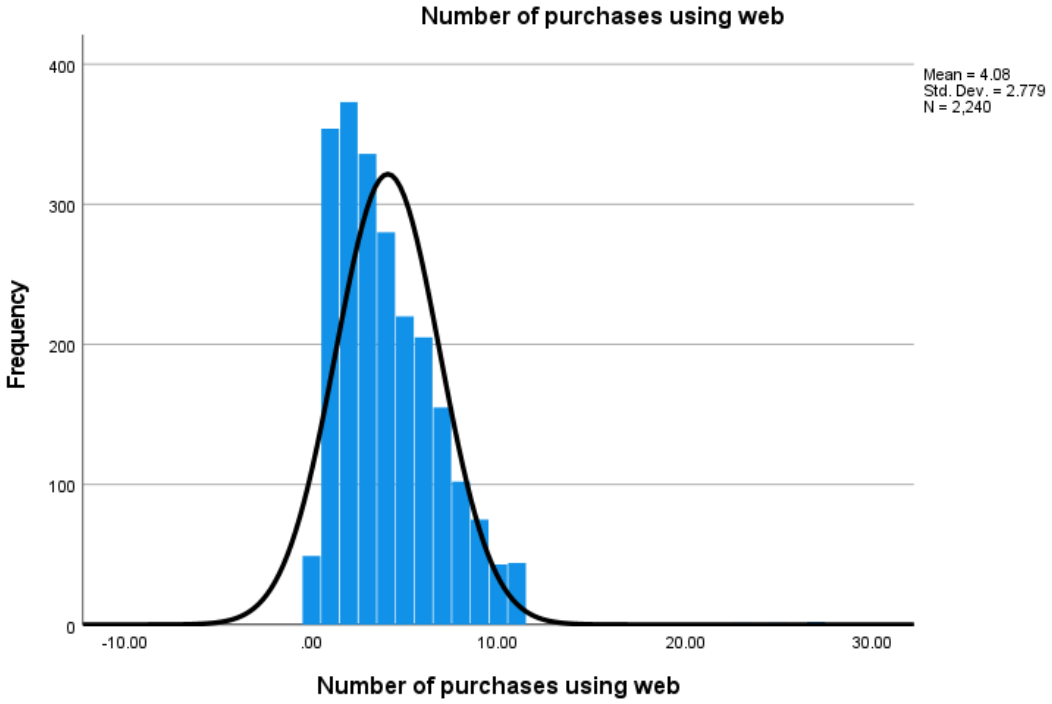
Count

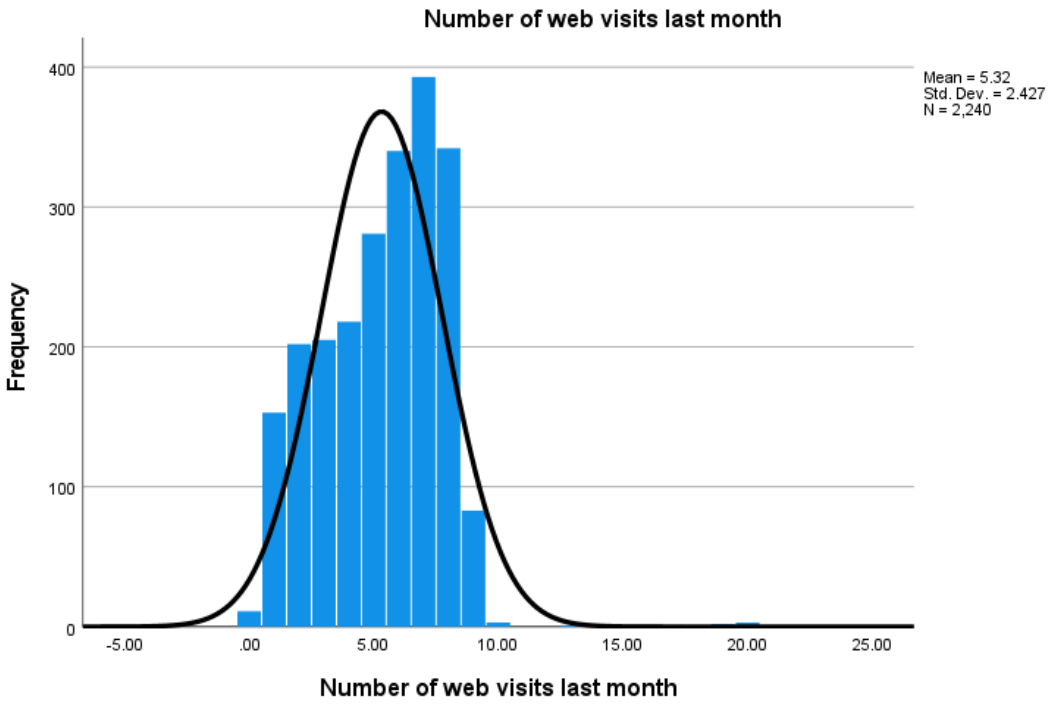
		Accepted the offer in the last campaign		
		No	Yes	Total
Marital status	Divorced	184	48	232
	Married	766	98	864
	Single	376	107	483
	Together	520	60	580
	Widow	58	19	77
Total		1904	332	2236

In this crosstab, we can see the marital status of the people who accepted the offer in the last campaign or not. It is interesting to see that 32.22% (107/332) out of the total amount of people who accepted the offer in the last campaign's marital status is single. However, single people only make up 21.6% of the total customers in the campaign. Widowed people make up the lowest group

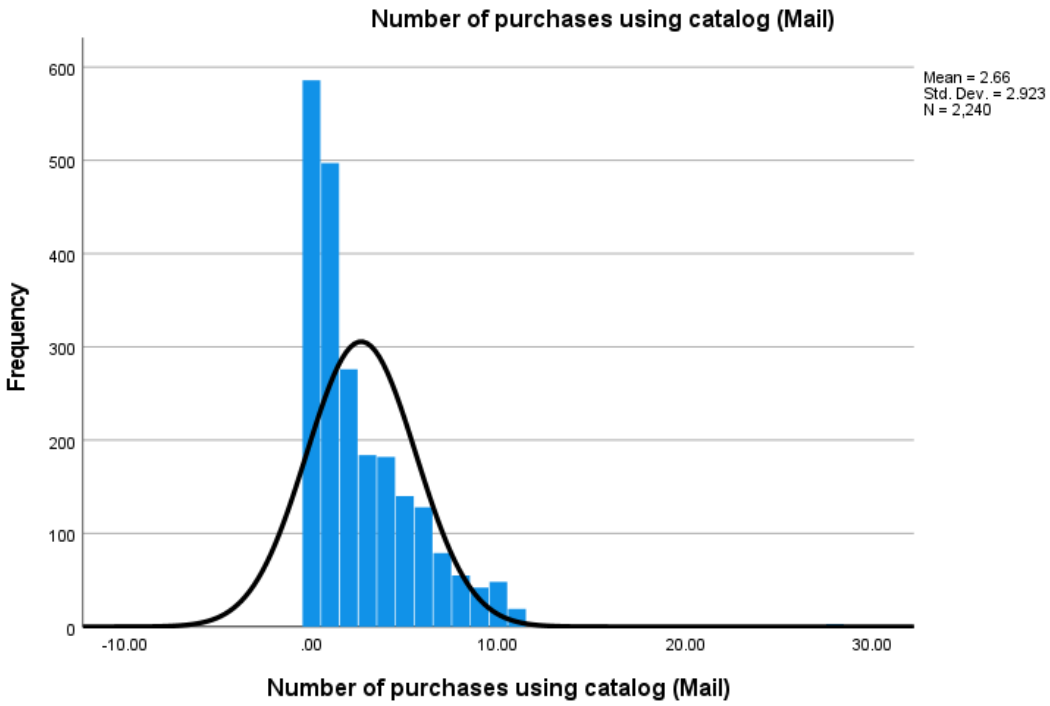
of people who accepted the offer in the last campaign, with 5.72%, but 32.76% of the widowed people accepted the offer.

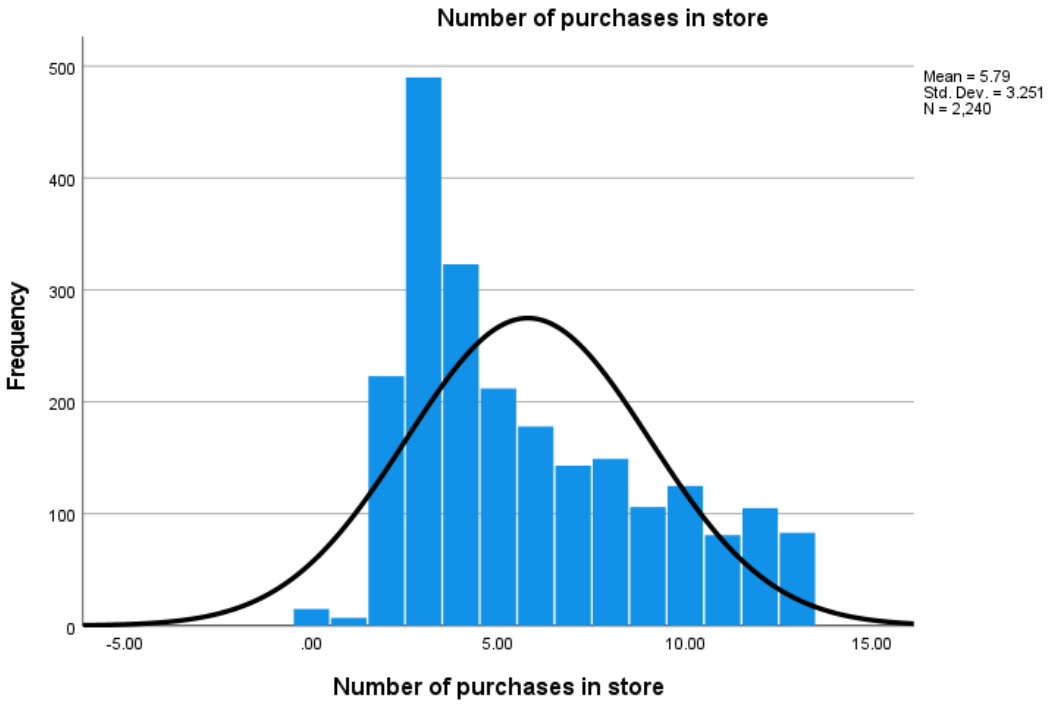
The following 2 charts show the distribution of the number of purchases made using the website, and the number of visits to the website respectively. The first one is slightly skewed to the left, while the second is slightly skewed to the right. This stands to reason as not everyone that logs a visit to the website is guaranteed to buy something. However, the means are close, which is a good sign for the website.





The following two graphs show the numbers of purchases using the catalog and the number of purchases in the store. Both are skewed to the left. However, the second one shows high kurtoses on the right. Moreover, the number of purchases made in store is the highest of all other modes with an average of 5.79 purchases per person.


Number of purchases using catalog (Mail)
Mean = 2.66
Std. Dev. = 2.923
N = 2,240
Frequency
600
500
400
300
200
100
0
-10.00
.00
10.00
20.00
30.00
Number of purchases using catalog (Mail)

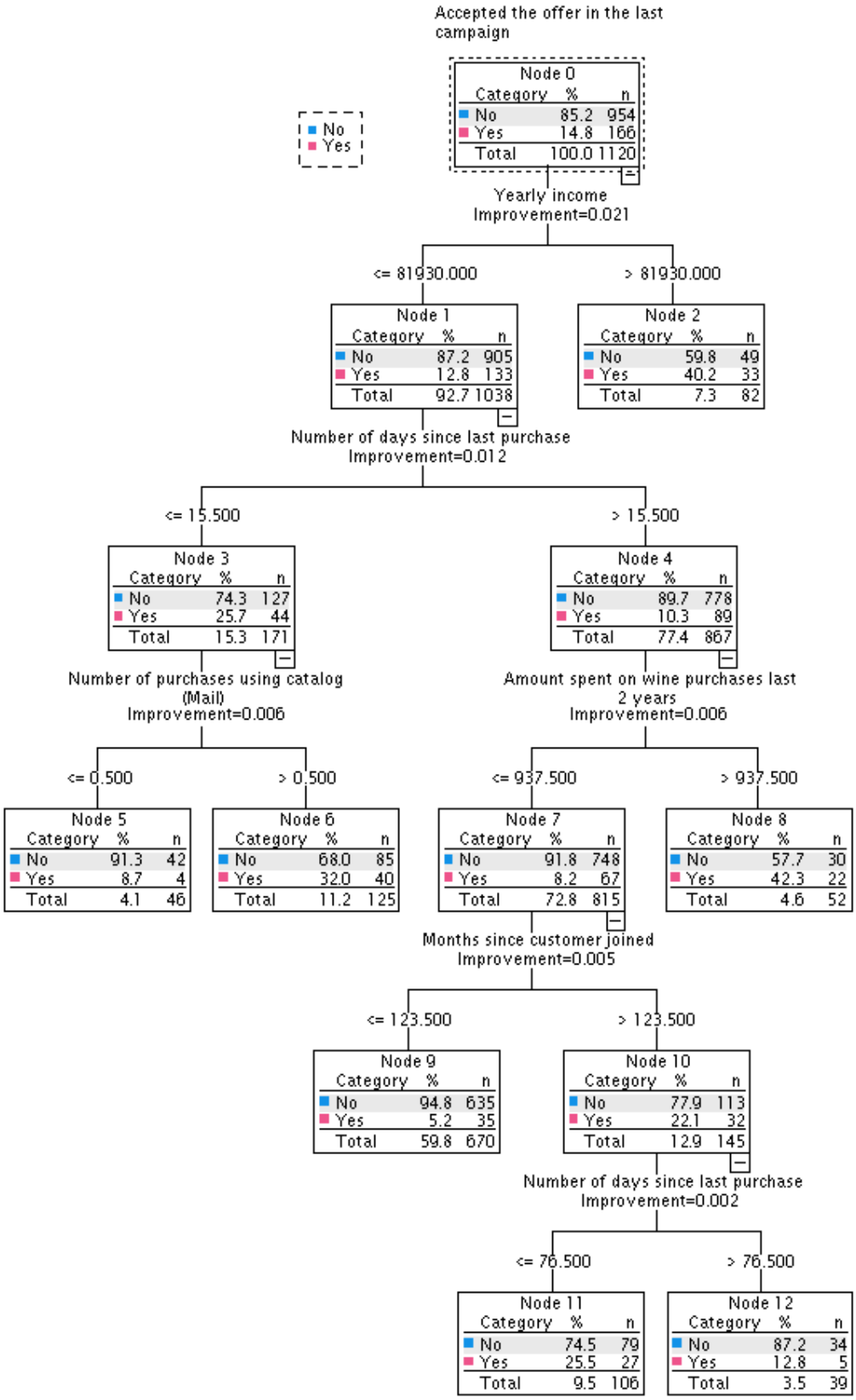

Number of purchases in store
Mean = 5.79
Std. Dev. = 3.251
N = 2,240
Frequency
500
400
300
200
100
0
-5.00
.00
5.00
10.00
15.00
Number of purchases in store

CART

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	Accepted the offer in the last campaign
	Independent Variables	Year of birth, Amount spent on meat product purchases last 2 years, Number of purchases in store, Number of purchases using catalog (Mail), Number of web visits last month, Number of purchases using web, Amount spent on fish product purchases last 2 years, Number of purchases made with discount, Amount spent on fruit purchases last 2 years, Amount spent on gold product purchases last 2 years, Education level, Amount spent on sweet product purchases last 2 years, Amount spent on wine purchases last 2 years, Yearly income, Months since customer joined, Number of days since last purchase, Marital status, Number of teens at home, Number of kids at home
	Validation	Split Sample
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	80
	Minimum Cases in Child Node	40
Results	Independent Variables Included	Yearly income, Number of purchases made with discount, Amount spent on meat product purchases last 2 years, Amount spent on sweet product purchases last 2 years, Amount spent on fruit purchases last 2 years, Number of purchases using catalog (Mail), Number of web visits last month, Amount spent on wine purchases last 2 years, Number of days since last purchase, Number of purchases using web, Amount spent on gold product purchases last 2 years, Number of purchases in store, Amount spent on fish product purchases last 2 years, Education level, Months since customer joined, Year of birth, Number of kids at home, Marital status
	Number of Nodes	13
	Number of Terminal Nodes	7
	Depth	5

The best CART model was found by running several different models, and obtaining an equilibrium between a low risk, high sensitivity and specificity, and high index. After running several trees, the best model was achieved by setting the minimum cases in parent nodes to 80, and child nodes to 40, and changing the threshold for classifying successes from 50% to 35%. Our overall response rate contained only 15% of the successes. The following page shows the tree structure.


Accepted the offer in the last campaign
No
Yes
Node 0
Category % n
No 85.2 954
Yes 14.8 166
Total 100.0 1120
Yearly income
Improvement=0.021
<= 81930.000
> 81930.000
Node 1
Category % n
No 87.2 905
Yes 12.8 133
Total 92.7 1038
Node 2
Category % n
No 59.8 49
Yes 40.2 33
Total 7.3 82
Number of days since last purchase
Improvement=0.012
<= 15.500
> 15.500
Node 3
Category % n
No 74.3 127
Yes 25.7 44
Total 15.3 171
Node 4
Category % n
No 89.7 778
Yes 10.3 89
Total 77.4 867
Number of purchases using catalog (Mail)
Improvement=0.006
Amount spent on wine purchases last 2 years
Improvement=0.006
<= 0.500
> 0.500
<= 937.500
> 937.500
Node 5
Category % n
No 91.3 42
Yes 8.7 4
Total 4.1 46
Node 6
Category % n
No 68.0 85
Yes 32.0 40
Total 11.2 125
Node 7
Category % n
No 91.8 748
Yes 8.2 67
Total 72.8 815
Node 8
Category % n
No 57.7 30
Yes 42.3 22
Total 4.6 52
Months since customer joined
Improvement=0.005
<= 123.500
> 123.500
Node 9
Category % n
No 94.8 635
Yes 5.2 35
Total 59.8 670
Node 10
Category % n
No 77.9 113
Yes 22.1 32
Total 12.9 145
Number of days since last purchase
Improvement=0.002
<= 76.500
> 76.500
Node 11
Category % n
No 74.5 79
Yes 25.5 27
Total 9.5 106
Node 12
Category % n
No 87.2 34
Yes 12.8 5
Total 3.5 39

The Overall Tree & Nodes 0-4:

The overall tree consists of 5 levels and 12 nodes (13 including the parent node).

- The best predictor for whether the customer accepted the offer in the last campaign or not is the yearly income. 40.2% of those who have a yearly income of 81930 or higher accepted the offer.
- For those who earned less than or equal to 81930 annually, 12.8% accepted the offer.
- Out of those who earned less than or equal to 81930, the best predictor was the number of days since the last purchase. For those who had the number of days since last purchase less than or equal to 15.5 days, 25.7% accepted the offer. For those who had more than 15.5 days, only 10.3% accepted the offer.

Nodes (5-8):

- Out of those whose last purchase was less than or equal to 15.5 days, the best predictor was the number of purchases using the catalog. For those who had a purchase of more than 1, (Since one can't make 0.5 purchases) 32% accepted the offer, compared to 8.7% of those who did not make any purchase using the catalog.
- For those whose last purchase was more than 15.5 days ago, the best predictor was the amount spent on wine purchases in the last 2 years. Of those who spent more than 937.50 dollars, 42.3% of them accepted the offer, compared to 8.2% of those who spent less than or equal to 937.50 dollars.

Nodes (9-12):

- Out of those who spent less than or equal to 937.50 on wine in the last two years, the best possible predictor was months since the last purchase. For those customers who joined the company more than 123 months ago, 22.1% accepted the offer compared to 5.2% of those who joined the company less than or equal to 123 months ago.
- For those who joined the company more than 123 months ago, the best predictor was the number of days since their last purchase. Of the customers whose last purchase was more than 76.5 days ago, 12.8% accepted the offer, compared to 25.5% of those whose last purchase was less than or equal to 76.5 days ago.

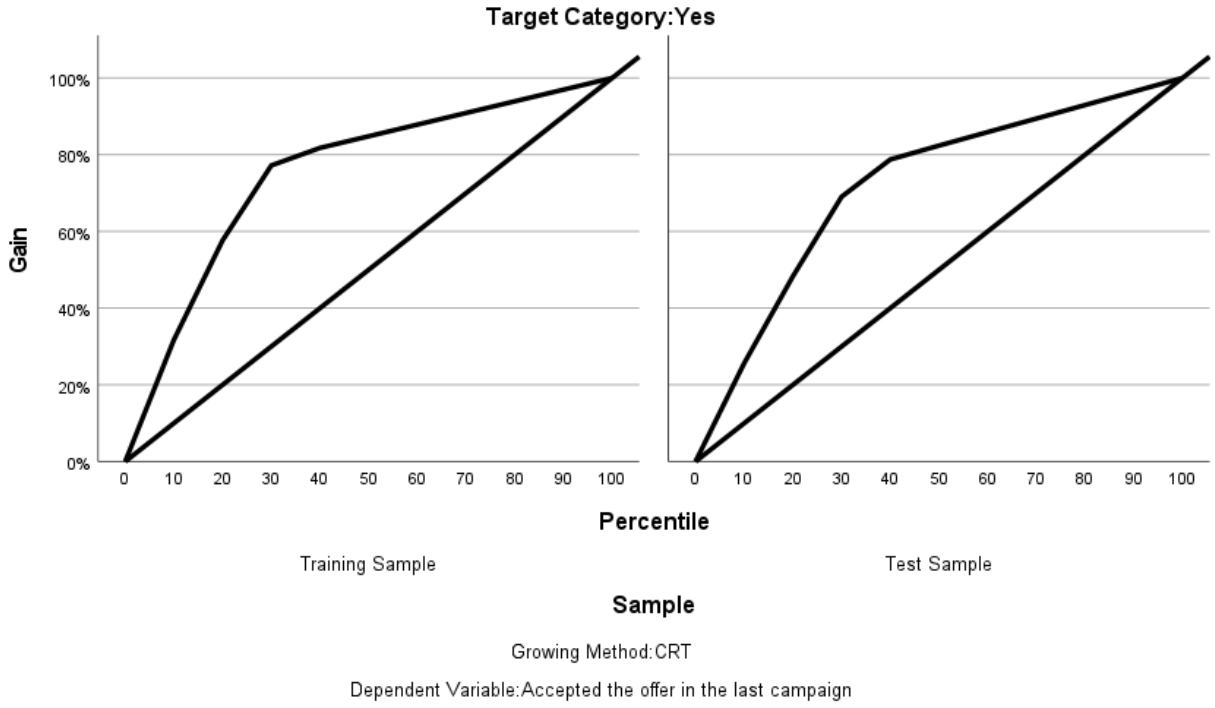
Gains Table and Gains Plots: Target Category Yes:

Gains for Nodes

Sample	Node	Node N	Node Percent	Gain N	Gain Percent	Response	Index
Training	2	84	7.5%	42	25.3%	50.0%	336.1%
	6	124	11.1%	48	28.9%	38.7%	260.2%
	8	40	3.6%	15	9.0%	37.5%	252.1%
	11	85	7.6%	23	13.9%	27.1%	181.9%
	5	49	4.4%	4	2.4%	8.2%	54.9%
	12	48	4.3%	3	1.8%	6.3%	42.0%
	9	686	61.5%	31	18.7%	4.5%	30.4%
Test	2	82	7.3%	33	19.9%	40.2%	271.5%
	6	125	11.2%	40	24.1%	32.0%	215.9%
	8	52	4.6%	22	13.3%	42.3%	285.4%
	11	106	9.5%	27	16.3%	25.5%	171.9%
	5	46	4.1%	4	2.4%	8.7%	58.7%
	12	39	3.5%	5	3.0%	12.8%	86.5%
	9	670	59.8%	35	21.1%	5.2%	35.2%

Growing Method: CRT
Dependent Variable: Accepted the offer in the last campaign

- While the Node section of the gains table shows the percentage of the customers in the node compared to the total customers in the dataset, the response percentage was sorted by the target category, which is Yes.
- From the index, we could interpret that the individuals who fell in the node 2, they responded at a rate 171.5% higher compared to the overall (testing). The ones who fell under node 9 terminal node had the worst rate of accepting the offer with a rate of 4.5%.



- From the Gains Plots, we can see that there is a slight difference in the model benefit area between the testing and training samples, but it's a decent size.
- The top 50% of the file contained around 82% of all the customers who accepted the offer (testing sample).

Risk

Sample	Estimate	Std. Error
Training	.149	.011
Test	.148	.011

Growing Method: CRT
Dependent Variable: Accepted the offer in the last campaign

14.9% of the training dataset is classified incorrectly, compared to 14.8% of testing. Since they are almost identical, this tells us there are no issues with overfitting.

Classification

Sample	Observed	Predicted No	Predicted Yes	Percent Correct
Training	No	950	0	100.0%
	Yes	166	0	0.0%
	Overall Percentage	100.0%	0.0%	85.1%
Test	No	954	0	100.0%
	Yes	166	0	0.0%
	Overall Percentage	100.0%	0.0%	85.2%

Growing Method: CRT
Dependent Variable: Accepted the offer in the last campaign

This is the original classification table with the threshold set to 0.5. Even though our specificity was 100%, but our sensitivity was 0% for both testing and training.

Predicted Probability * Accepted the offer in the last campaign Crosstabulation[a]

			Accepted the offer in the last campaign: No	Accepted the offer in the last campaign: Yes	Total
Predicted Probability	.00	Count	807	61	868
		% within Accepted the offer in the last campaign	84.9%	36.7%	77.8%
	1.00	Count	143	105	248
		% within Accepted the offer in the last campaign	15.1%	63.3%	22.2%
Total		Count	950	166	1116
		% within Accepted the offer in the last campaign	100.0%	100.0%	100.0%

a. Sample Assignment = Training Sample

When we changed the threshold to 0.35, we predicted that those who did not accept the offer 84.9% correctly (specificity) and we predicted those who accepted the offer correctly 63.3%. This is a better option because we would want to predict those who accepted the offer correctly, without lowering the percentage for those we predicted incorrectly too much.

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
SQUEadjusted	1116	.00	.39	.0818	.13994
SQUEoriginal	1116	.00	.25	.0478	.07773
Valid N (listwise)	1116				

a. Sample Assignment = Training Sample

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
SQUEadjusted	1120	.00	.39	.0869	.14277
SQUEoriginal	1120	.00	.25	.0501	.07750
Valid N (listwise)	1120				

a. Sample Assignment = Testing Sample

The Average Squared Error for the adjusted classification threshold is higher than the original .50, However, the ASE for both testing and training samples are very close for both cases, which tells us that the model does not have any issues with overfitting.

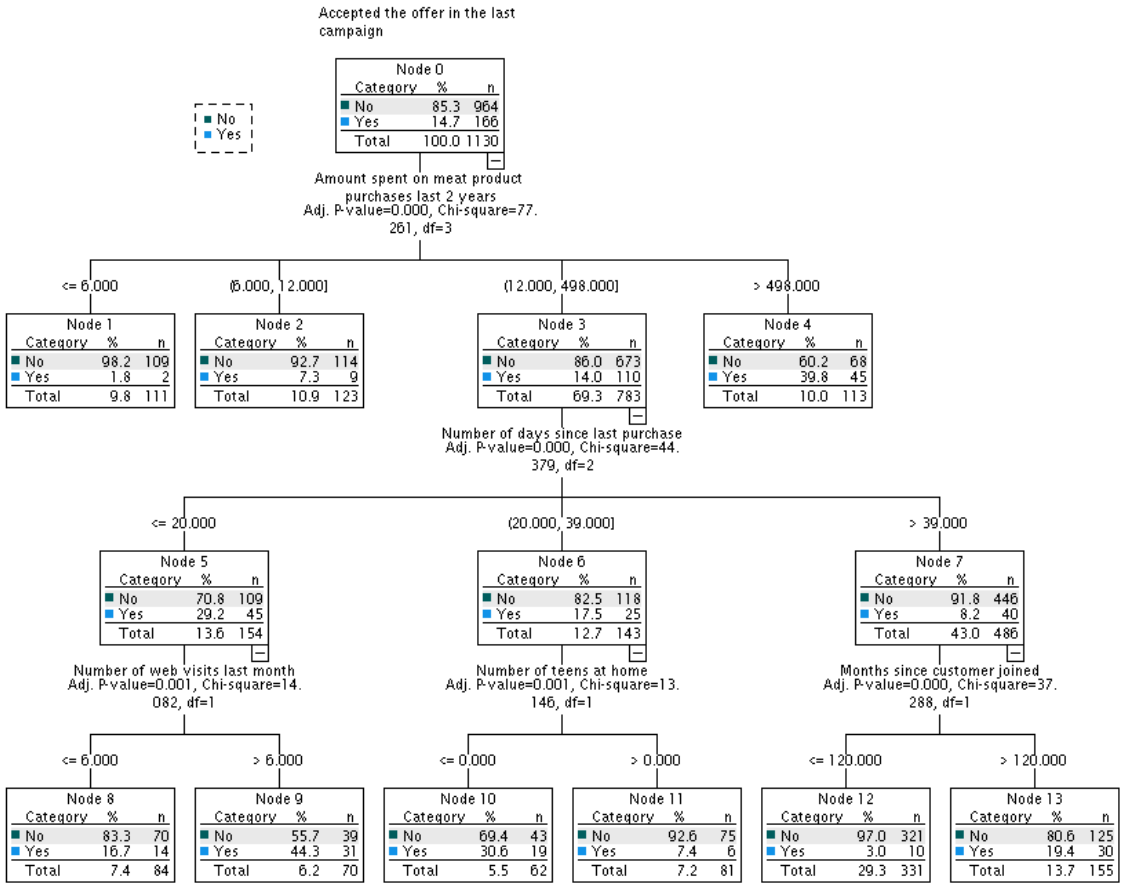
As we could see from the independent variable importance table, the yearly income was the most important predictor for if a person accepted the offer or not, with 100% normalized importance, compared to the number of kids at home with 0.7% normalized importance.

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
Yearly income	.021	100.0%
Number of days since last purchase	.015	70.0%
Amount spent on wine purchases last 2 years	.009	43.1%
Number of purchases using catalog (Mail)	.006	29.3%
Months since customer joined	.005	25.8%
Amount spent on meat product purchases last 2 years	.005	25.6%
Number of purchases made with discount	.004	17.0%
Number of purchases using web	.003	12.1%
Amount spent on sweet product purchases last 2 years	.002	11.2%
Amount spent on gold product purchases last 2 years	.002	9.2%
Amount spent on fish product purchases last 2 years	.001	4.2%
Amount spent on fruit purchases last 2 years	.001	4.0%
Number of web visits last month	.000	1.9%
Education level	.000	1.6%
Year of birth	.000	1.4%
Number of purchases in store	.000	0.9%
Marital status	.000	0.7%
Number of kids at home	.000	0.7%

Growing Method: CRT
Dependent Variable: Accepted the offer in the last campaign

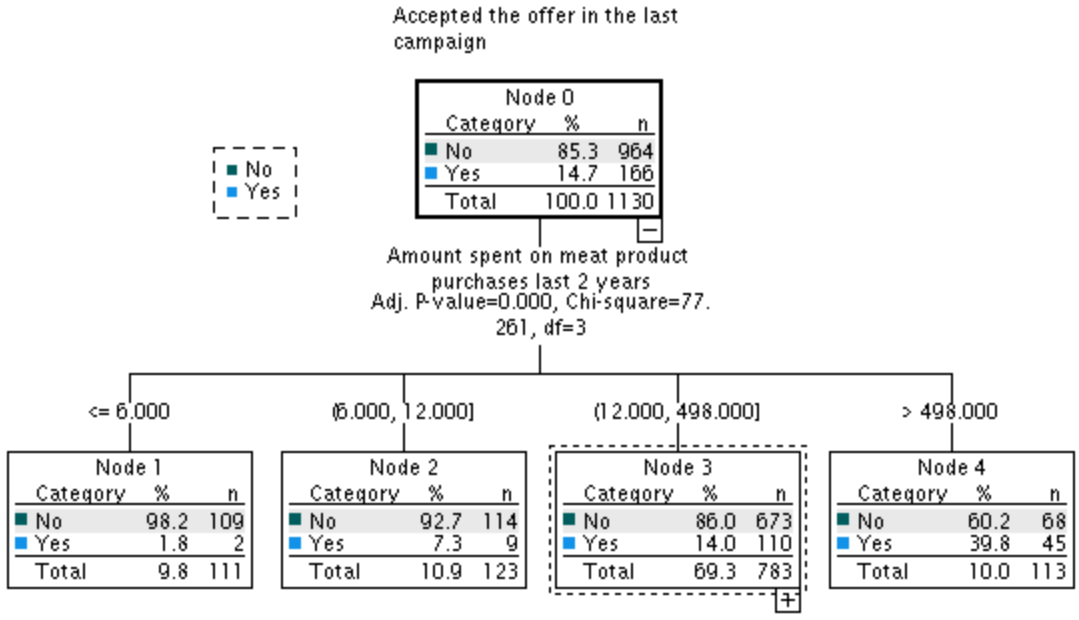
CHAID and exhaustive CHAID



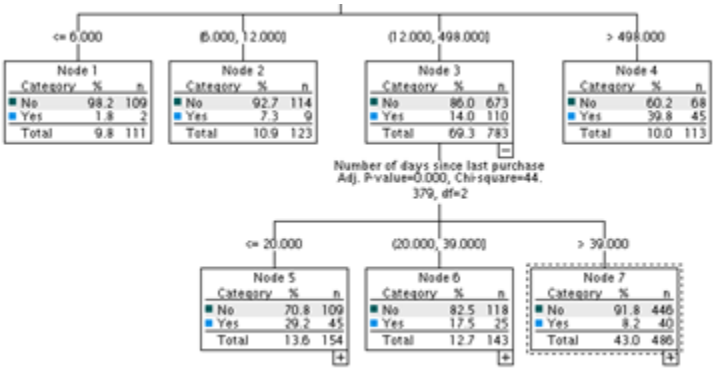
Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	Accepted the offer in the last campaign
	Independent Variables	Year of birth, Education level, Marital status, Yearly income, Number of kids at home, Number of teens at home, Months since customer joined, Number of days since last purchase, Amount spent on wine purchases last 2 years, Amount spent on fruit purchases last 2 years, Amount spent on meat product purchases last 2 years, Amount spent on fish product purchases last 2 years, Amount spent on sweet product purchases last 2 years, Amount spent on gold product purchases last 2 years, Number of purchases made with discount, Number of purchases using web, Number of purchases using catalog (Mail), Number of purchases in store, Number of web visits last month
	Validation	Split Sample
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Amount spent on meat product purchases last 2 years, Number of days since last purchase, Number of web visits last month, Number of teens at home, Months since customer joined
	Number of Nodes	14
	Number of Terminal Nodes	9
	Depth	3

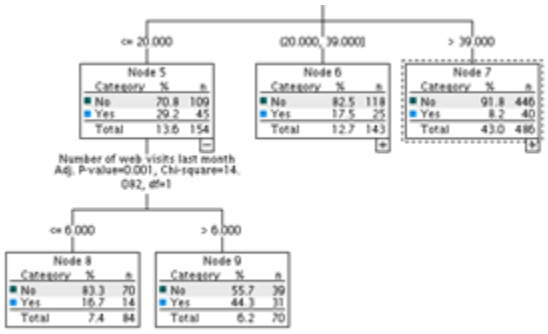
This is the final CHAID model included in the project, it has the default settings of 100 parent size and 50 child size and a depth of 3. The model started with 19 predictor variables but only ended up including five in the final model. There are a total of 14 nodes and 9 terminal nodes in the tree. (Overall picture of tree included below.) The variables in the results section are listed in order of importance from most to least helpful at predicting whether or not a customer (existing customer, not trying to recruit new ones) will accept an offer for a discounted price on a gold membership that gives them 20% on all products from the store.



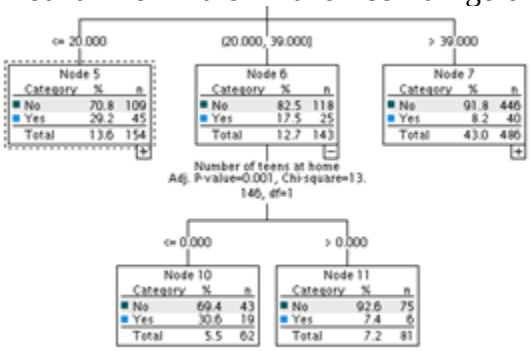
This is the first level of the tree resulting in four child nodes only one of which becomes a parent node for the next levels. The variable most helpful when predicting whether or not a customer accepts the offer is the amount spent on meat product purchases in the last two years. This is an interesting choice for the first split but there is some sense in the choice; if customers are buying meat product from this superstore it follows that this superstore is what they use for their regular grocery store. The three terminal nodes on this level imply that the customer made a one-time purchase of a produce, either a relatively small amount to try a new produce or a very large order possibly for a party. But the third node, many customers fall into that category and they are the customers who are probably frequently at that store making semi regular grocery purchases. The first two nodes are really good at predicting customers who won't accept the sale on a gold membership. The fourth node is a little more split evenly and was able to predict that 45 of the 113 customers who purchases over $498 in meat product would accept the offer. This makes sense as well since it is possible that those customers represent a person buying groceries regularly for a large family. Node three predicts customers who won't accept the offer more so than customers who would.



The next three nodes 5, 6, and 7 make up level two, determined by the variable number of days since last purchase (recency) which is often an important variable in determine how customers response to any type of sales campaign. The general trend is the more recent a purchase the more likely the customer is to accepting the sales offer and we can see that in this level. The node that found the highest number of customers predicted to accept the offer was node 5, those customers were people who purchased between $12 and $498 worth of meat product in the last two years and their last purchase was within 20 days and 45 of those 154 people were predicted to accept the offer. The other two nodes are lower, node 6 is all the people who purchased between $12 and $498 worth of meat product in the last two years but their last purchase was between 21 and 39 days ago and only 25 of those people were predicted to accept the offer. Node 7 only has 8% of its people predicted to accept the offer but a much larger number of customers included than the other two at 486. All three of these nodes are the parent nodes to the next and final level of the tree.

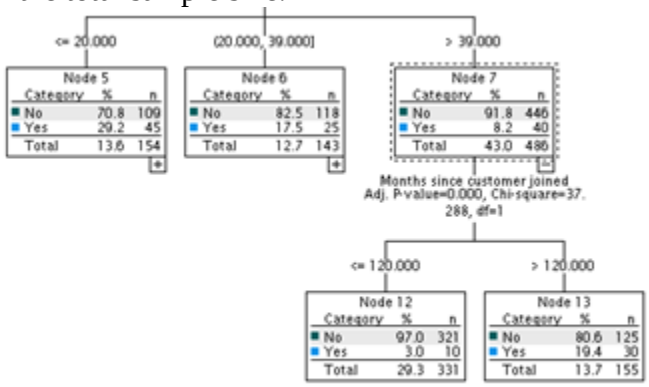


Node 8 and 9 are terminal nodes and the variable that determined them was number of visits last month. Node 9 is of particular interest because it is one of the best performing nodes predicting 44% of its 70 total people as customers who would accept the offer. Cases in node 9 are customers who purchased between $12 and $498 worth of meat product in the last two years and visited the store in the past 20 days and who have visited more than 6 times in the last month. Given all that information it would make sense that almost 50% of that group would response to an offer for a discounted gold membership. These people obviously frequent the store a lot and buy a good amount from them and so a gold membership would benefit them most of all.

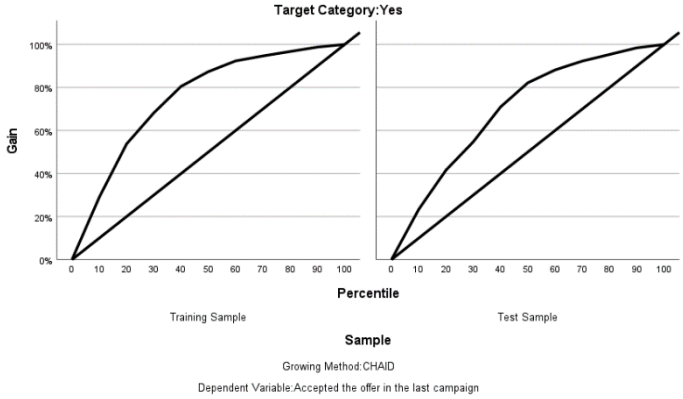


Node 10 and 11 both come from node 6 with the variable number of teens at home. Node 10 does not have any teens at home, which is interesting because 30% of those customers responded to the

offer but it would make more sense if the bigger the family the more the need for discounts on groceries. It is possible that the bigger families have less money and don't shop at the bigger superstores. Either way, node 11 predicted that all but 6 of its total cases would not accept an offer even though they have at least 1 teenager in the household. Both nodes have a very small percent of the total sample size.



This is the last section of the tree with node 12 and 13. The variable that created this split was months since customer joined. It seems like the longer they've been a customer they are a little more likely to response to the offer but in general neither of these nodes are very good at predicting success.



This is the gains chart showing by how much the model improved things. So if we were to capture 50% of the data we would get around 80% of the successes.

Risk

Sample	Estimate	Std. Error
Training	.147	.011
Test	.150	.011

Growing Method: CHAID
Dependent Variable: Accepted the offer in the last campaign

Classification

Sample	Observed	Predicted No	Predicted Yes	Percent Correct
Training	No	964	0	100.0%
	Yes	166	0	0.0%
	Overall Percentage	100.0%	0.0%	85.3%
Test	No	940	0	100.0%
	Yes	166	0	0.0%
	Overall Percentage	100.0%	0.0%	85.0%

Growing Method: CHAID
Dependent Variable: Accepted the offer in the last campaign

Accepted the offer in the last campaign * misclass_0.35 Crosstabulation[a]

			misclass_0.35 no	misclass_0.35 yes	Total
Accepted the offer in the last campaign	No	Count	857	107	964
		% within Accepted the offer in the last campaign	88.9%	11.1%	100.0%
	Yes	Count	90	76	166
		% within Accepted the offer in the last campaign	54.2%	45.8%	100.0%
Total		Count	947	183	1130
		% within Accepted the offer in the last campaign	83.8%	16.2%	100.0%

a. Sample Assignment = Training Sample

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
ASE_1	1130	.00	.96	.1053	.20138
Valid N (listwise)	1130				

a. Sample Assignment = Training Sample

This is the risk, misclassification and ASE output, for the misclassification table I lowered the threshold from 0.5 to 0.35 since the overall success rate was only 0.15 to begin with. I did include the first table to show the 0% sensitivity but when I lowered the threshold the sensitivity was then about 46% which isn't too bad, and the specificity was 89%. Risk was just under 15% for the

training data set and there is not a large gap between training and testing so there is no indication of overfitting in this model. ASE for this model is 0.103 which is good because we are always looking for a low number here.

Gains for Nodes

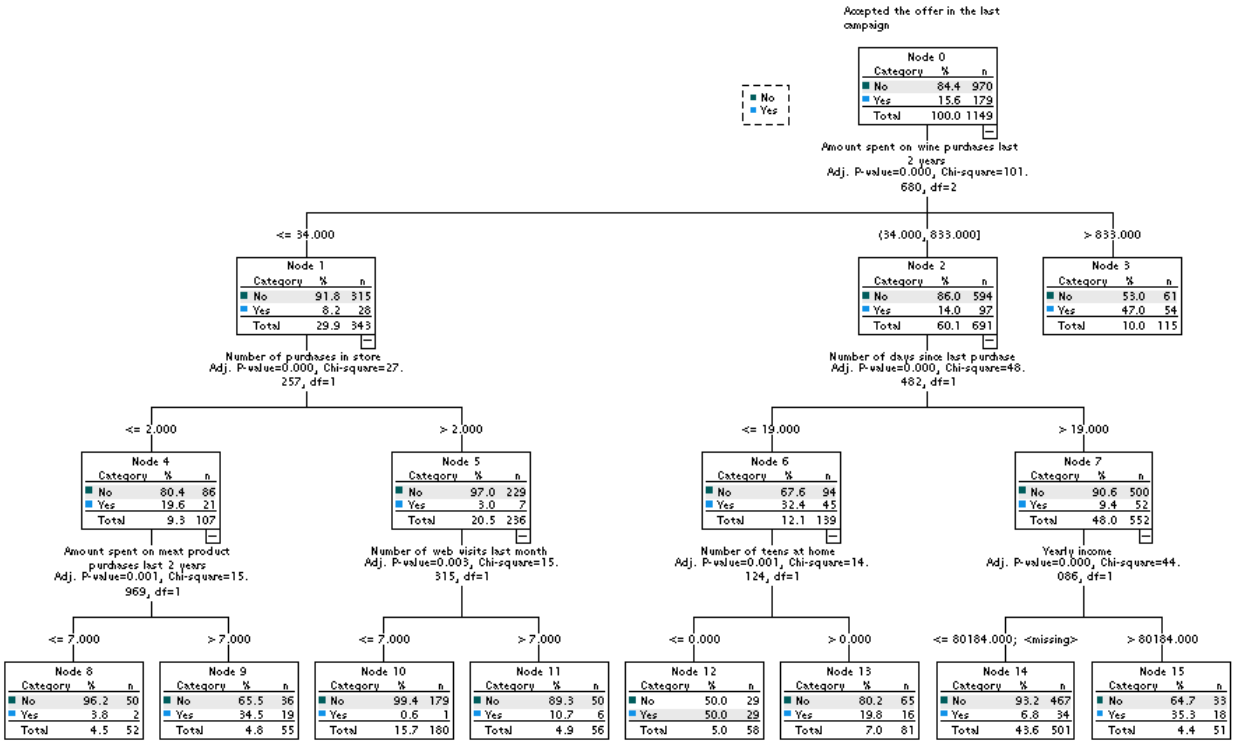
Sample	Node	Node N	Node Percent	Gain N	Gain Percent	Response	Index
Training	9	70	6.2%	31	18.7%	44.3%	301.5%
	4	113	10.0%	45	27.1%	39.8%	271.1%
	10	62	5.5%	19	11.4%	30.6%	208.6%
	13	155	13.7%	30	18.1%	19.4%	131.8%
	8	84	7.4%	14	8.4%	16.7%	113.5%
	11	81	7.2%	6	3.6%	7.4%	50.4%
	2	123	10.9%	9	5.4%	7.3%	49.8%
	12	331	29.3%	10	6.0%	3.0%	20.6%
	1	111	9.8%	2	1.2%	1.8%	12.3%
Test	9	57	5.2%	18	10.8%	31.6%	210.4%
	4	111	10.0%	43	25.9%	38.7%	258.1%
	10	53	4.8%	8	4.8%	15.1%	100.6%
	13	152	13.7%	30	18.1%	19.7%	131.5%
	8	111	10.0%	30	18.1%	27.0%	180.1%
	11	84	7.6%	9	5.4%	10.7%	71.4%
	2	138	12.5%	12	7.2%	8.7%	57.9%
	12	303	27.4%	14	8.4%	4.6%	30.8%
	1	97	8.8%	2	1.2%	2.1%	13.7%

Growing Method: CHAID
Dependent Variable: Accepted the offer in the last campaign

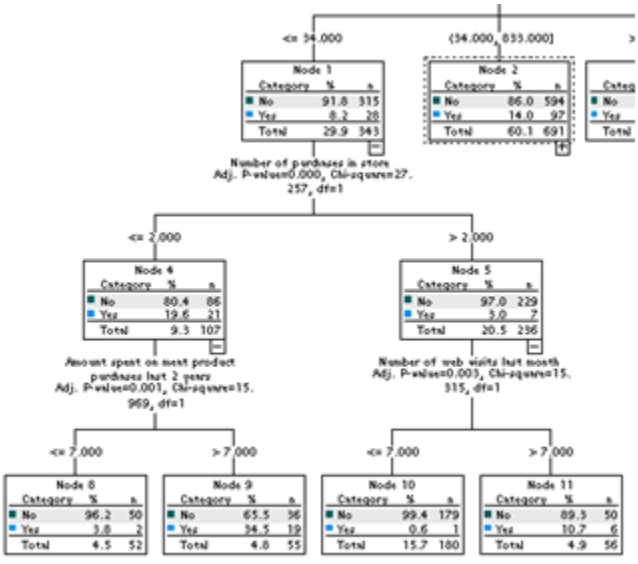
Some of the top performing nodes were nodes 9 and 4. Node 9 had 6.2% of the observations and a response rate of 44% meaning it predicted that almost half of its cases were successes. These customers are people who spent between $12 and $498 on meat products in the last two years, made a purchase in the last 20 days and visited more than 6 times in the last month. It makes sense then that this would be the best performing node because all of those categories indicate a frequent, loyal customer who would greatly benefit from a gold membership. They frequent the store enough to offset the cost of membership with the savings they would receive from it. Node 4 had 10% of the data and just under a 40% response rate. Node 1 was the worst performing node in both the testing and training data sets with under a 2% response rate. These are customers who purchased under $6 in meat products in the past two years which is an indication that they are a very infrequent customer.

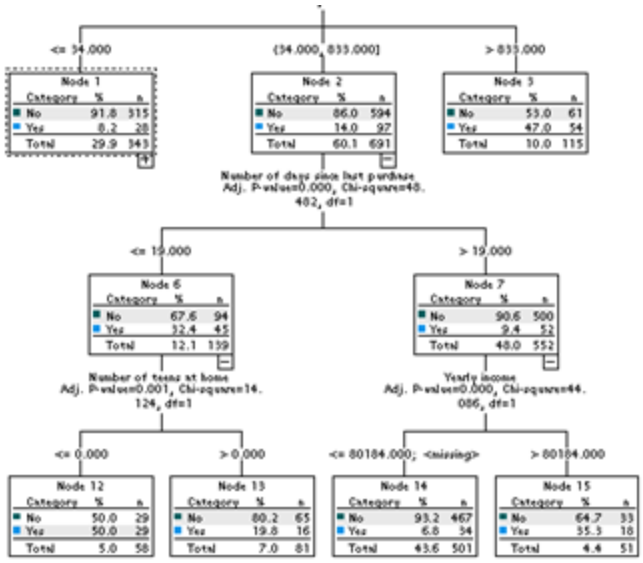
Model Summary

Specifications	Growing Method	EXHAUSTIVE CHAID
	Dependent Variable	Accepted the offer in the last campaign
	Independent Variables	Year of birth, Education level, Marital status, Yearly income, Number of kids at home, Number of teens at home, Months since customer joined, Number of days since last purchase, Amount spent on wine purchases last 2 years, Amount spent on fruit purchases last 2 years, Amount spent on meat product purchases last 2 years, Amount spent on fish product purchases last 2 years, Amount spent on sweet product purchases last 2 years, Amount spent on gold product purchases last 2 years, Number of purchases made with discount, Number of purchases using web, Number of purchases using catalog (Mail), Number of purchases in store, Number of web visits last month
	Validation	Split Sample
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Amount spent on wine purchases last 2 years, Number of purchases in store, Amount spent on meat product purchases last 2 years, Number of web visits last month, Number of days since last purchase, Number of teens at home, Yearly income
	Number of Nodes	16
	Number of Terminal Nodes	9
	Depth	3



I also ran the exhaustive CHAID model and though it looks a little different from the CHAID model I determined in the end it did not do much better or worse than the CHAID model. It ended in the same number of terminal nodes with the same depth. Even though after the first level there will only be two child nodes from any of the parent nodes. Variables, amount spent on meat product in the past two years, number of days since last purchase and number of teens in the home were used in both the CHAID and exhaustive CHAID models.





This just shows the left and right sides of the tree. The third node on the first level is a terminal node and the rest of the terminal nodes come in the third level. Some of the best performing nodes here were node 12 with a 50% response rate but only 5% of the data, these were customers who spent between 34 and $833 on wine in the past two years, purchased something in the past 19 days, and had no teens at home. Node 3 had 10% of the observations and a 47% response rate. The worst performing node was node 10 with an under 1% response rate. These were customers who spent less than $34 on wine in the past two years, made more than two in store purchases and visited the website under 7 times in the past month.

Risk

Sample	Estimate	Std. Error
Training	.156	.011
Test	.157	.011

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: Accepted the offer in the last campaign

Classification

Sample	Observed	Predicted No	Predicted Yes	Percent Correct
Training	No	941	29	97.0%
	Yes	150	29	16.2%
	Overall Percentage	95.0%	5.0%	84.4%
Test	No	894	40	95.7%
	Yes	131	22	14.4%
	Overall Percentage	94.3%	5.7%	84.3%

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: Accepted the offer in the last campaign

Accepted the offer in the last campaign * misclass2_0.35 Crosstabulation[a]

			misclass2_0.35 no	misclass2_0.35 yes	Total
Accepted the offer in the last campaign	No	Count	824	140	964
		% within Accepted the offer in the last campaign	85.5%	14.5%	100.0%
	Yes	Count	84	82	166
		% within Accepted the offer in the last campaign	50.6%	49.4%	100.0%
Total		Count	908	222	1130
		% within Accepted the offer in the last campaign	80.4%	19.6%	100.0%

a. Sample Assignment = Training Sample

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
ASE_2	1130	.00	.99	.1095	.20934
Valid N (listwise)	1130				

a. Sample Assignment = Training Sample

This again is the risk, misclassification, and ASE output. Risk is only very slightly higher than in the CHAID model and a very very small difference between training and testing so again no indication of overfitting. Again, I changed the misclassification threshold to 0.35 moving it closer to the original which was around 15%. Exhaustive CHAID only did slightly better in sensitivity at 50% and 85% for specificity. ASE was also nearly the same as the CHAID model.

Gains for Nodes

Sample	Node	Node N	Node Percent	Gain N	Gain Percent	Response	Index
Training	12	58	5.0%	29	16.2%	50.0%	320.9%
	3	115	10.0%	54	30.2%	47.0%	301.4%
	15	51	4.4%	18	10.1%	35.3%	226.6%
	9	55	4.8%	19	10.6%	34.5%	221.7%
	13	81	7.0%	16	8.9%	19.8%	126.8%
	11	56	4.9%	6	3.4%	10.7%	68.8%
	14	501	43.6%	34	19.0%	6.8%	43.6%
	8	52	4.5%	2	1.1%	3.8%	24.7%
	10	180	15.7%	1	0.6%	0.6%	3.6%
Test	12	62	5.7%	22	14.4%	35.5%	252.1%
	3	99	9.1%	34	22.2%	34.3%	244.0%
	15	48	4.4%	14	9.2%	29.2%	207.2%
	9	46	4.2%	12	7.8%	26.1%	185.3%
	13	70	6.4%	17	11.1%	24.3%	172.5%
	11	57	5.2%	4	2.6%	7.0%	49.9%
	14	471	43.3%	40	26.1%	8.5%	60.3%
	8	61	5.6%	3	2.0%	4.9%	34.9%
	10	173	15.9%	7	4.6%	4.0%	28.7%

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: Accepted the offer in the last campaign

Lastly, this is the gains table for the exhaustive CHAID model, node 12 performed best in both the testing and training. Node 12 contained people who spent between $38 and $833 on wine, visited within the last 19 days don't have any teenagers at home. Those first two variables are indicators of being a frequent customer which would then make sense that there are a higher number of people who would response to an offer among that group. The same logic can be applied to Node 10 the worst performing node. All three of the splits that lead to it are an indication that those customers don't use this superstore regularly so they would have no need for an expensive membership, the initial cost to them would not be worth it.

Logistic Regression

Multicollinearity

The next figure shows the correlation matrix between the variables. There is a moderately high correlation between yearly income and the amount spent on all products. There is also some negative correlation between yearly income and the number of kids at home. The number of kids at home is negatively correlated with wine purchases. Highlighted in yellow are correlations higher than ±0.400.

Correlations

	Year of birth	Yearly income	Number of kids at home	Number of teens at home	Months since customer joined	Number of days since last purchase	Amount spent on wine purchases last 2 years	Amount spent on fruit purchases last 2 years	Amount spent on meat product purchases last 2 years	Amount spent on fish product purchases last 2 years	Amount spent on sweet product purchases last 2 years	Amount spent on gold product purchases last 2 years	Number of purchases made with discount	Number of purchases using web	Number of purchases using catalog (Mail)	Number of purchases in store	Number of web visits last month	Accepted the offer in the last campaign
Year of birth	--																	
Yearly income	-.163**	--																
Number of kids at home	.231**	-.429**	--															
Number of teens at home	-.353**	.020	-.036	--														
Months since customer joined	.026	-.018	-.056**	.005	--													
Number of days since last purchase	-.020	-.004	.008	.018	.030	--												
Amount spent on wine purchases last 2 years	-.159**	.579**	-.496**	.005	.146**	.016	--											
Amount spent on fruit purchases last 2 years	-.019	.430**	-.373**	-.175**	.059**	-.005	.390**	--										
Amount spent on meat product purchases last 2 years	-.030	.585**	-.438**	-.261**	.074**	.022	.563**	.543**	--									
Amount spent on fish product purchases last 2 years	-.043*	.439**	-.388**	-.202**	.079**	.000	.401**	.594**	.569**	--								
Amount spent on sweet product purchases last 2 years	-.019	.441**	-.371**	-.162**	.076**	.022	.387**	.567**	.525**	.582**	--							
Amount spent on gold product purchases last 2 years	-.065**	.325**	-.349**	-.019	.145**	.016	.389**	.391**	.351**	.418**	.371**	--						
Number of purchases made with discount	-.061**	-.083**	.223**	.387**	.200**	.001	.011	-.131**	-.122**	-.139**	-.119**	.050*	--					
Number of purchases using web	-.146**	.388**	-.361**	.155**	.169**	-.009	.543**	.298**	.295**	.296**	.349**	.424**	.233**	--				
Number of purchases using catalog (Mail)	-.123**	.589**	-.503**	-.109**	.091**	.024	.636**	.487**	.725**	.533**	.491**	.435**	-.007	.380**	--			
Number of purchases in store	-.129**	.529**	-.500**	.051*	.102**	.001	.642**	.462**	.480**	.461**	.449**	.383**	.069**	.503**	.519**	--		
Number of web visits last month	.122**	-.553**	.449**	.133**	.252**	-.020	-.321**	-.417**	-.539**	-.444**	-.423**	-.248**	.347**	-.057**	-.519**	-.429**	--	
Accepted the offer in the last campaign	.019	.133**	-.079**	-.155**	.173**	-.198**	.247**	.124**	.238**	.110**	.117**	.137**	.002	.148**	.220**	.039	-.003	--

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Categorical variables

To prepare the two categorical variables in this dataset for logistic regression, the number of categories had to be reduced. To do that, similarities were found using crosstabs between each category and the response variable. Showing below, is the crosstab for the education level variable.

Education level * Accepted the offer in the last campaign Crosstabulation

			Accepted the offer in the last campaign		
			No	Yes	Total
Education level	Associate	Count	181	22	203
		% within Education level	89.2%	10.8%	100.0%
	High School	Count	52	2	54
		% within Education level	96.3%	3.7%	100.0%
	Bachelor's	Count	975	151	1126
		% within Education level	86.6%	13.4%	100.0%
	Master	Count	312	57	369
		% within Education level	84.6%	15.4%	100.0%
	PhD	Count	384	100	484
		% within Education level	79.3%	20.7%	100.0%
Total		Count	1904	332	2236
		% within Education level	85.2%	14.8%	100.0%

It is noticeable that the higher the education level is the more people in this dataset purchased the gold offer. To make things simpler, associates and bachelors were combined. Master and PhD were combined as well. The following crosstab is for the combined categories. Three distinct logical categories were made.

Education level * Accepted the offer in the last campaign Crosstabulation

			Accepted the offer in the last campaign		
			No	Yes	Total
Education level	High School	Count	52	2	54
		% within Education level	96.3%	3.7%	100.0%
	BA/AS	Count	1156	173	1329
		% within Education level	87.0%	13.0%	100.0%
	Post-grad	Count	696	157	853
		% within Education level	81.6%	18.4%	100.0%
Total		Count	1904	332	2236
		% within Education level	85.2%	14.8%	100.0%

The following crosstab shows the original categories in the marital status variable. As can be seen, it seems that those who are with someone are less likely to accept the offer compared to those who are single.

Marital status * Accepted the offer in the last campaign Crosstabulation

			Accepted the offer in the last campaign		
			No	Yes	Total
Marital status	Divorced	Count	184	48	232
		% within Marital status	79.3%	20.7%	100.0%
	Married	Count	766	98	864
		% within Marital status	88.7%	11.3%	100.0%
	Single	Count	376	107	483
		% within Marital status	77.8%	22.2%	100.0%
	Together	Count	520	60	580
		% within Marital status	89.7%	10.3%	100.0%
	Widow	Count	58	19	77
		% within Marital status	75.3%	24.7%	100.0%
Total		Count	1904	332	2236
		% within Marital status	85.2%	14.8%	100.0%

So, the new categories were chosen based on the idea to separate the categories based on whether the person was single or with someone. The next table shows the new categories. There are few differences in the response rate between the old categories and the new.

Marital status * Accepted the offer in the last campaign Crosstabulation

			Accepted the offer in the last campaign		
			No	Yes	Total
Marital status	Single	Count	618	174	792
		% within Marital status	78.0%	22.0%	100.0%
	With someone	Count	1286	158	1444
		% within Marital status	89.1%	10.9%	100.0%
Total		Count	1904	332	2236
		% within Marital status	85.2%	14.8%	100.0%

When modelling, the model made the dummy variables as follows:

Categorical Variables Codings

			Parameter coding	
		Frequency	(1)	(2)
Education level	BA/AS	666	1.000	.000
	High Sch	30	.000	1.000
	Post-Grad	425	.000	.000
Marital status	Single	397	1.000	
	With Someone	724	.000	

'Post-Grad' and 'With Someone' categories were their variables' default categories. The other categories were made into dummy variables.

Modelling

A training subset (50% of total) was selected randomly. A forward likelihood ratio algorithm was chosen to select the variables with the highest likelihood ratio that would affect the model. The cut-off value for the prediction was made to 0.35 to match the other models. The algorithm starts with block 0, with only the constant. Which in most cases is useless. However, the algorithm then adds the variables one step at a time, each time assessing Pseudo R-square, sensitivity, specificity, and risk.

Step 10 was chosen as the optimal step as it had 42% sensitivity, 93.1% specificity and 14.3% risk. It also had the highest Pseudo R-square values. The following table shows the model result:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 10	Education level			7.193	2	.027			
	Education level(1)	-.432	.211	4.199	1	.040	.649	.429	.981
	Education level(2)	-2.113	1.061	3.966	1	.046	.121	.015	.967
	Marital status(1)	1.092	.200	29.729	1	<.001	2.981	2.013	4.415
	Number of teens at home	-1.208	.215	31.497	1	<.001	.299	.196	.456
	Months since customer joined	.050	.014	13.959	1	<.001	1.052	1.024	1.080
	Number of days since last purchase	-.030	.004	60.350	1	<.001	.970	.963	.978
	Amount spent on wine purchases last 2 years	.001	.000	14.996	1	<.001	1.001	1.001	1.002
	Amount spent on gold product purchases last 2 years	.005	.002	8.187	1	.004	1.005	1.002	1.009
	Number of purchases using web	.179	.046	15.097	1	<.001	1.196	1.093	1.308
	Number of purchases using catalog (Mail)	.071	.031	5.168	1	.023	1.073	1.010	1.141
	Number of purchases in store	-.226	.042	29.536	1	<.001	.798	.735	.866
	Constant	-6.573	1.576	17.399	1	<.001	.001		

As seen in the table above, there are a few variables that were not included in the model, these are: meat, fish, and sweet products, as they were highly correlated to each other with low correlation to the response. age of the customer, which had the least correlation with the response, and number of kids at home that did not add to the model performance.
The final logit equation can be written as:

$$\begin{aligned} logit(\hat{\pi}) = {} & -6.573 - 0.432\left(\frac{BA}{AS}indicator\right) - 2.113\,(High\ School\ indicator) + 1.092\,(single\ indicator) - 1.208\,(\#\ teens) \\ & + 0.050\,(Months\ since\ customer\ joined) - 0.030\,(Number\ of\ days\ since\ last\ purchase) + 0.001\,(\$Wine) \\ & + 0.005\,(\$Gold) + 0.179\,(\#\ of\ web\ purchases) + 0.071\,(\#\ of\ mail\ purchases) - 0.226\,(\#\ of\ store\ purchases) \end{aligned}$$

Overfitting

To check for overfitting the training and testing performance parameters were compared.

Training

ASE: 0.1425

Descriptive Statistics

	N	Mean
Square_err	1130	.1425
Valid N (listwise)	1130	

Sensitivity: 42%
Specificity: 93.1%
Risk: 14.3%

Classification Table[a]

Observed			Predicted: Accepted the offer in the last campaign — No	Yes	Percentage Correct
Step 10	Accepted the offer in the last campaign	No	893	66	93.1
		Yes	94	68	42.0
	Overall Percentage				85.7

a. The cut value is .350

Testing:
ASE: 0.1456

Descriptive Statistics

	N	Mean
Square_err	1106	.1456
Valid N (listwise)	1106	

Sensitivity: 44.1%
Specificity: 92.9%
Risk: 14.6%

Accepted the offer in the last campaign * Predicted_value Crosstabulation

			Predicted_value: No	Yes	Total
Accepted the offer in the last campaign	No	Count	870	66	936
			92.9%	7.1%	100.0%
	Yes	Count	95	75	170
			55.9%	44.1%	100.0%

The model was deemed not overfit.

Comparing all models

Model	Sensitivity (%)	Specificity (%)	ASE	Risk (%)
CHAID	45.8	88.9	0.1053	14.7
Exhaustive CHAID	49.4	85.5	0.1095	15.6
CART	63.3	84.9	0.0869	14.8
Logistic regression	44.1	92.9	0.1456	14.6

The CART model had the highest sensitivity, and the least square error. A high sensitivity might be more desirable in the context of this study because a yes response is the desired outcome. The logistic regression had the highest sensitivity and least risk of misclassification, however it had the highest ASE. As the sensitivity is more desirable than specificity, and as the risk of misclassification between the models is similar, CART algorithm was determined to be the best model.