Target Market Decision Tree Analysis

Mrudula Anumala

ADTA 5120 Introduction to Data Analytics

University of North Texas

Introduction

This report discusses the analysis and decision tree regression model generated to understand what kind of customers they have. Aim is to identify customers that would tend to have a higher value to Wonderful Wines of the World using a sample of customers provided. And the principal objective is to help WWW to identify a customer profile to use a targeted marketing approach more efficiently. And analyzed five hypothetical customers of interest to determine if WWW should shift the direction of their marketing efforts by predicting their lifetime value. From the forecasts, as well as any other factors or data decided to include, recommended one of the five customer profiles for a targeted marketing attempt.

Dataset Description

The dataset Master customer file is provided by the Wonderful Wines of the World and it consists of the information of 10,000 people from its database who have purchased something from them in the past four years. It provided 66 potentially relevant parameters that can be used in generating a prediction model. It includes a customer estimated lifetime value variable (also called CLV) which is dependent and the rest all 65 variables are the independent ones which are grouped into categories(Demographics, Magazine Subscriptions, Financial Relationship with WWW, Wine preferences, Problems and Compliments, Questionnaire Answers, Specific Marketing Attempts). The variables contain nominal, ordinal, integer and ratio data.

Descriptive Statistics

Summary statistics were computed on the Age groups, Income groups, Customer has friended the WWW Facebook Page, Customer paid for the test wine tasting, and three other variables

mentioned below which are used to separate the different branches in our final decision tree model.

- TBOTTLES - Total number of bottles purchased by customer
- AVGPURCH - Average amount of a purchase by this customer
- AVGDISC - Average discount received by this customer, in percent

Computed the Univariate summaries below,

K	L	M	N	O	P	Q	R
	AGE	*INCOME*	*TBOTTLES*	*AVGPURCH*	*AVGDISC*	*FACEBOOK*	*WINETAST*
Mean	50.1274	98048.1766	17.8483	37.191935	20.702	0.0988	0.0807
Standard Error	0.143947869	467.9646812	0.347402674	0.311196179	0.282539064	0.002984082	0.002723875
Median	50	97165.5	3	27.905	0	0	0
Mode	64	135557	1	10.18	0	0	0
Standard Deviation	14.39478691	46796.46812	34.74026735	31.11961792	28.25390636	0.298408218	0.272387462
Sample Variance	207.2098902	2189909428	1206.886176	968.4306196	798.2832243	0.089047465	0.074194929
Kurtosis	-1.26748728	-0.6814723	17.94844656	1.629837831	-1.006850445	5.234305819	7.483699179
Skewness	-0.00672884	0.26945868	3.660770741	1.257093553	0.865839113	2.689471881	3.079318509
Range	66	249285	406	232.79	75	1	1
Minimum	18	20216	1	4.17	0	0	0
Maximum	84	269501	407	236.96	75	1	1
Sum	501274	980481766	178483	371919.35	207020	988	807
Count	10000	10000	10000	10000	10000	10000	10000

Computed the Multivariate summaries below,

L	M	N	O	P	Q	R
	AGEGRP1	*AGEGRP2*	*AGEGRP3*	*AGEGRP4*	*AGEGRP5*	*AGEGRP6*
Mean	-48.85967742	-35.03843137	-27.00242978	81.32703837	300.5187581	349.7939943
Standard Error	17.29421718	2.915244419	3.295039004	10.20667467	14.98310736	15.85780455
Median	-24.27	-28.05	-28.05	-14.02	53.63	78.33
Mode	-19.63	-28.05	-16.83	-19.63	-14.02	-14.02
Standard Deviation	96.29012609	123.1669205	156.0541048	416.8523895	694.7381077	730.6645853
Sample Variance	9271.788383	15170.09029	24352.88361	173765.9146	482661.0383	533870.7362
Kurtosis	4.042751944	27.34273477	33.69453757	15.51479776	7.815691982	5.771477299
Skewness	-0.411153531	0.601162526	2.103839561	2.54314853	1.576185129	1.737354449
Range	570.09	2316.15	3260.43	6277.78	9627.23	7739.01
Minimum	-329.87	-1149.97	-1517.6	-2401.62	-4293.4	-2618.04
Maximum	240.22	1166.18	1742.83	3876.16	5333.83	5120.97
Sum	-1514.65	-62543.6	-60566.45	135653.5	646115.33	742612.65
Count	31	1785	2243	1668	2150	2123

M	N	O	P	Q	R	S
	INCGRP1	*INCGRP2*	*INCGRP3*	*INCGRP4*	*INCGRP5*	*INCGRP6*
Mean	-59.86766667	-42.9293	-38.2271	4.008831	98.11429	452.4625
Standard Error	13.40039632	1.896859	3.767416	6.387712	9.249001	14.60949
Median	-33.66	-30.175	-28.05	-21.88	7.98	149.02
Mode	-39.27	-28.05	-19.63	-22.44	-16.83	-14.02
Standard Deviation	103.7990235	83.11622	147.074	263.6044	390.4349	801.6606
Sample Variance	10774.23729	6908.305	21630.78	69487.29	152439.4	642659.8
Kurtosis	23.83852367	41.87565	16.9751	13.2235	6.891441	4.749795
Skewness	-4.340245677	-4.18774	-0.34846	0.531432	0.984256	1.380524
Range	790.67	1701.97	2273.69	3957.65	5046.48	9627.23
Minimum	-692.31	-1257.41	-1149.97	-1986.27	-2401.62	-4293.4
Maximum	98.36	444.56	1123.72	1971.38	2644.86	5333.83
Sum	-3592.06	-82424.2	-58258.1	6827.04	174839.7	1362364
Count	60	1920	1524	1703	1782	3011

H	I	J
	Friended the WWW Facebook Page	*Haven't friended the WWW Facebook Page*
Mean	-57.88016194	161.6669308
Standard Error	7.973072478	5.776531589
Median	-32.84	-6.68
Mode	-19.63	-16.83
Standard Deviation	250.6133399	548.3751221
Sample Variance	62807.04612	300715.2746
Kurtosis	20.66119365	13.63560262
Skewness	-1.588394149	2.637732106
Range	3679.16	9627.23
Minimum	-2208.09	-4293.4
Maximum	1471.07	5333.83
Sum	-57185.6	1456942.38
Count	988	9012

H	I	J
	Paid for Wine Tasting	*Haven't paid for Wine Tasting*
Mean	769.5645601	84.70773197
Standard Error	34.94259349	4.448883838
Median	486.9	-16.83
Mode	-16.83	-16.83
Standard Deviation	992.6403002	426.5595763
Sample Variance	985334.7657	181953.0722
Kurtosis	1.97975457	17.89128417
Skewness	1.077497322	2.462122469
Range	7888.34	9414.37
Minimum	-2554.51	-4293.4
Maximum	5333.83	5120.97
Sum	621038.6	778718.18
Count	807	9193

Anomalous Data

Decision trees are robust to outliers, because they segregate points by lines. So how much ever a point is far from the lines, it doesn't make any difference. Apparently, as the nodes are determined based on the portion of the sample in each split section and not on their absolute values, outliers or the anomalous data will have negligible effect.

Decision Tree Methods

There are three Growing methods available in SPSS. They are CHAID, Exhaustive CHAID and CRT.

CHAID: CHAID stands for Chi-squared Automatic Interaction Detection which is a method used for constructing decision trees (O'Connor, 2015). It can create nonbinary trees indicating

that it can have splits of more than two branches. This method identifies optimal splits. This method is best used when there are a lot of categorical variables present in the dataset. A predictor variable is chosen, and splitting is performed based on that variable and finds out whether it is resulting in statistically significant value or not (McCormick & Salcedo & Peck & Wheeler, 2017). If we get the statistically significant value, then that significant factor is considered to make a split for the next part of the tree. This process continues until no statistically significant result is obtained. For example, if input contains more than two categories, these are compared, and categories which show no differences in the outcome are collapsed together. Then pairs of least significant differences are joined. In the case of nominal input fields, any categories can be combined. For ordinal data, contiguous categories are merged.

Exhaustive CHAID: In this method, all possible splits for each predictor are considered. As every possible split is considered, it takes more time to compute (Song & Lu, 2015).

Significance Level (CHAID/Exhaustive CHAID):

When it comes to significance level, for the CHAID and Exhaustive CHAID methods, we can control the tree's splitting nodes and merging categories. We know that default level of significance is 0.05 but for splitting nodes it can be greater than 0 and less than 1. If the value is less i.e., near to 0 then less nodes will be produced in the tree and vice versa. For merging categories, the value must be greater than 0 and less than or equal to 1 and if we want to prevent merging of categories then value of 1 must be taken which in the case of scale independent variable means that number of intervals in the final tree is equal to the number of categories for the variable. (Ye & Chen & Chen & Liu & Zhang & Fan & Wang, 2016).

Scores:

For both CHAID and Exhaustive CHAID the scores of ordinal dependent variables for each category of the dependent variable can be assigned and these scores define the distance between categories and the order of the dependent variable.

Method-Dependent Rules:

CHAID and Exhaustive CHAID view all system and user missing values as a single category for each independent variable. This category may or may not subsequently be combined with other categories of that independent variable, depending on the criteria, for scale and ordinal independent variables.

CRT: CRT stands for classification and regression trees. In this method, it splits the data which are homogeneous with respect to the dependent variable. This method solves the problem of surrogate splits that are handling missing values. Out of all the trees obtained, this method is providing the tree with less risk value and good prediction results (Bagozzi, 1994).

Methods Used

We have tried with all the above three methods, but the risk values are low while using CHAID and CRT method comparatively. In the CHAID method, the risk values are 60408.684 for the training sample and 48399.611 for the test sample. In the CRT method, the risk values are 43280.864 for the training sample and 44972.037 for the test sample. As predicting the LTV is our objective, we used the CRT method because it has low risk value comparatively.

Derivation of Tree

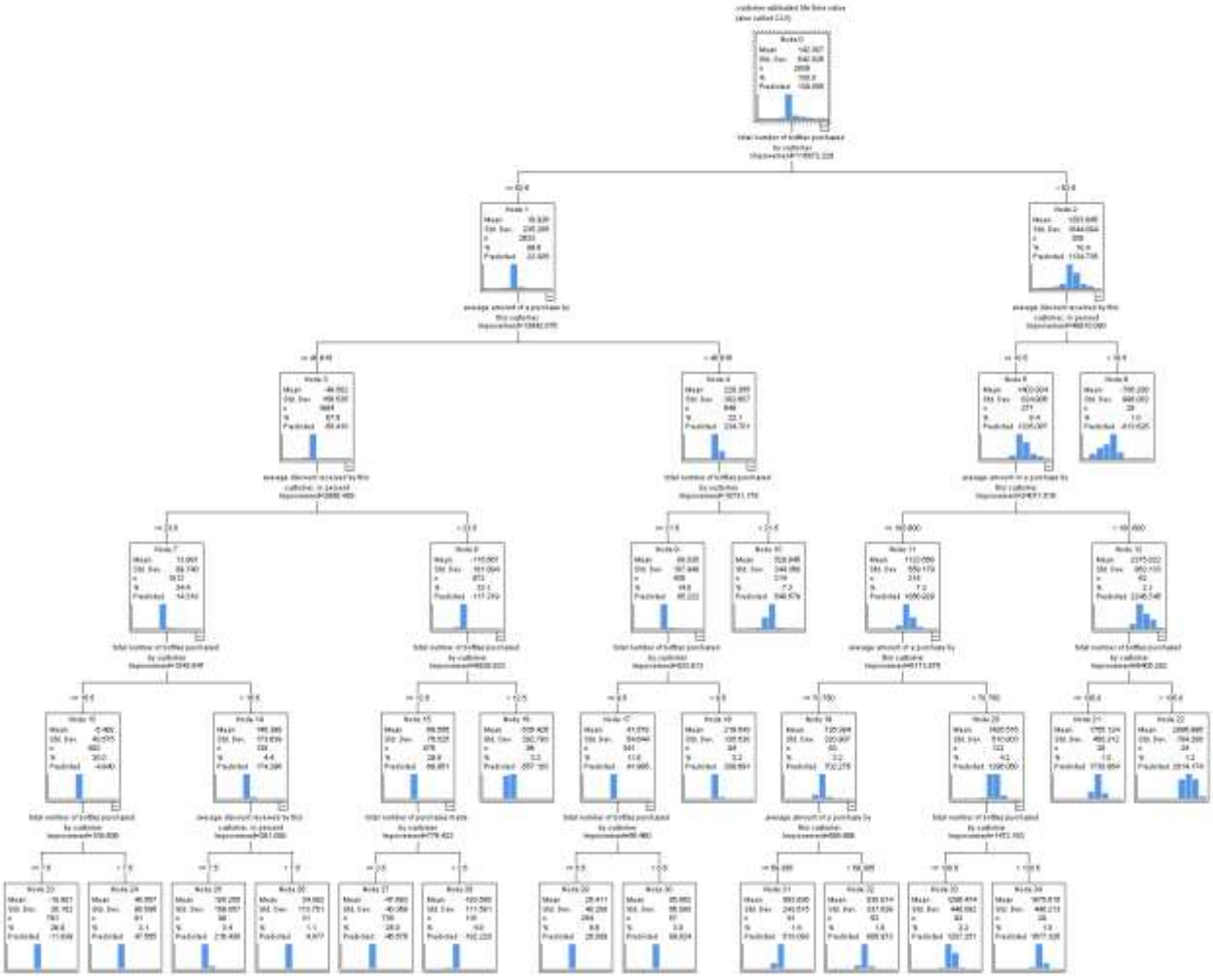
For building a decision tree, the key focus is on splitting the data into training and test samples and inputting the number of cases for the parent node and child node. In our case of obtaining a decision tree, we took training and test samples with a split 70% and 30% of the dataset. While using CHAID and Exhaustive CHAID method, the levels under the root node are limited to three and using CRT method, it is limited to five. Based on the 100/50 number of cases for the nodes, decision is made whether a split should be made from that node or not. From the final tree, it is observed that the important variable on which decision tree is built is the total number of bottles purchased by the customer.

Model Summary		
Specifications	Growing Method	CRT
	Dependent Variable	customer estimated life time value (also called CLV)
	Independent Variables	customer estimated age between 18 and 24?, customer estimated age between 25 and 34?, customer estimated age between 35 and 44?, customer estimated age between 45 and 54?, customer estimated age between 55 and 64?, customer estimated age 65 or over?, estimated customer household income, customer estimated household income $0 to $25,000, customer estimated household income $25,000 to $50,000, customer estimated household income $50,000 to $75,000, customer estimated household income $75,000 to $100,000, customer estimated household income $100,000 to $125,000, customer estimated household income above $125,000, at least one child under 13 at home?, at least one child aged 13-19 at home?, number of subscriptions to gourmet magazines, number of subscriptions to specifically wine magazines, number of subscriptions to home decor magazines, number of subscriptions to sports magazines, total number of purchases made by customer, total number of bottles purchased by customer, total number of catalogs sent to customer, average amount of a purchase by this customer, average discount received by this customer, in percent, average quality rating of wine purchased by this customer, total number of complaints about late deliveries by customer, total number of complaints about wrong wines by customer, total number of positive comments about wine by customer, I trust Wonderful Wines to care about me as a customer (on 10% random sample), Customer has friended the WWW Facebook page, Customer paid for the test wine tasting
	Validation	Split Sample
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	total number of bottles purchased by customer, total number of purchases made by customer, estimated customer household income, average amount of a purchase by this customer, total number of complaints about late deliveries by customer, total number of complaints about wrong wines by customer, total number of positive comments about wine by customer, average quality rating of wine purchased by this customer, customer estimated household income above $125,000, Customer paid for the test wine tasting, number of subscriptions to gourmet magazines, customer estimated age 65 or over?, number of subscriptions to specifically wine magazines, number of subscriptions to home decor magazines, average discount received by this customer, in percent, at least one child under 13 at home?, customer estimated household income $25,000 to $50,000, number of subscriptions to sports magazines, at least one child aged 13-19 at home?, customer estimated age between 25 and 34?, customer estimated age between 35 and 44?, Customer has friended the WWW Facebook page, customer estimated household income $100,000 to $125,000, customer estimated age between 55 and 64?, customer estimated household income $50,000 to $75,000, total number of catalogs sent to customer, I trust Wonderful Wines to care about me as a customer (on 10% random sample), customer estimated household income $0 to $25,000, customer estimated age between 18 and 24?, customer estimated household income $75,000 to $100,000, customer estimated age between 45 and 54?
	Number of Nodes	35
	Number of Terminal Nodes	18
	Depth	5

Risk

Sample	Estimate	Std. Error
Training	43280.864	2877.758
Test	44972.037	3996.033

Growing Method: CRT
Dependent Variable: customer estimated life time value (also called CLV)



Relative Importance of Variable

Variable importance score is computed within the CRT method itself by using the improvement measure attributable to each parameter in its role as either a primary or substitute splitter. The values of all these improvements are aggregated over each node and totaled.

- TBOTTLES - Total number of bottles purchased by customer: 145469.495
- AVGPURCH - Average amount of a purchase by this customer: 44677.033
- AVGDISC - Average discount received by this customer, in percent: 44057.607

Now they are scaled relative to the best performing variable. The variable with the highest sum of improvement i.e., TBOTTLES is scored 100 and the other two will have decreasingly lower scores. (Machuca & Vettore & Krasuska & Baker & Robinson, 2017).

Informative Discussion of Variables in Tree

In the derived decision tree, the parent node is the total number of bottles purchased by the customer. If this value is >52.5, the variable on which parent node is split is the average discount received by this customer in percent. If this value is >10.5, the predicted LTV is -813.625 which is lower, so ignoring this case. If the value is <=10.5, the predicted LTV is 1335.087 which is one of the highest top 5 predicted LTV values and further split is made based on the variable average amount of purchase by this customer. If that variable value is >100.600, then the predicted LTV value is 2245.745 which is the one of the highest among all the LTV values. Then further split is made based on the variable total number of bottles purchased by the customer. We can say that it is the variable which has the highest influence because we have predicted LTV higher in that case. The Customer LTV values are 2814.174 and 1738.954. So, we can conclude that the

variables which are highly affecting the lifetime values in our model are the average amount of purchase by this customer and the total number of bottles purchased by the customer.

Characterization of Solution

Provided below the Predicted Lifetime Customer Value and corresponding node number along with the total number of bottles purchased by customer, Average amount of a purchase by this customer and Average discount received by this customer, in percent.

CUSTID	TBOTTLES	AVGPURCH	AVGDISC	PREDICTED LTV	NODE NO.
20001	16	-9.41	119	-557.10	16
20002	1	-17.23	4	-11.84	23
20003	6	-32.61	80	-46.58	27
20004	3	.68	31	-46.58	27
20005	3	15.94	86	-46.58	27

Other Decision Tree Model Developed

The second-best model we achieved with low risk values is by using the CHAID model. In this case, parent node is the total number of bottles purchased by the customer. The highest values of LTV are obtained when splitting occurs in the following manner: If the total number of bottles >56.0, then lifetime value is 1195.505 which is one of the top five lifetime values. Then further splitting is done by the variable average amount of purchase by this customer and if that value is >81.140, then the predicted LTV is 1885.842 and if it lies between [63.190,81.140], then the predicted LTV is 975.393. The top two predicted lifetime values are for the case when the customer paid for the test wine tasting. If he has paid, the LTV is 2062.122 and if he hasn't paid, then LTV is 1686.052. Based on all the results obtained, we can say that the variables total number of bottles purchased by the customer, average amount of purchase by this customer and

customer paid for test wine tasting are the most influential in predicting the lifetime variable of the customer in this method. Even in this case, split of the training and test sample are 70% and 30% of dataset. The minimum number of cases considered for the parent and child node are 100 and 50 respectively.

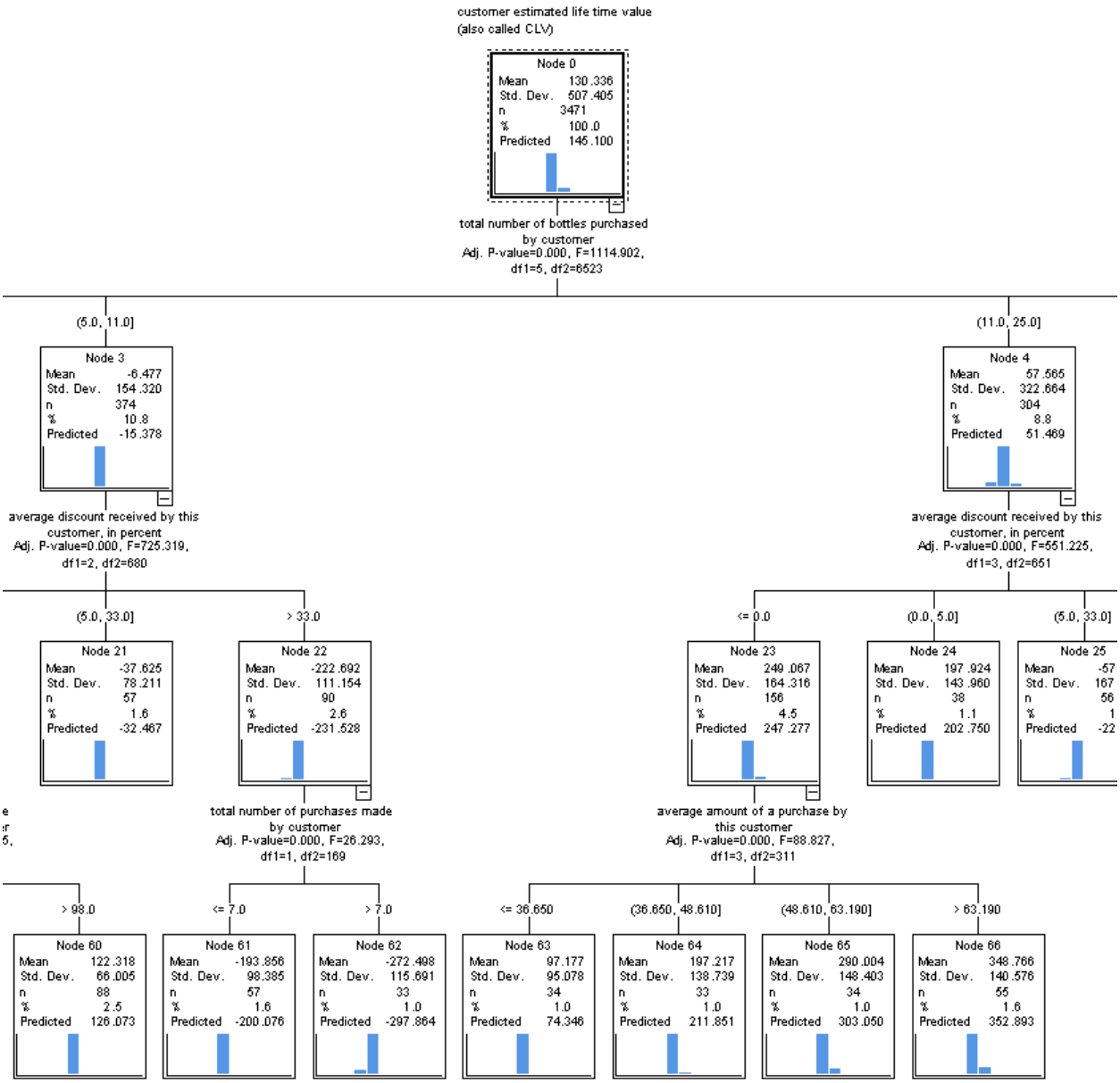
Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	customer estimated life time value (also called CLV)
	Independent Variables	customer ID number, customer estimated age, estimated years education, estimated customer household income, number of subscriptions to specifically wine magazines, number of days since customer made first purchase, number of days since customer made last purchase, total number of purchases made by customer, total number of bottles purchased by customer, total number of catalogs sent to customer, average amount of a purchase by this customer, average discount received by this customer, in percent, average quality rating of wine purchased by this customer, total number of complaints about late deliveries by customer, date of order for most recent complaint about late delivery (number of days ago), total number of complaints about wrong wines by customer, date of order for most recent complaint about wrong wine (number of days ago), total number of positive comments about wine by customer, date of order for most recent compliment about wine by customer (number of days ago), customer estimated household income $0 to $25,000, customer estimated household income $25,000 to $50,000, customer estimated household income $50,000 to $75,000, customer estimated household income $75,000 to $100,000, customer estimated household income $100,000 to $125,000, customer estimated household income above $125,000, at least one child under 13 at home?, at least one child aged 13-19 at home?, customer estimated age between 18 and 24?, customer estimated age between 25 and 34?, customer estimated age between 35 and 44?, customer estimated age between 45 and 54?, customer estimated age between 55 and 64?, customer estimated age 65 or over?, customer sex or imputed sex , Customer paid for the test wine tasting, Customer has friended the WWW Facebook page, number of subscriptions to sports magazines, number of subscriptions to home decor magazines, Customer Age Group Code, Customer Income group code, number of subscriptions to gourmet magazines
	Validation	Split Sample
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	total number of bottles purchased by customer, average amount of a purchase by this customer, total number of purchases made by customer, number of days since customer made first purchase, average discount received by this customer, in percent, total number of catalogs sent to customer, average quality rating of wine purchased by this customer, Customer paid for the test wine tasting
	Number of Nodes	77
	Number of Terminal Nodes	54
	Depth	3

Risk

Sample	Estimate	Std. Error
Training	60408.684	5051.465
Test	48399.611	4468.649

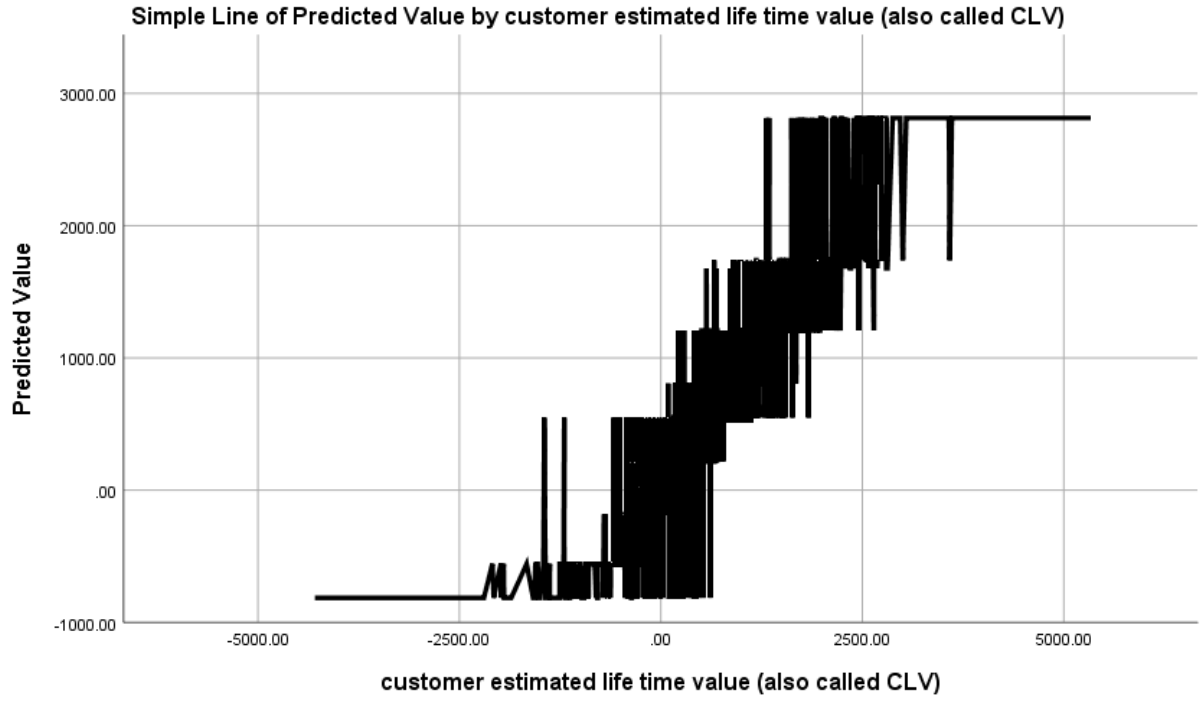
Growing Method: CHAID
Dependent Variable: customer estimated life time value (also called CLV)



CHAID Decision Tree.png

Predictive Capability (Risk)

Models have been validated using the testing dataset. Predictive capability depends significantly on the quality of data. And in our model, risk is estimated as 43280.864 for training data and 44972.037 for the test data.



As customers are assigned with the values only from the set of predicted LTV values obtained at the nodes of the decision tree based upon the node conditions, we won't have continuous values and so is the reason why we obtained the comparison graph as above which is almost linear which in turn implies that the predicted LTV is almost same as that of actual value. Therefore, we have a high predictive capability.

Validation Procedure

There are two validation methods available: cross validation and split-sample validation. Validation will allow us to assess how well our tree structure generalizes to a larger population.

We used the split-sample validation in the decision tree regression model which is generated by using a training sample and tested on a hold-out sample. And within the split-sample validation, we generated the model using a custom variable to split the sample but then as we have a fixed set of test and training data, in all the models, risk is almost around 65,000 for both the samples and we didn't find much variance in multiple runs. So, we tried using random assignment by varying the sample's training, test split as - (50%,50%), (60%,40%), (70%,30%) and achieved a better model for 70-30 split. Apart from these, with less training sample, parameter estimates have greater variance and with less test sample, performance statistics have greater variance.

Extra Insight

From the final decision tree model by CRT method, we predicted the lifetime values of the five hypothetical customers based on the predicted values. The variables which are affecting the LTV are total number of bottles purchased by customer, average discount received by this customer, average amount of purchase by this customer. And there are some other variables which can be taken into consideration by using different growing methods like customers paid for wine tasting and the total number of catalogs sent to the customer. The values of the risk vary when we change the split among training and test samples. So, we tried the possibilities and when we got a better risk value, we fixed the data split percentage and predicted the LTV.

From the model generated we can suggest that the best customer profile would be the one with,

- Total number of bottles purchased by customer > 52 (and >135 more specifically)
- Average discount received by this customer in percent <= 10.5
- Average amount of purchase by this customer <= 100.6

Therefore, customers with the above properties have high probability to show inclination towards wonderful wines of the world. And if the number of bottles purchased by the customer are high, it automatically indicates that they are happy with the services offered by the wonderful wines of the world.

References

[1] O'Connor, A (2015) An analysis of the predictive capability of C5.0 and Chaid decision trees and Bayes net in the classification of fatal traffic accidents in the UK. DIT 2015

[2] McCormick, K., Salcedo, J., Peck, J., & Wheeler, A. (2017). *Spss statistics for data analysis and visualization*. Indianapolis, IN: John Wiley & Sons, Inc.

[3] Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

[4] Ye, F., Chen, Z. H., Chen, J., Liu, F., Zhang, Y., Fan, Q. Y., & Wang, L. (2016). Chi-squared Automatic Interaction Detection Decision Tree Analysis of Risk Factors for Infant Anemia in Beijing, China. *Chinese medical journal*, *129*(10), 1193–1199. https://doi.org/10.4103/0366-6999.181955

[5] Bagozzi, R. P. (1994). *Advanced methods of marketing research*. Cambridge, MA: Blackwell Business.

[6] Machuca, C., Vettore, M. V., Krasuska, M., Baker, S. R., & Robinson, P. G. (2017). Using classification and regression tree modelling to investigate response shift patterns in dentine hypersensitivity. BMC medical research methodology, 17(1), 120. https://doi.org/10.1186/s12874-017-0396-3