Database Marketing and CRM:
A Case on Predictive Modeling for Ayurveda Product Offerings

Purba Rao

Washington Sycip Graduate School of Business
Asian Institute of Management, Philippines.

Abstract

This paper deals with an overview of predictive modeling approaches adopted by marketers to identify segments in the target market which would have the largest response rates pertaining to a certain product or service being offered.

The three approaches in the predictive modeling, which have been discussed are Heuristic approach, Statistical approach and Data Mining approach. However, with reference to a case study on which the predictive modeling is applied , essentially the case of a company manufacturing and marketing ayurveda products to a large target population, the specific predictive models considered are (1) RFM model … Recency, Frequency, Monetary Model ; (2) Logistics Regression Model ; and (3) Decision Tree model, in particular the CHAID

(Chi Square Automatic Interaction Detection) approach. These models helped to identify the profile of typical consumers belonging to the target segment having the most significant predicted response rate.

***Keywords**: Predictive modeling, consumer, market, target segment*

Introduction

Among the fastest developing marketing approaches which are applied in today's world, Direct Marketing is growing at a phenomenal rate. In the United States more than half of the population order merchandize over phone or email (Direct Marketing Association,1995, Spring) . Keeping up with this trend, sales volume by Direct Marketing has risen in a huge proportion too, and so has the advertising expenditure in the United States over Direct Marketing. Along with Direct Marketing, Database Marketing, which forms an intrinsic component of Direct Marketing, is assuming an increasingly critical role. This is happening because through Database Marketing, organizations offering products and /or services are able to reach their primary target markets in a much more effective manner and are able to develop marketing strategy in a personalized and customized manner at the marketing segment level or even at the individual level, if needed. This, in turn, helps the companies and organizations to develop and maintain a relationship with the customers far more effectively and even evaluate measure and monitor the impacts of marketing efforts. In effect, it leads to the organizations adopting Customer Relationship Management (CRM) to reach out and satisfy customers in a far more integrated process than they had ever done before.

However, to have effective Data base marketing to help firms to strategize, reach and acquire customers, in a manner referred to above, one needs to use the power of data and information technology to arrive at a focused target market from among the huge list of potential customers in a database called the house-list or house-file. The specially created promotional offerings would, thereafter, be directed to the focused target market instead of targeting such exercises over the entire list. This would help significantly to cut down on the expenditure because the focused market segment would have a much smaller list of recipients of the promotion. Also, since this process would enhance efficiency, the expected response rate (potential customers actually purchasing) would be much higher than that in the overall general market. This is actually the start of predictive modeling which helps identify customer segments with the highest future potential, who are the most profitable customers, who are likely to purchase again and who would spend the most amount of money .Predictive modeling would also help to prioritize prospects who start as potential customers and later a certain proportion of them graduate to actual customers.

This approach of predictive modeling is possible to achieve these days because today's technology enables organizations to collect and analyze massive amounts of customer related data. For instance, in predictive modeling, the marketer has to estimate and understand the relevant market segments, in terms of demographics, psychographics, lifestyle variables, operating & decision making styles and product use behavior. Such a characterization and segmentation is carried out using and analyzing large matrices having many data fields of key variables coming under the above categories and using various multivariate segmentation processes, starting with simple recency-frequency-monetary (RFM) analysis, to full scale logistics regression and Chi-square automatic interaction detection (CHAID) procedures . In other words, the segmentation carried out for predictive modeling encompasses heuristic approaches, statistical as well as Data Mining analytical procedures.

Such analysis enables the database marketer to identify as well as evaluate target market segments, develop positioning strategy and brand personality, choose key selling as also propositions and offer marketing strategy based on product or service itself.

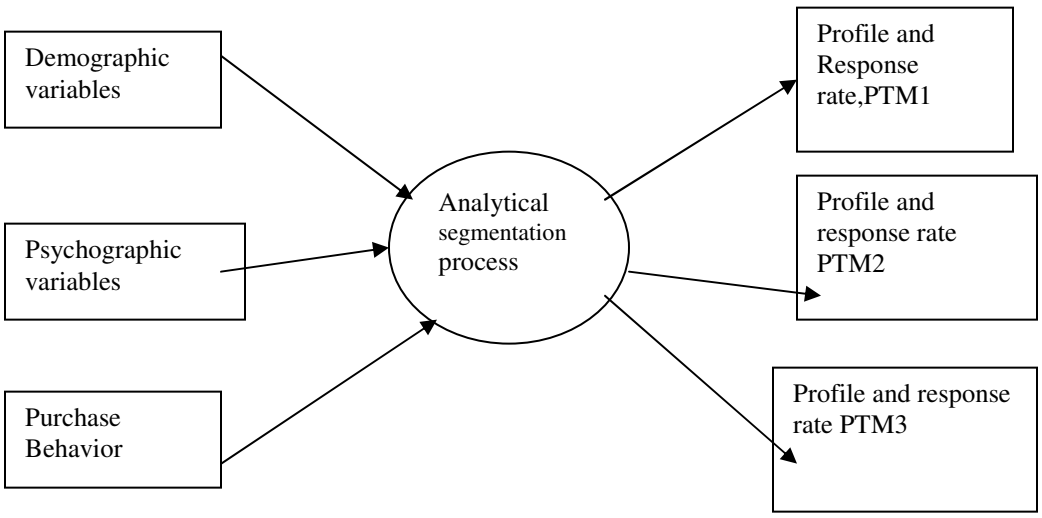


Figure1: Predictive modeling segmentation process under Database Marketing

The identification and profiling of target market segments which emerge from the predictive modeling thus result in much more focused and efficient marketing campaigns and promotions. These efforts cut down on ineffective marketing endeavors leading to better bottom line, more effective use of promotional budgets and thus higher profits.

Predictive Modeling under Data Base Marketing

Essentially, database marketing is concerned with building and processing of a marketing database which includes relevant, and sometimes even seemingly not so relevant, information about the firm's customers .The purpose of building such a database is to enable the firm to learn about the customers' likes, dislikes, preferences and wants and about their purchasing habits and patterns so as to offer appropriate products and services. The database, therefore, includes purchasing history, demographic characteristics, lifestyle attributes, financial and credit data, psycho-graphic data and all other related information. The database marketing then coordinates the data in a coherent format which can then be stored, retrieved, processed, analyzed and used for generating insights for decision making by managers and marketers.

The common database marketing packages, which are currently available for such purposes, use relational database architecture to represent and link data elements in the database. Here the raw data is stored in tables with each table representing specific type of information, such as customers' name, address, purchase history etc which again are linked by one or more key fields such as the customers account number or some other identification number.

Harit Ayurvedic Products Inc.

Harit Ayurvedic Products Inc. was established at the very end of 1996 for the purpose of selling specialty ayurveda products through different marketing efforts involving a variety of channels including media advertising, such as TV, magazines, newspapers, billboards and posters on DTC/ private buses and lamp posts.

Right from the start, the company has been involved in manufacturing and selling of ayurvedic and herbal products such as ayurvedic medicines, ayurvedic cosmetics, herbal oil, herbal shampoos, ayurvedic slimming medicines and ayurvedic drug formulations. Because of having so many product lines in their offering, coordination and inventory control of all products have always been a cause of worry and, on and off, Harit Ayurvedic Products Inc. has considered computerization of its entire database of products.

Initially, they had only thought of creating a house-list of all such products and giving a category number for them. However, they are currently considering creating a much more comprehensive database which would help them track their product sales by category and by the last date of purchase.

Tanuleen… the new product

The need for a comprehensive database arose because of a new product category of weight loss facilitating drugs, Tanuleen, which Harit is thinking of introducing to its market. This new drug needs to be marketed and promoted to its potential customers and Harit is

thinking of sending a promotional brochure to a customer segment which would have high predicted response rate. This weight loss formulation, Tanuleen, would be naturally, ayurvedically and organically produced and would not have the perceived associated side effects of increasing chances of cancer, which many regular weight loss formulations are feared to have. All the same, this information needs to be communicated to Harit's market through rigorous promotional activities. As such, Harit now is keen to identify the specific target market segment it would send the promotional brochure to and also find out the profiles of people who would be interested in purchasing such natural remedy for people wanting to reduce weight. However, before deciding on the Database features the company first wanted to decide on what kind of predictive modeling approaches it could take recourse to.

Predictive Modeling Endeavors at Harit Ayurvedic Products.

In applying predictive modeling under Database Marketing, if the marketer already has access to a full-scale customer data base then his or her job is simpler. The marketer then just has to access the database and apply predictive modeling techniques.

In case of Harit, they did not have that privilege because it was just in the process of starting to set up the database. However, instead of looking at the lack of data scenario as a shortcoming, Harit looked upon this as an opportunity to set up the database in a manner as would be most relevant to the decision-making situation in the company. So to start with, Harit looked at the prospective decision making processes it would be involved in, which would constitute the foundation for the response model. Towards this Harit decided to consider the details of database marketing they would need to engage in for marketing / promoting their new product.

As mentioned earlier, predictive modeling encompasses the following approaches:

Heuristic approach … essentially the common sense, trial and error approach

Statistical approach … using statistical modeling theory, and

Data-mining approach… which is a combination of the above two approaches.

Harit Ayurvedic learnt about the predictive modeling and the approaches it follows through Vishnu, a young professional junior manager, actually the owner's son, who had attended the MBA program from a much reputed international management school.

During the MBA, he had opted for an elective on CRM - customer relationship management, where he learnt about predictive modeling. After coming back to India to join his family business, Harit Ayurvedic, Vishnu was determined to use modern management approaches to streamline his father's business for superior effectively and dynamism. While considering the introduction and marketing of the new product category of weight loss

facilitating drugs, Vishnu convinced his father to support the setting up of full-scale database for predictive modeling which expectedly would elevate the company to world class level.

RFM approach at Harit Ayurvedic Inc.

Vishnu decided to consider first the RFM approach under the Heuristic modeling, where respondents' previous purchase behavior is expected to determine the future purchase behavior. The primary variables here are recency of last purchase from the company (how long age the customer last made a purchase), frequency of purchase and monetary amount (cumulative) spent on purchasing products from the company. Frequency is how many purchases the customer has made within a specified time period, such as, average number of purchases per year. Monetary is the total rupees spent by the customer within a specified time period. Marketers consider these three variables to be critical in determining the likely probability of responding to direct mailing or promotions by an individual customer.

Vishnu decided on the following data fields for each respondent for RFM analysis:

Customer no.	Month since last Purchased from Harit	total purchases last 5 yrs from Harit	total Rs. spent last 5 yrs on Harit products
002351	2 months ago	54	Rs.42,000
…			

Once the database is operational, every time any customer makes a purchase all of the relevant fields would be automatically filled up. However, to start with, Vishnu hoped to collect a random sample of 400 respondents (margin of error 5%, level of confidence 95%) from the sampling frame comprising the houselist of customers of Harit .This list of customers would be generated from the stack of invoices which Harit stored and filed regularly. For RFM analysis, initially Vishnu would only need the three variables concerned, but Vishnu observed that since he was collecting primary data through sampling, he might as well collect additional data fields which would enable him to carry out statistical as well as data mining approaches to predictive modeling. So, he set up his questionnaire to collect data on all of the following data fields.

Table 1

The data fields used for predictive modeling:

Customer identification number
Name
Address
(X1) Income category
(X2) Age category
(X3) Gender
(X4) Educational background further categorized into low medium and high
(X5) Occupation
(X6) Month last purchased from Harit (R)
(X7) Total number of times customer purchased from Harit over last 5 years (F)
(X8) Total amount of money spent on Harit products over the last 5 years (M)
(X9) Total amount spent on Ayurvedic pain relieving oils over last five years
(X10) Total amount spent on Ayurvedic herbal cosmetic products, over last 5 years
(X11) Total amount spent on cough, cold, liver, headache, blood sugar, indigestion medicines
(X12) Total amount spent on medicines womens' problems and sexual problems
(X13) Total amount spent on cardiac and anti cholesterol medicines and so on.

In addition, Vishnu decided to include the critical data field which would determine the response rate as the willingness to purchase the new product, Tanuleen, the weight loss facilitator. This would be obtained as a dichotomous or binary variable (Yes/No type) from the respondents. Vishnu planned to compare the consumer profile of those who exhibited willingness to purchase the new product from those who did not. For this he would create deciles or quintiles for each of the R, F and M variables, so that first decile for Recency variable would constitute those who purchased most recently, second decile for Recency variable being those who purchased next recently, and so on. Similarly, he would create deciles for Frequency variable, where first decile would comprise people who purchase with the greatest frequency, second decile those with next greatest frequency, and so on. Similarly, deciles would be created on the Monetary variable with the first decile comprising people with largest monetary amount spent and so on.

Thereafter, a composite index would be created where the people with most purchasing potential would be those who belong to first decile for Recency, first decile for Frequency and first decile for Monetary .The promotion, or direct marketing campaign, could be directed at these people first. Of course, for each composite index, he would also determine the response

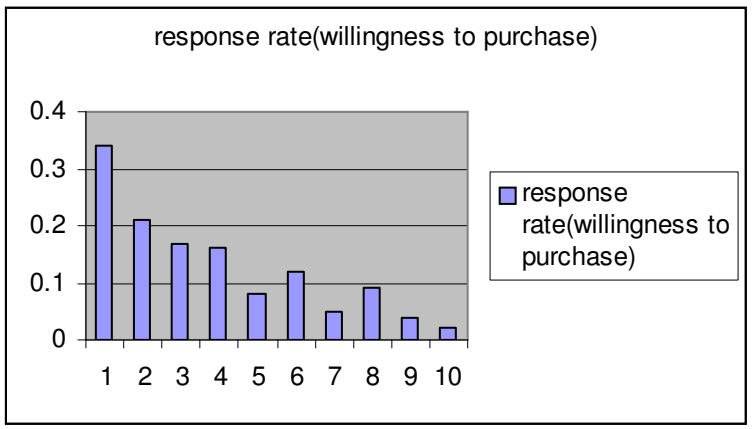
rate as computed from the willingness to purchase the new product. This would help to score segments of prospective respondents in terms of expected response rate.

Vishnu knew that RFM analysis could be carried out with regular Microsoft Excel software. However, in his MBA class he was exposed to SPSS and he found this software to be very elegant and comprehensive to use. Thus, for Harit he planned to acquire SPSS and use it for his RFM analysis.

Results of RFM analysis for Harit on the test sample data

On the regular basis the RFM analysis would be carried out in the full scale once the database builds up. All the same, Vishnu wanted to start the analysis on the random test sample he proposed to obtain. Implementing the sampling procedure, although the objective was to get a sample of size 400, the sample finally obtained was 240 respondents with sampling frame as the list of customers generated from the stack of invoices available with the company . The questionnaire used for the data included all the data fields referred to in Table 1, including the binary variable capturing willingness to purchase the new product Tanuleen. An attractive brochure for Tanuleen was also included as an attachment to the questionnaire, as is the case whenever willingness to purchase is to be ascertained.

Using only the RFM variables from the sample data, Vishnu carried out the analysis and came out with deciles on Recency, Frequency and Monetary variables. Thereafter, he computed composite scores. For instance, a composite score of 112 for a respondent would tell him that the respondent belongs to recency decile 1, frequency decile 1 and monetary decile 2. The response rates (willingness to purchase) for the composite scores were obtained as given below:



In the above diagram the horizontal axis refers to the RFM composite deciles and the vertical axis gives the response rate. Vishnu observed that the first two deciles appeared to be the most promising in terms of response rate measured by willingness to purchase.

Statistical approach … Logistics Regression under predictive modeling at Harit Ayurvedic

So, Vishnu established the foundation of RFM analysis in Harit Ayurvedic. Having done the RFM and predicting the response, Vishnu thought of utilizing the entire array of data fields he had acquired (Table 1) and go for more advanced modeling for predicting the response rate. This would lead him to explore the statistical modeling approach of logistics regression. In logistics regression, one is directly able to estimate the probability of an event occurring.

Conceptually, logistics regression is similar to multiple linear regression, but the dependent variable in logistics regression can only take on two values represented by Yes/No, 0/1, Purchase/not purchase etc. The mean of a 0-1 variable, referred to as dummy variable, can be interpreted as the proportion of 1 and thus can be interpreted as probability. This allows the predicted values in logistics regression to be interpreted as probabilities of purchase. In the case of a single predictor or an independent variable, the logistics regression model can be written as:

$$\text{Probability (event)} = \frac{e^{b0 + b1X}}{1 + e^{b0 + b1X}}$$

where b0, b1 are coefficients estimated from the data, X is the independent variable or predictor and e is the base of natural logarithm, approximately 2.718. If there are more than one independent variable, the model can be extended.

So that

$$\text{Probability (event)} = \frac{e^{b0 + b1Z}}{1 + e^{b0 + b1Z}}$$

where Z is the linear combination $Z = b0 + b1X1 + b2X2 + \ldots. BkXk$.

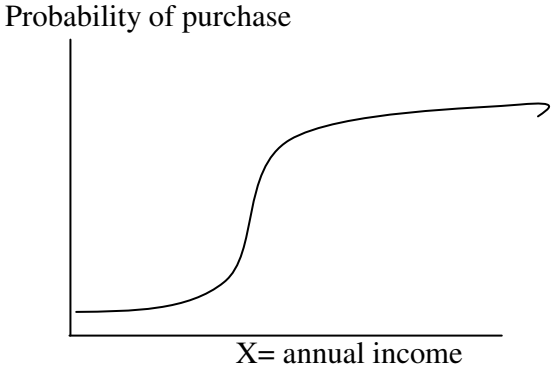
In logistics regression analysis, the parameters of the model are estimated using maximum likelihood method. That implies that the coefficients are estimated, ensure that the observed results are most likely. Logistics regression model is non linear and so an iterative algorithm is necessary for parameter estimation.

Data fields for logistics regression

Customer account no	X1	X2	X3	…probability of purchase

Please note that the last field, probability of purchase is the outcome and not input to the predictive model.

If y=1 refers to the purchase event, then in logistics regression, the probability of purchase is depicted on the y-axis, with x-axis providing, say, annual income of the respondent. Then the probability of purchase is given by an S-shaped curve as follows:



The curve has been drawn with only one independent variable whereas, in reality, the probability of purchase is determined by a host of independent variables, as given in Vishnu's data base. While running the logistics regression, Vishnu would use the foreward Likelihood Ratio method where significant variables are introduced one at a time until the best fit is obtained.

This approach would also score each prospective customer in terms of the highest probability of purchase, or in terms of highest willingness to buy. Once these probabilities are obtained, Vishnu can then create deciles on these probabilities and determine the profile of consumers who belong to the most attractive decile with the highest probability of purchase.

In this approach, the dependent variable Y was the binary (Yes/No) variable, where

The Probability (Y=1) is given as follows:

$$\text{Probability } (Y=1) = \frac{e^{b0 + b1X1 + b2X2 + b3X3 + \ldots}}{1 + e^{b0 + b1X1 + b2X2 + b3X3 + \ldots}}$$

or Probability (Y=1) = $\frac{1}{1+ e^{-(b0 +b1x1 + b2X2 + b3 X3 +.....)}}$

where X1, X2, X3… are the same independent variables as provided in Table 1.

This probability equation, along with foreward LR method would determine which independent variables significantly affect the probability of purchase which would help Vishnu to find out the profiles and other characteristics of the segment with the highest probability of purchase.

Assessing the goodness of fit of logistics regression model.

To assess the overall fit of the model, logistics regression provides different statistical measures and the associated significance. One such measure is -2 times the natural logarithm of the likelihood (-2LL) which has a Chi-square distribution. The smaller the value of -2LL the higher is the prediction of log odds, odds being defined as

Odds = probability (event)/ probability (no event).

If the significance of the associated Chi-square is less than 5%, the model fit is considered acceptable.

Logistics regression also provides two measures that are analogous to R-square in multiple regression. These are Cox and Snell pseudo R-square and Nagelkere R-square, either of which indicate a measure of the variance explained by the predictors in the model. Another measure of how well the model performs is in its ability to accurately classify cases into the two categories contained in the binary dependent variable. The assigning of predicted classes is done under the criterion that if a customer has a probability of purchase as more than 50%, the customer is assigned a predicted class of 'purchase'. The customer whose observed class matches with the predicted class has accurate prediction. The ratio of total number of accurate prediction in the sample to total number of cases is considered another measure of accuracy and the fit of the model.

Applying logistics regression to test sample from Harit Ayurvedic.

Upon applying the logistics regression approach to the test sample of 240 respondents, the model converged in 5 iterations with acceptable Chi-square and level of significance being less than 5%. The iteration started with Block 0 (not shown) where there were no independent variables in the model. Thereafter the Foreward LR method inputted the independent variables one by one.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	8.114	1	.004
	Block	8.114	1	.004
	Model	8.114	1	.004
Step 2	Step	5.813	1	.016
	Block	13.928	2	.001
	Model	13.928	2	.001
Step 3	Step	5.747	1	.017
	Block	19.674	3	.000
	Model	19.674	3	.000
Step 4	Step	4.701	1	.030
	Block	24.376	4	.000
	Model	24.376	4	.000

Please note that at every stage the Chi-square significance was very much acceptable.
The other indicators like –2LL, Nagelkerke R-square and Cox and Snell R-square were significant too. One may observe that -2LL decreased with each iteration; Cox and Snell R-square and Nagelkerke R-square increased with succeeding iterations.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	692.963	.016	.021
2	687.150	.027	.036
3	681.403	.038	.051
4	676.702	.047	.063

As mentioned earlier, for each customer in the sample, if the probability of purchase is >.05, the customer is classified in the 'will purchase' predicted category. A measure of accuracy of prediction is given by the percentage of customers correctly classified .In the case of Harit, the final model had the percentage of accuracy as 78.4 %.

Finally, the iteration summary below observes that the variables which most significantly affect the predicted probability of purchase are:

Gender (sex), income as categorical variable, recency and educational background.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	SEX	.316	.115	7.478	1	.006	1.371
	Constant	-.157	.117	1.793	1	.181	.854
Step 2(b)	INCOMCAT	.370	.155	5.676	1	.017	1.447
	SEX	.329	.116	8.032	1	.005	1.390
	Constant	-.664	.243	7.472	1	.006	.515
Step 3(c)	INCOMCAT	.371	.156	5.650	1	.017	1.450
	MONTHLAS	.045	.019	5.553	1	.018	1.046
	SEX	.320	.117	7.473	1	.006	1.378
	Constant	-.922	.269	11.759	1	.001	.398
Step 4(d)	EDUCAT	.121	.057	4.515	1	.034	1.129
	INCOMCAT	.368	.157	5.481	1	.019	1.444
	MONTHLAS	.043	.019	5.015	1	.025	1.044
	SEX	.323	.117	7.564	1	.006	1.381
	Constant	-1.177	.297	15.682	1	.000	.308

In the above table if B > 0, it implies that the probability of purchase would increase as X increases with exp (B) signifying the extent of increase. The Wald statistic gives the significance of independent variables in the model. If the significance of Wald is < .05, it implies that the independent variable has significant impact on the probability of purchase.

Data Mining approach … the CHAID method for predictive modeling.

While logistics regression held significant potential for determining the most promising market segment with the highest estimate of response rate prediction, Vishnu thought that while recommending its application to the board of directors of Harit Ayurvedic, he would also present another approach, the data mining approach, towards determination of the segment with highest predicted response rate. Essentially, data-mining approaches are used when the predictors or independent variables are too many.

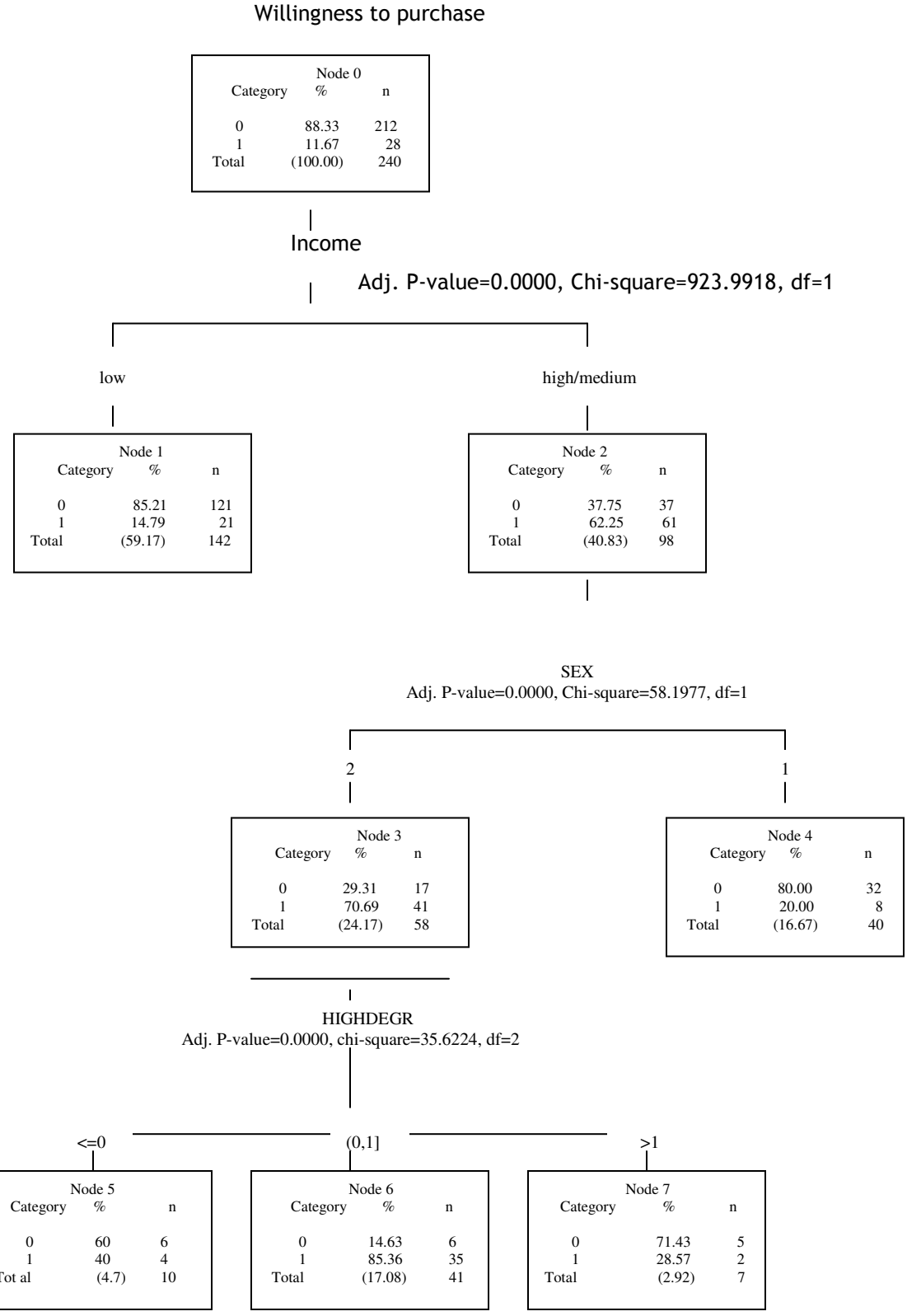
Data Mining approach, which is a combination of heuristic and statistical modeling, includes many algorithms which work efficiently in organizations having extensive data bases. Of these, recursive partitioning algorithms, or decision trees, are a versatile tool for finding out trends, patterns or relationships in the data. Vishnu knew that if he went for data mining approach, as above, he would have to have all the data fields, mentioned in Table 1 filled up by the respondents .In other words the data set must be complete with all the independent variables or predictors such as X1, X2…., as also the dependent variable Y, willingness to purchase, which in this algorithm we call the target variable.

Decision Tree with CHAID (Chi Square Automatic Interactive Detection) Algorithm in Harit Ayurvedic.

CHAID is a combination of heuristic as well as statistical method, which examines relationships between many categorical predictor variables and a categorical, usually nominal, target variable. It applies the Chi-square test on independence, also called Contingency table between the target variable and each of the predictor independent variable using the multi-way cross tab table. The null hypothesis H0: the two variables are independent.

This iterative process works with repeated application of Chi square test between target variable Y and each one of the different predictor variables. The predictor variable which gives the smallest p-value provides the basis for first partition from the root node. Thereafter the tree 'grows' following the same iterative process of partitioning by the Chi-square testing. The process of identifying the predictor variable with the smallest p-value is called the Bon Ferroni approximation.

The following application of the CHAID algorithm to Harit would clarify the process.
When Vishnu tried to apply the CHAID algorithm to his database, he did it on an experimental basis on his sample of 240 respondents.


Willingness to purchase
Node 0
Category % n
0 88.33 212
1 11.67 28
Total (100.00) 240
Income
Adj. P-value=0.0000, Chi-square=923.9918, df=1
low
high/medium
Node 1
Category % n
0 85.21 121
1 14.79 21
Total (59.17) 142
Node 2
Category % n
0 37.75 37
1 62.25 61
Total (40.83) 98
SEX
Adj. P-value=0.0000, Chi-square=58.1977, df=1
2
1
Node 3
Category % n
0 29.31 17
1 70.69 41
Total (24.17) 58
Node 4
Category % n
0 80.00 32
1 20.00 8
Total (16.67) 40
HIGHDEGR
Adj. P-value=0.0000, chi-square=35.6224, df=2
<=0
(0,1]
>1
Node 5
Category % n
0 60 6
1 40 4
Tot al (4.7) 10
Node 6
Category % n
0 14.63 6
1 85.36 35
Total (17.08) 41
Node 7
Category % n
0 71.43 5
1 28.57 2
Total (2.92) 7

Looking at the CHAID tree, Vishnu realized that the base rate for willingness to purchase was 11.67 % for the entire sample representing the target population. The first split below the root node is income category. This implies that of all variables income had the most significant relationship with willingness to purchase Tanuleen. He also saw that in the low income category the propensity to purchase Tanuleen was very small, only 14.79%, whereas in the medium and high income category, the propensity is far higher, 70.69%. However, the size of medium to high income category is only 40.83% of the total market and the size of the low income category is 59.17% of the market.

The medium to high income category is further split into two nodes based on the sex of the respondent. Of the remaining predictor variables, sex was the most significant predictor for this subgroup. For the male segment under the high/medium income class, the response rate is only 20%, whereas for the female segment under this income category, the response rate is much higher, 81.03%.

Vishnu thereby concluded that his weight control formulation, Tanuleen, has to be targeted to women in the high/medium category.

The tree further split once again, from the women segment with the educational background, called high degree, as the most significant predictor. This time there were three splits, the first split for high degree less than or equal to college degree, the second split was for high degree more than college degree but less than or equal to masters' degree, and the third split was for high degree greater than masters' degree, perhaps for doctoral or post-doctoral background. The response rates for the three segments were 40%, 85.36% and 28.57% respectively and sizes of the segments were 4.7 %, 17.08 % and 2.92 % of the entire sample representing the entire population.

Hence, Vishnu observed that the segment he should focus on, which would help him get the highest response rate of 85.36% comprises women belonging to an income category of high or medium and having an educational background of more than college degree, including but nor exceeding masters degree. This segment would approximately constitute 17% of the total target market.

Concluding remarks

There are many approaches and many algorithms which allow direct marketers and other researchers to carry out predictive modeling to determine market segments having most significant response rates .Of these approaches, only a few have been mentioned above. Also, within these approaches there are various individual approaches which work. For instance, within Data Mining approaches, only decision tree has been mentioned. Again, within decision tree, in addition to CHAID, there are so many others, like Exhaustive CHAID, C & RT (Classification and Regression Trees), QUEST (Quick, Unbiased, Efficient, Statistical Tree) etc. are also becoming popular and are applied depending on the perspective of the decision making situation. Of course, there are some new softwares which are slowly entering the market where all these approaches are available to the marketer together and one can try out all of these approaches and select the one which offers the closest fit.

However, the greatest challenge in many of these situations is not so much the applicability of the models but the availability of data, in the form of comprehensive databases, without which even a simple application of predictive modeling leads nowhere. So, how can this difficulty be overcome?

I feel it could be overcome as more and more Vishnu's get exposed to the advantages of predictive modeling, learn how to use them in a classroom scenario and then try to apply these concepts, plan, design, and build up the databases which would be required for such applications.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984).Classification and Regression Trees. Belmont, California: Wadsworth.

Lim, T., Loh, W., & Shih, Y. (1997). An empirical comparison of decision trees and other classification methods. Technical Report 979: 1-31. Madison.Wisconsin: Department of Statistics .University of Wisconsin.

Loh, W., & Shih, Y. (1997). Split selection methods for classification trees. Statistica Sinica, 7:815-840.