Analyzing the customer's personality to provide the best services Using decision tree

Nibras Talib Mohammed,

nbrass.t@uokerbala.edu.iq
Department of statistics, University of Kerbala , Kerbala , Iraq.

Abstract: *This research will analyze a dataset containing 22 important customer attributes, such as education, marital status, income, and the amount spent on various products on a regular basis, such as fruits, meat, and gold. The dataset also includes catalogs, whether a promotion or discount affects customers, as well as the number of purchase sources customers use, such as decision tree analysis is a powerful tool that can be used to analyze customer personality and understand Using* ***.****stores, and locations their behavior. This technique helps in deducing relationships between various attributes and assessing how these relationships might influenceresponses. By examining the results extracted from the decision tree analysis, it is clear that there is a variation in positive behavior among different customer groups. Some nodes showed high gain rates when asked specific questions about responses or indicators, indicating that the groups within these nodes were the mostresponsive. Optimistic are the common characteristics of these customers. All customers in these nodes, which are generally more responsive, have similar characteristics needs or consumption patterns. Identifying and capitalizing on these characteristics can represent a and may share common s significant opportunity for a company when developing marketing and communicationstrategies. On the other hand, there are other nodes with low profitability but a large number of customers. These nodes may represent customer groups with different structures or characteristics that make them less likely to respond. It may be useful to study these nodes in more detail to understand what motivates these people and learn how to attract them.*

Keywords: decision tree, Customer personality

Introduction

Customer personality analysis is an important practice in marketing and sales, whereby the individual traits and characteristics of customers are studied and deduced with the aim of providing services or products tailored to their specific needs and desires. The process involves collecting a variety of customer data, such as demographic, behavioral, psychological, and even consumer patterns[1].

Studying customer behaviors and preferences helps companies design personalized their target audience experiences and tailored content that resonates with[2]Additionally, . personality analysis can help identify the most effective media channels and develop marketing messages that motivate customers topurchase[3].

competitive markets, where This strategy is particularly important in today's increasingly targeted marketing or "personalized marketing" creates additional value, enabling brands to enhance customer loyalty and build strong, sustainable relationships.

2- Research problem

The research problem focuses on how to analyze customer personality to improve marketing planning and effective communication. Companies require innovative methods to understand customers, not only through purchase data, but also through their behavior, interactions, and personal preferences. This topic explores the challenges of predicting purchasing behavior and examines advanced methods, such as big data analysis and psychographics, to analyze customer personality more deeply and thus Create a more personalized and convenient experience.

3-The concept of decision tree

It is a graphic representation of the elements and relationships that make up a decision problem business organization in order to address a specific problem in the practical reality of[4]

It is a quantitative, pictorial and graphic method for the elements and relationships that make up the problem, in light of different risk situations and natural situations[5]In light of these . definitions, we must point out an important issue: the graphic form of the decision tree is considered a guide and a guide for the decision maker in terms of the natural situation or investment opportunity that achieves the best results and the lowest costs and risks

3.1 Types of decision trees

1-and no/yes This method relies on the presence of variables of the type :Classification trees aims to divide the data into groups and then choose between them. In this type of data, the type, the results are answers are not limited to the answers mentioned above. Rather, in this presented in the form of two categorical options, with no other options. For example, determining the gender of individuals male/female or determining the opening of a new branch opening the branch/not opening the branch

2-This type targets variable objectives. A variable objective :continuous change Trees of cannot be defined or answered with a definitive answer. That is, the value placed in the tree example, an can change rapidly, and the value cannot be limited to specific probabilities. For individual's income value may be a variable objective depending on their age, job position , and other factors. In such cases, continuous variable trees, also known as regression trees, are used

4-Components of a decision tree

tree can be made simply and hierarchically, but in general there is a relatively fixed A decision form for the components of a decision tree

Different forms are used to indicate

- .Indicates the decision node to be made :Square
- .Circle: represents the opportunity or possible outcome of the decision
- .Lines: Lines in a decision tree indicate possible or existing actions
- -Crossed lines: Crossed lines are placed on the base lines to indicate imposed and non .implementable measures
- .tree Triangle: The final result of the decision

5-Steps to draw a decision tree

The decision tree is not drawn randomly, but according to specific rules and steps in light of the data available about the problem. The more expressive and accurate the graphic form A contributing and essential and its branches, the more is about the origin of the problem factor in reaching a solution.In general, there are sequential steps used in the process of Decision tree .drawing and analysis

It can be explained in the following steps:

1. (Determine decision points and the number of available alternatives (strategies
2. Determine the probability points and the number of states of nature available on the .The tree .origin and branches
3. Attaching information to the root and branches of the tree, including the expected returns for each state of nature, as well as stating the probability of these states being realized.
4. each of the Calculate the amount of achieved return or expected financial value for existing branches in order to clarify the idea of its use or application as one of the Decision making methods.

Mstage decision tree-ulti

making -This type of decision tree is more complex than the previous case, as the decision process takes place in several stages. This method is used to address complex problems, as situations in the stages of providing a solution to the multiple maker faces-the decision problem, in which it is necessary to take decisions subsequent to the first decision that was solving process, and it is necessary to take -adopted at the beginning of the problem each branch, the expected results decisions subsequent to the first decision, and at the end of are calculated based on the probability of achieving that branch or the state of nature, as The following forms and formulas shown in:

Regular network

Decision points and probability points lie on the same vertical plane and form a symmetry where each decision point is associated with an equal number of probability points... and likewise each probability point is associated with an equal number of decision points.

Irregular network

Decision and Decision points and probability points are not similar in their relation to the Common in practice probability points, noting that this type of decision tree is the most

6-Practical side

adopted in the research, where a sample of 2240 observations was chosen. A set of data was The data set was taken From the open source data reference [6]the variable chosen is the dependent variable

while (0) ,The value 1 is chosen if the customer responds to the advertising campaign does not respond to the campaig represents if the customer

Some variables were defined with a value of 1 if the customer agrees and a value of 0 if he disagrees

deleted Note that the number of variables was 29 variables, the unimportant variables were and only 22 variables remained

The data were analyzed in SPSS statistical program.

Analysis outputs-1

Model Summary Table 1

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	Response
	Independent Variables	Complain, Z_Revenue, AcceptedCmp2, AcceptedCmp1, NumWebVisitsMonth, NumStorePurchases, NumCatalogPurchases, NumWebPurchases, NumDealsPurchases, MntGoldProds, MntSweetProducts, MntFishProducts, MntMeatProducts, MntFruits, Recency, Teenhome, Kidhome, Income, Marital_Status, Education, Year_Birth
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50

Model Summary

Results	Independent Variables Included	AcceptedCmp1, Recency, NumWebVisitsMonth, Marital_Status, Income
	Number of Nodes	14
	Number of Terminal Nodes	8
	Depth	3

.The first table shows a summary of each model

1-CHAID Method:

The CHAID algorithm, proposed by statistician Cass in the late 1970s, is one of the most popular statistical methods supervised decision learning, trees are essentially grown as a multivariate dependency For method

To detect the association between the categorical dependent variable and multiple independent variables that can be categorical, use CHAID.

Or metric (in this case, encoding and converting them to categorical variables must be done in advance

CHAID .refers to the automatic and recursive tree procedure

Maximum Tree Depth ,,,….(Where the maximum tree depth was equal to 3

Minimum Cases in Parent The minimum number of cases in the parent node is equal to (100

Minimum Cases in Child nodes is equal to 50-The minimum number of cases in the sub

The results included the following three independent variablesResponse The tree was divided : into5

Number of Nodes The number of nodes is equal to 14

Number of Terminal Node The number of nodes is equal to 8

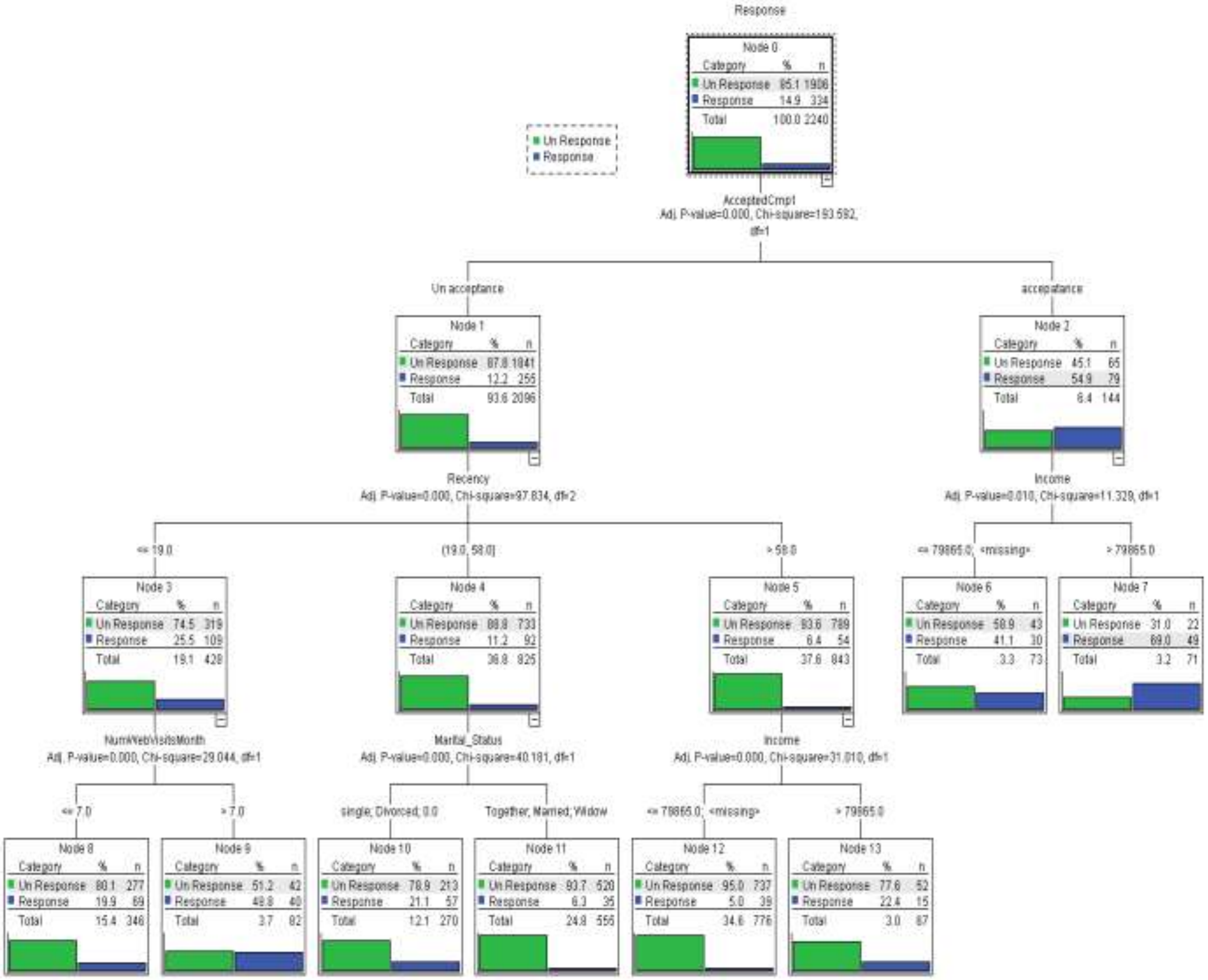
Depth Depth equals 3

Dependent variable (Response) Dependent variable

Independent Variables Independent variables are the following table

T	Independent Variables	Definition
1	Complain	complains
2	Revenue	profit
3	AcceptedCmp1	Advertising Campaign 1
4	AcceptedCmp2	Advertising Campaign 2
5	Web Visits Month	Web Visits Month
6	Store Purchases,	Store purchases
7	Catalog Purchases,	Catalog purchases
8	Web Purchases,	Web purchases
9	Deals Purchases,	Purchase deals
10	Gold Products	gold products
11	Sweet Products,	Sweet products
12	Fish Products,	fish products
13	Meat Products	meat products
14	Fruits	Fruit
15	Recency	Modernity

16	Teen home	adolescence
17	Kid home,	Children's age
18	Marital _Status	marital status
19	Year Birth	the age
20	Education	Academic qualification
21	Income	income



The above figure represents a decision tree or regression tree[2] .

Response means responding to the advertising campaign :
.un Responseresponse to the campaign Advertising-Non

dependent \The first section of the tree Response rate is equal to 85.1, response-The non : the response rate is equal to 14.9, the total rate is equal to 100%, and the total number of cases for the variables is equal to 2240

The second section of the tree: The independent variableAccepted Cop1 .)
into two parts) Node1, Node2

Node1 response rate equals 87.8, response rate equals 12.2, and the total rate equals 93.6-Non :
The total number of cases for the variables is 2096, theChi-square value is 97.837, and df =2.

Node2 equals 54.9, and the total rate equals 6.4 response rate equals 45.1, response rate-Non :
The total number of cases for the variables is equal to 144 and theChi-square value is equal to 11.329 anddf = 1

The third section of the tree: The independent variableRecency was divided into
into three sections Node3, Node4, Node5

The fourth section of the tree: The independent variable Income was divided .
into two partsNode6, Node7

web visits month .was divided, meaning the month of web visits
into two partsNode8, Node9,

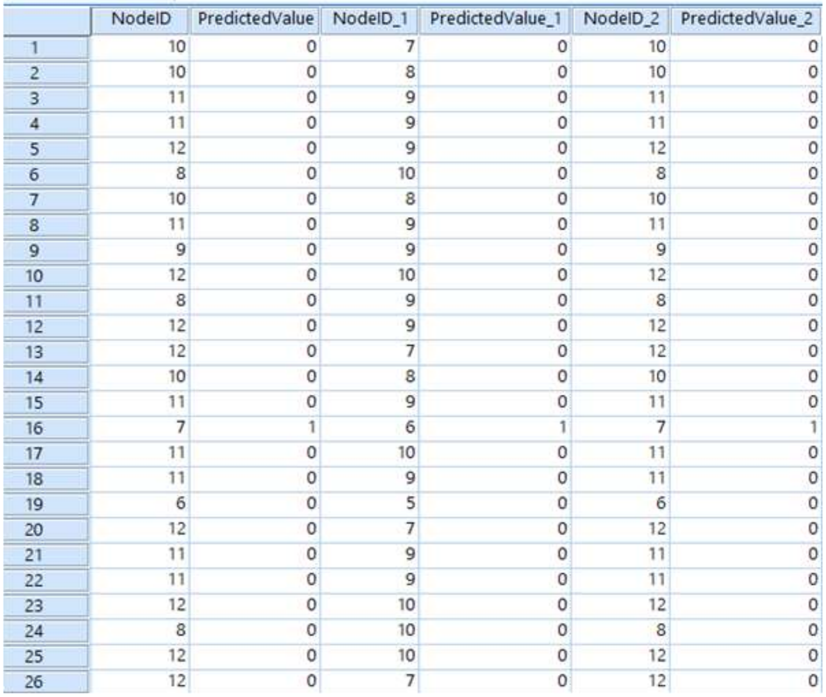
The sixth section of the tree: The independent variableMarital-status marital status means into two partsNode10, Node11

Income was divided into two sectionsNode 12; Node 13

Note (note that the p-value = 0.000 (in all tree divisions

reviewing the data, we find that the program created two variables, the first being AfterNode ID.and predicted value

Node ld .means the number of the tree branch to which each case belongs :
predicted value .means the expected value for each case :



	NodeID	PredictedValue	NodeID_1	PredictedValue_1	NodeID_2	PredictedValue_2
1	10	0	7	0	10	0
2	10	0	8	0	10	0
3	11	0	9	0	11	0
4	11	0	9	0	11	0
5	12	0	9	0	12	0
6	8	0	10	0	8	0
7	10	0	8	0	10	0
8	11	0	9	0	11	0
9	9	0	9	0	9	0
10	12	0	10	0	12	0
11	8	0	9	0	8	0
12	12	0	9	0	12	0
13	12	0	7	0	12	0
14	10	0	8	0	10	0
15	11	0	9	0	11	0
16	7	1	6	1	7	1
17	11	0	10	0	11	0
18	11	0	9	0	11	0
19	6	0	5	0	6	0
20	12	0	7	0	12	0
21	11	0	9	0	11	0
22	11	0	9	0	11	0
23	12	0	10	0	12	0
24	8	0	10	0	8	0
25	12	0	10	0	12	0
26	12	0	7	0	12	0

Table 1: Contract summary and expected values -:3

Node 6, fromIncome, responses-has 58.9 non.
Node 7 fromIncome .has a response rate of 69.0
Node 8, coming fromthe web visits map, .response cases-has 80.1 non
Node 9, coming fromthe web visits map, .response cases-has 51.2 no
Node 10, which comes fromMarital-status, .response cases-consists of 78.9 non
Node 11, which comes fromMarital-status, .response cases-consists of 93.7 non
Node 12, fromIncome, .response cases-consists of 95.0 non
Node 13, coming fromIncome, .response cases-consists of 77.6 non

Note (These interpretations were taken based on the tree and the previous table)

Target Category: Response

Gains for Nodes

Node	Node		Gain		Response	Index
	N	Percent	N	Percent		
7	71	3.2%	49	14.7%	69.0%	462.8%
9	82	3.7%	40	12.0%	48.8%	327.2%
6	73	3.3%	30	9.0%	41.1%	275.6%
13	67	3.0%	15	4.5%	22.4%	150.1%
10	270	12.1%	57	17.1%	21.1%	141.6%
8	346	15.4%	69	20.7%	19.9%	133.7%
11	555	24.8%	35	10.5%	6.3%	42.3%
12	776	34.6%	39	11.7%	5.0%	33.7%

:Table II: nodes Earnings Summary

Gain is a measure of the effectiveness of a node in improving the predictive ability and is It used to construct and prune the tree. Gain helps in choosing the questions or conditions that .separate the data in the best possible way

Gain here refers to a measure of the improvement in the model's predictive ability.
Indicates the number of nodes or points that have been analyzed: Node
Number of times a particular data partition has occurred at that node: N
Percentage of each node to the total number of nodes:
Number of responses or results that occurred in this branch: Response
Index An additional indicator or value that guides the arrangement or comparison of contracts

Node 7
=N 71
Gain= 69.0
.shows a high gain, indicating that this node has a significantly positive influence This node On the predictions in the model.

Node 9 :
=N 82
Gain =48.8
This node has good gain and relatively large response value and is considered an important results node in improving

Node 6
N=73
Gain: 41.1%
The profit of this node is relatively high compared to other nodes, which is evidence that the node has a significant influence on the model Compared to contracts that have lower profits

Node 13
N= 67
Gain =22.4%
node shows less gain, which means less influence in the model This

Node 10
N = 270
Gain 21.1%:
This node has a large response rate, making it important in the model

Node 8
N =346
Gain: 19.9%
means it has a big impact on This node has low gain but has the highest response ratio which .the model

Node 11
N=555
Gain: 6.3%
Although it is a large node, it has a relatively small gain, which means it has less impact on .improving results

Node 12
N= 776
Gain: 5.0%
gain This is the biggest knot but with the least Of all the previous contracts, which means that it has the least impact on the results

Risk

Estimate	Std. Error
.137	.007

Risk -Table III

The estimate is equal to 0.137, meaning the percentage is 13.7% ::.
.which is relatively small compared to the estimate itself 0.007 The error value is equal to :
Std. Error
This indicates that we are relatively confident in this estimate. That is, the smaller the More accurate and less susceptible to random .standard error, the better the estimate fluctuations in data

Classification

Observed	Predicted		
	Unanswered	Response	Percent Correct
Unanswered	1884	22	98.8%
Response	285	49	14.7%
Overall Percentage	96.8%	3.2%	86.3%

Table 4_ Prediction accuracy rate

The table shows the results of classification between two types of responses“Un Response” Response

1 The first row means that the model successfully predicted that there were 1884 cases that : did not respond correctly.
The number 22 means that the model was wrong 22 times when it predicted a response but in fact there was no response.
A percentage of 98.8% means that the model was very accurate in predicting "no response".

For the "response":

no “The second row means that the model made a mistake in 285 cases, as it predicted a : 2 when in fact it was a response ”response.
The number 49 means that the model correctly predicted 49 response cases.
slightly accurate in predicting the response A percentage of 14.7% means that the model was .

The overall prediction accuracy rate is 86.3%, meaning the model was accurate in about 86 .out of every 100 cases
.When predicting whether there will be a response or not

Conclusions

the data using decision tree analysis, it was found that the last decade During the analysis of was the one in which the prediction occurred
-The nodes are 13, 12, 11, 10, 9, 8, 7, 6 and the response status was in node 7 and the non response status was in the rest of the nodes

Node 7 is the best node in terms of response, with the lowest percentage of clients (3.2%) and achieving the highest response rate 14.7%
Its response index 69.0% and gain response index 462% are the highest, indicating that marketing
.was the most effective compared to other contracts In this category it
Node 12 is the largest node in terms of percentage 34.6%, but it has the lowest gain 5.0%, : which indicates that this node
It has less impact on improving the accuracy of predictions

References

[1]Publisher: -Saqa -Author Mohamed Hassan Al -The Comprehensive Guide to Marketing Krakeeb in Arabic 2024

[2]-Zouzou Fatima Al (The role of service quality in achieving customer satisfaction 2011-2010 Qassimi University profitable/Zahra

[3]A Conceptual Framework for a Modern Marketing - Customer Relationship Management Adel Hadi - Philosophy Ghazwane Salim
8-Page 7

[4]Prof. Dr. Saad Ali) - Decision Tree Analysis (Strategic Investment of Human Resources) Anzi-Al 89 Page - 2020 - Yazouri-Dar Al

[5]136-Page 133 -Ahmed Ragab) 2023 -Management -Decision Tree Analysis

[6]https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis