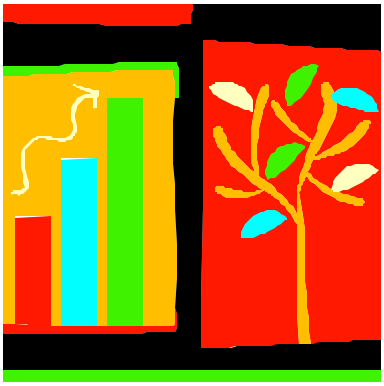
Kenan-Flagler Business School
The University of North Carolina



Predicting Response at BookBinders: Decision Trees

Recursive partitioning algorithms (or decision trees) are a versatile tool for uncovering patterns or relationships in data. They are especially useful when there is a large set of potential predictors and when you are not sure which are most important or what the relationships between the predictors and the target (dependent) variable are. In the case of a binary target variable, decision tree algorithms iteratively search through the data to find which predictors best separate the two categories of the target variable.

So far, we have used RFM segmentation and logistic regression to predict the response to the mailing offer for "The Art History of Florence." Now we will see how decision trees compare as an alternative.

Tree using Exhaustive CHAID

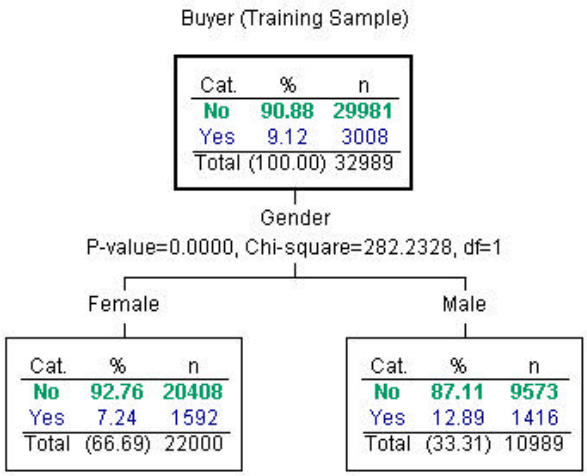
We'll start with a tree using exhaustive CHAID which is one of the algorithms in SPSS's AnswerTree software package. Our target variable is BUYER (whether or not they bought *The Art History of Florence*) and all other variables will be potential predictors. To see how the tree grows, let's take it one step at a time, beginning with the 'root node'. Because decision trees are prone to 'overfitting', we will split the dataset in two: two-thirds of the observations will be used to develop the model (the training sample) and the remaining one-third will be used to test the model (the validation or test sample) to see how well the model performs on 'new' observations.

Professor Charlotte Mason prepared this case to provide material for class discussion rather than to illustrate either effective or ineffective handling of a business situation. Names and data may have been disguised to assure confidentiality. The assistance of the Direct Marketing Educational Foundation in supplying data is gratefully acknowledged.

Exhibit 1 "Root" node for CHAID Tree (Training Sample)

The root node contains all the observations in the training sample. We see that 29981 or 90.88% are not buyers and the remaining 3008 or 9.12% are buyers. Next, we'll grow the tree one level. After specifying this, the AnswerTree software searches through the potential predictor variables to see which one 'best' separates the buyers from the non-buyers, and we see that gender is selected. These results are shown in Exhibit 2.

Exhibit 2 One Level CHAID Tree (Training Sample)



To see how well this one-level tree classifies buyers and non-buyers we can look at the classification table and 'risk estimate'. In AnswerTree, the risk estimate is the percent of customers incorrectly classified. Exhibit 3 shows the misclassification for this one-level tree.

Exhibit 3 Misclassification Matrix for One Level CHAID Tree (Training Sample)

Misclassification Matrix

		Actual Category		
		No	Yes	Total
Predicted Category	No	29981	3008	32989
	Yes	0	0	0
	Total	29981	3008	32989

	Training Sample
Risk Estimate	0.091182
SE of Risk Estimate	0.001585

Note that the tree predicts a total of zero buyers – meaning, so far, there are no nodes in the tree where the buyers outnumber non-buyers. Because our sample is so dominated by non-buyers, it's not surprising that the computer predicts non-buyers for many or even all nodes. Although the classification is correct for 91% of the cases, its predictions are not very useful for distinguishing good prospects from bad ones. To identify segments worth targeting (i.e., those with a relatively high probability of responding), we will need to look at the specific response rates for the nodes. First, though, let's grow the tree another level. Once again, the AnswerTree software searches the available predictor variables and selects which variable(s) to branch on for this level.

Exhibit 4 Two Level CHAID Tree (Training Sample)

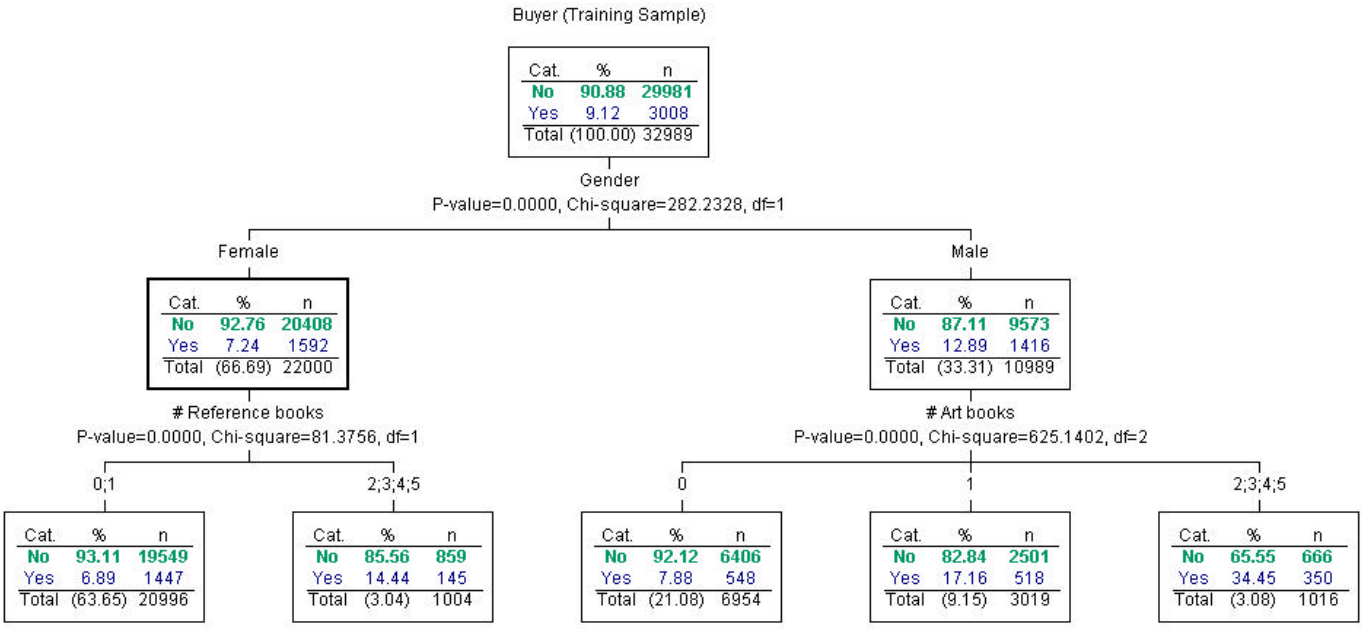


EXHIBIT 5

CHAID Tree

Based on 2/3rds of customer records used for 'training' sample

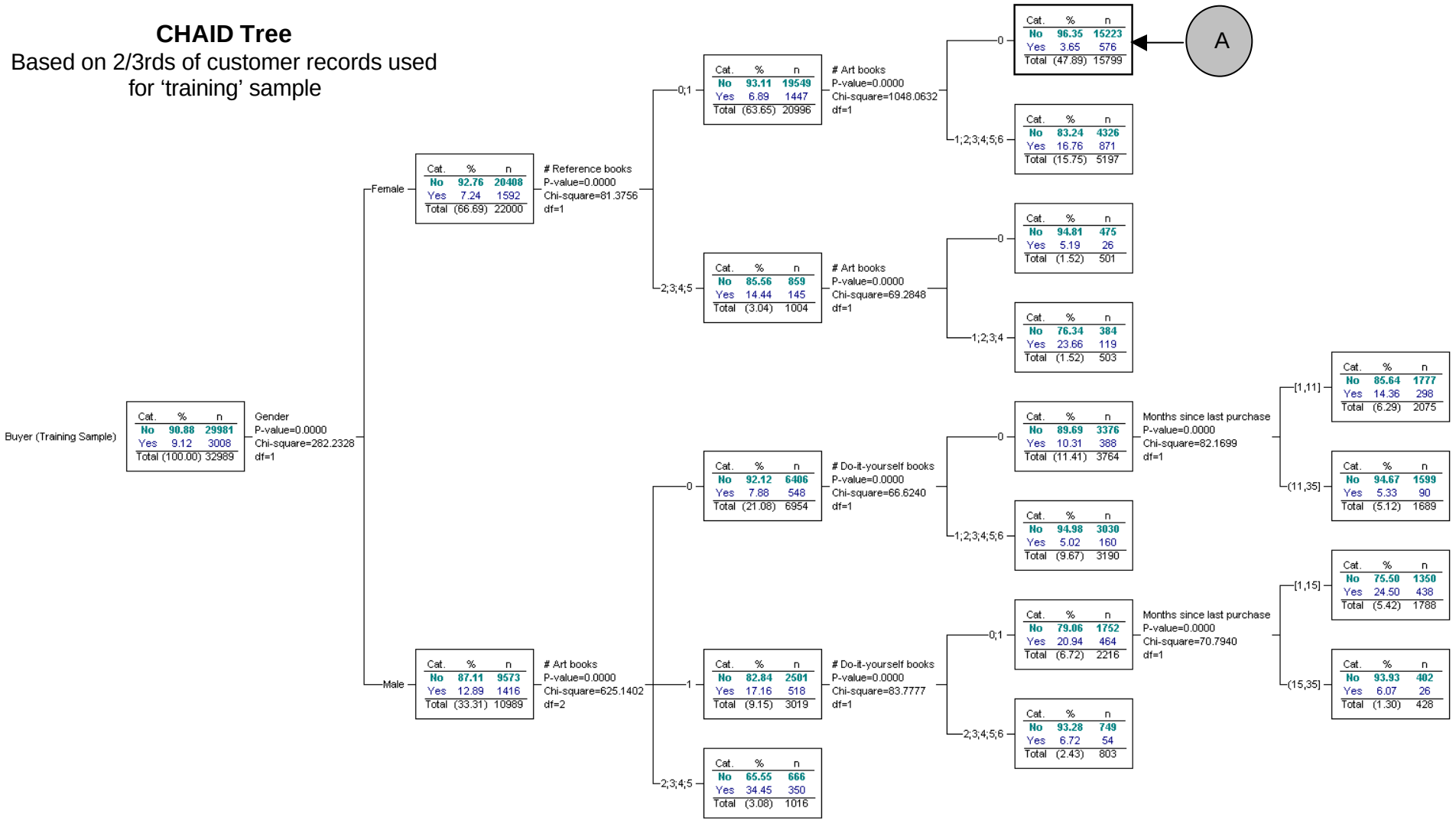
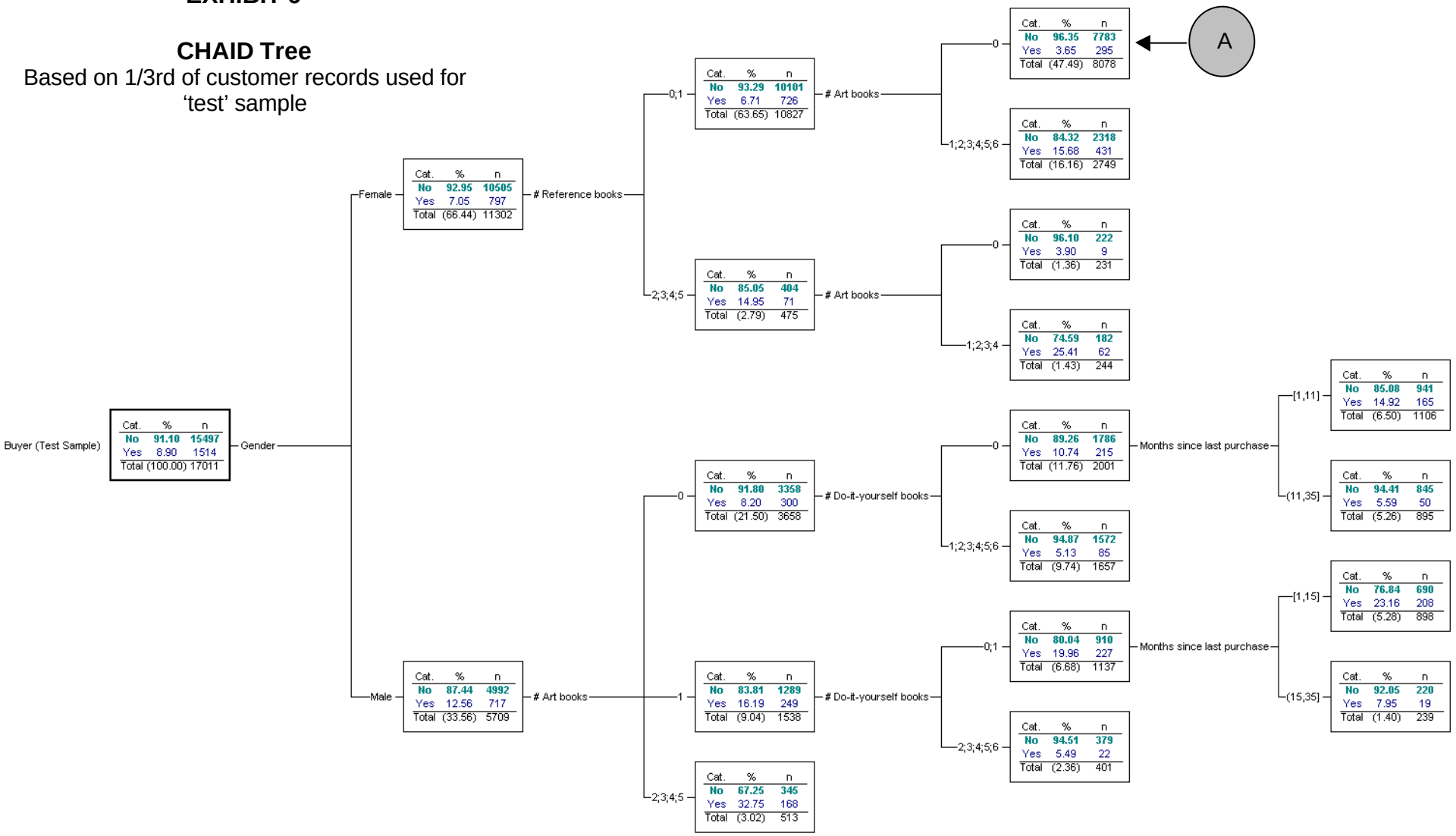


EXHIBIT 6

CHAID Tree

Based on 1/3rd of customer records used for 'test' sample



Note that for females the tree branches on the number of reference books purchased, but for males the tree branches on the number of art books purchased. Females who have purchased 2 or more reference books had more than twice the response rate (14.44% compared with 6.89%) than females who had purchased one or no reference books. Also, with CHAID algorithms there can be binary or multi-way splits (e.g. three or more branches from a single node) as seen above. The tree divides males into three categories or nodes based on the number of previous art books purchased. The response rate increases substantially with the number of prior art book purchases.

Exhibit 5 shows a complete tree after an iterative process of growing and pruning branches. So far, the analysis has used a random sample of 32,989 (equal to 66%) of the 50,000 customers in the dataset. These 32,989 records comprise the *training sample.* The remaining one-third, or 17,011 records, form the *validation* or test sample.

The results using the test sample are shown in Exhibit 6. The tree in Exhibit 6 uses the same branching rules as the tree we just developed using the training sample – and was used solely to classify the 17,011 customers in the test sample. Each customer in the test sample is 'put through' the tree starting at the root node and branching until a terminal node is reached. For example, if the first customer is a female with no past purchases of reference or art books, she will end up in the node labeled 'A' in Exhibit 6. We can see that there were a total of 8078 customers who fit that profile (females with no prior reference or art book purchases) and that 7783 of them did not purchase 'The Art History of Florence'.

Exhibit 7 shows a summary of the number of customers, the number of buyers and the response rate in each 'leaf' or terminal node for both the training sample and the test sample. The nodes in Exhibit 7 are sorted by response rate from highest to lowest.

Exhibit 7 Response Rates by Node

Training Sample			Test Sample		
# Customers	# Buyers	Response Rate	# Customers	# Buyers	Response Rate
1016	350	34.45%	513	168	32.75%
1788	438	24.50%	244	62	25.41%
503	119	23.66%	898	208	23.16%
5197	871	16.76%	2749	431	15.68%
2075	298	14.36%	1106	165	14.92%
803	54	6.72%	239	19	7.95%
428	26	6.07%	895	50	5.59%
1689	90	5.33%	401	22	5.49%
501	26	5.19%	1657	85	5.13%
3190	160	5.02%	231	9	3.90%
15799	576	3.65%	8078	295	3.65%

For example, node "A" in Exhibit 5 includes 15,799 customers from the training sample. Of the 15,799 customers in this node, 576 belong to the target category of Yes (i.e. they are buyers), which is a response rate of 3.65%.

.

The terminal nodes of the tree and the summary statistics in Exhibit 7 are used to identify which segments to target and which to avoid. An important decision is how 'deep' in the customer base to go. This decision may be based on the number of prospects wanted, a desired response rate or a desired proportion of potential buyers you want to contact, or profitability.

For the BookBinders' mailing offering *The Art History of Florence*, we know the following about profits for the two groups:

- Non-responder: -$0.50 for the cost of mailing
- Responder: $5.50 ($18 revenue less $9 COGS, $3 shipping and $0.50 for the mailing)

We can use this information to determine which nodes are profitable to target. An equivalent approach is to target customers in nodes with a response rate greater than or equal to the breakeven response rate.

Validating the Model

Decision trees are prone to overfitting – meaning that the tree is overly tailored or customized to the dataset used to create the tree. If this is the case, then the tree will do a substantially poorer job of predicting or classifying on a new set of data. To assess the performance of the decision tree on 'new' data, we use the one-third of the dataset that forms the validation or test sample. We expect the model to perform slightly worse on the test sample compared with the training sample – although the difference should be slight. A large discrepancy between the two suggests that the tree has been overfit and needs to be pruned back.

Case Questions:

1. Using the information in Exhibits 5 and 6, summarize – for the Director of Marketing – which customer groups should be targeted with this mailing.

2. Use the information in Exhibit 7 to make a cumulative gains chart for both the training and test samples. Does the tree appear 'overfit'? Why or why not?

3. Using the same costs as before ($18 selling price, $9 wholesale price, $3 shipping and $0.50 mailing costs), estimate what the gross profit (in dollars and as a % of gross sales) as well as the return on marketing would be if the "The Art History of Florence' offer were only mailed to those predicted by the CHAID tree results to be good prospects for this offer.

4. Compare and contrast the results and insights from using RFM, logistic regression and CHAID decision tree analysis for targeting buyers for BookBinders offers.