IRONWOOD INSIGHTS group, LLC

Artificial Intelligence (AI) and Machine Learning in Market Research

There continues to be a great deal of hype surrounding the use of artificial intelligence (AI) and machine learning in market research. The use of machine learning, (unsupervised learning), continues to expand rapidly in today's work environment.

Basically, machine learning is a field of computer science that gives computers the ability to learn from the data without being explicitly programmed.

Machine learning utilizes algorithms that learn from the data, organize the data, and in some cases even make predictions from the data. Many in the industry today refer to this ability for machine learning as AI. Some of the machine learning algorithms available to market researchers include:

- **Cluster analysis** – including k-means, hierarchical, and two step clustering.
- **Decision tree analysis** – including CHAID, CRT, and QUEST decision trees.
- **Artificial Neural Networks** – including Multilayer Perceptron and Radial Basis Function.

In this paper we want to explore two of the most commonly used machine learning techniques used by market researchers: cluster analysis and decision tree analysis. In the area of cluster analysis, we will briefly look at the k-means clustering algorithm. In the area of decision tree analysis, we will explore decision trees using the CHAID algorithm.

Cluster Analysis

Clustering involves creating groups of observations, or individuals, (clusters) and complying with the condition that individuals in one group differ visibly from individuals in other groups. The output from any clustering algorithm is a label, which assigns each individual to the cluster of that label.

Clustering uses an unsupervised machine-learning algorithm to identify patterns found within unlabeled input data. For example, when creating a new marketing campaign, it would be useful to segregate potential customers into subgroups based on information associated with each person to create a specific number of clusters so that the people in each cluster have similar features and demographics to each other and differ from the people in other clusters.

In the figure on the next page, we see people are segregated into four different clusters of different sizes. One cluster has 95 people, a second cluster has 55 people, a third cluster has 35 people, and the final cluster has 25 people. Individuals in each cluster may share demographic patterns, socio-economic patterns, or consumption patterns. The machine-learning clustering algorithms will find these patterns from the data and group people together in the best way possible.

[Figure 1]

Cluster Analysis Example

In our evaluation of data using cluster analysis, we are going to look at the purchasing habits of 440 individuals over 12 months in their local grocery store1. We have data collected on how much each individual spent during a 12-month period in six areas of the store:

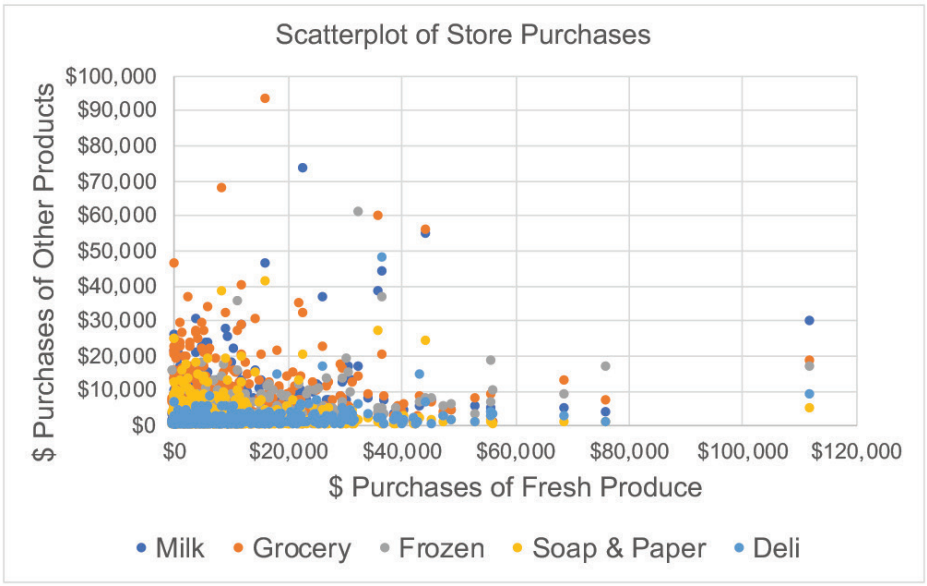
1. Fresh produce
2. Milk and dairy
3. General groceries
4. Frozen foods
5. Soap and paper products
6. Delicatessen

In the graph below we look at a scatter plot that shows the amount these 440 individuals spent in each of the six areas.

In Chart 1 below, we have created a scatter plot where the 12-month amount spent on fresh produce is shown along the horizontal (or X) axis in the chart. This is then matched along the vertical (or Y) axis with the 12-month amount spent on each of the five remaining types of products: milk and dairy, general groceries, frozen foods, soap and paper products, and delicatessen.

By just looking at the data as displayed in Chart 1 it is difficult to identify any pattern in purchasing habits among the 440 individuals. That is where cluster analysis becomes useful.

[Chart 1]



One of the most popular clustering algorithms is k-means. K-means clustering focuses on separating the instances into n clusters of equal variance by minimizing the sum of the squared distances between two points.

The k-means algorithm is used when we have data that has not previously been classified into different categories. The classification of data points into each cluster is done based on similarity, which for this algorithm is measured by the distance from the center (centroid) of each cluster. When the algorithm has completed running, we are presented with two major results:

1. every observation will have been assigned to a specific cluster, and

2. the centroid of each cluster will be identified.

The k-means algorithm works through an iterative process that begins with the analyst specifying a specific number of clusters to evaluate. An initial cluster centroid is randomly assigned and then all data points are assigned to the nearest cluster in the data space by measuring their distance from the centroid. The objective is to minimize the distance to each centroid. Centroids are calculated again by computing the mean of all data points belonging to the cluster. This process continues through an iterative process until the clusters stop moving. This is known as convergence of the clusters.

In our grocery market purchasing data, we end up setting the number of clusters to three. We initially tried 2, 3, 4, and 5 clusters before settling on three clusters.

In the table below we can see the results of the iterations of cluster centering after specifying three clusters.

[Table 1 – Iterations]

	Change in Cluster Centers		
Iteration	1	2	3
1	13754.326	32490.984	34613.308
2	431.326	18895.454	9694.856
3	766.749	14961.332	6854.911
4	762.472	8038.314	7093.278
5	664.632	5110.432	4948.693
6	455.630	3066.460	3965.664
7	452.483	2401.321	4010.549
8	234.016	758.631	2507.054
9	101.378	264.186	1454.365
10	216.873	268.103	2713.716
11	149.512	244.743	1521.349
12	101.194	0.000	922.667
13	99.297	0.000	878.497
14	104.033	0.000	858.424
15	198.471	0.000	1458.431
16	154.850	0.000	1008.344
17	106.801	244.743	633.913
18	93.314	268.103	336.662
19	0.000	0.000	0.000

Here we see that on iteration 19 the centroids of the 3 clusters stopped moving and the k-means clustering algorithm reached convergence.

One of the statistics that we always want to look at when conducting k-means clustering is the Analysis of Variance (ANOVA) shown in Table 2 below. From this table we can see that all six measures of purchasing habits significantly contributed to the assignment of the shoppers to one of the three clusters. This can be seen by the significant F statistic for each variable, as displayed in the far-right column of the table.

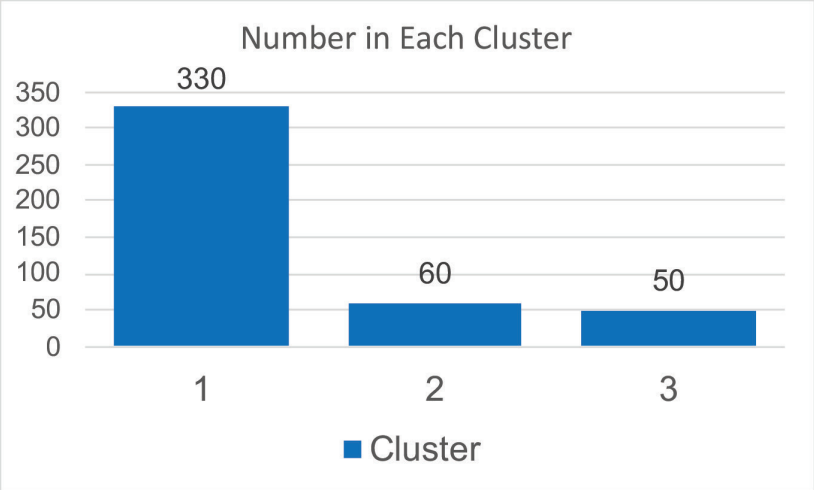
[Table 2 – ANOVA]

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Fresh Produce	19911731735.740	2	69557779.557	437	286.262	.000
Milk and Dairy	4685156442.931	2	33276894.124	437	140.793	.000
General Grocery	10886109792.836	2	40901409.525	437	266.155	.000
Frozen Foods	467966534.577	2	21533991.924	437	21.732	.000
Soap and Paper	2573047079.610	2	11060515.471	437	232.634	.000
Delicatessen	107710075.603	2	7496443.365	437	14.368	.000

As mentioned above, one of the results of the clustering process is that each shopper is assigned to one of the final clusters, in our case, one of three clusters.

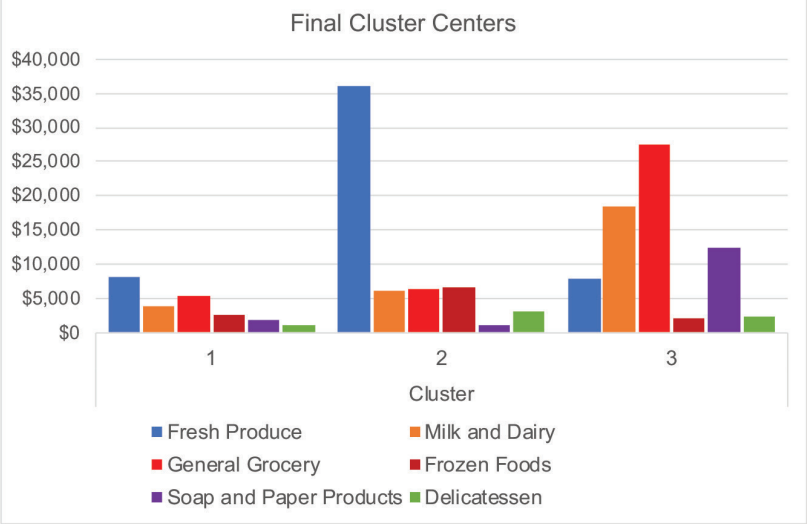
As we can see below, most shoppers fell into cluster 1, which represents 75% of all shoppers.

[Chart 2]



So how do these clusters differ and what does it tell us about the shopping habits of these 440 individuals? To help us with that, we now look at the final cluster centers for the six product types. This can be seen here in Chart 3.

[Chart 3]



From Chart 3 we see the different purchasing habits of the three cluster groups.

Cluster 1, which makes up 75% of all shoppers, spends about the same amount of money across the six product types. They purchase slightly more fresh produce and general grocery during a typical 12-month period.

Cluster 2, which makes up 14% of all shoppers, spends significantly more in the fresh produce section of the grocery store. As we can see in the chart above, the average shopper in cluster 2 spends more in the produce section of the store than in the entire rest of the store combined.

Cluster 3, which makes up 11% of all shoppers, spends more in the general grocery section of the store as well as the milk and dairy section of the store.

From the chart above we might come to the following conclusions:

Cluster 1 is made up of shoppers that are looking to get the best deal on all of their shopping needs. These types of shoppers are likely to frequent "big box" discount warehouse stores (such as Costco or Sam's Club) for the bulk of their shopping needs.

Cluster 2 is made up of shoppers who want to prepare mostly fresh produce for their family. These shoppers are going to frequent stores with the best selection of fresh produce.

Cluster 3 is made up of shoppers who spend most of their money purchasing general groceries along with milk and dairy products. These shoppers are going to frequent grocery stores where they can purchase a large selection of different products in a single location, such as Walmart.

Decision Tree Analysis

Our second machine learning tool is decision tree analysis. Decision tree learning creates a visual decision tree graph as a predictive model which maps observations, or individual respondents, that help to predict an individual's likely outcome. In medicine, decision trees can be used to predict who is likely to have a stroke or a heart attack. In finance, decision trees can be used to predict who is most likely to default on their mortgage. In customer experience, decision trees can be used to predict which customers are most likely to be unhappy and go to a competitor (or churn). For our decision tree example, we are going to look at the customer churn issue and we are going to focus on decision tree analysis using the CHAID algorithm. CHAID stands for Chi-Square Automatic Interaction Detector. CHAID:

- Is commonly used in the market research industry,
- Can easily handle a large number of independent variables,
- Is easy to understand, and
- Handles missing values well.

For our example, we are going to look at 1,000 telecommunications customers2 to see if we can find patterns within the data that will help us to predict which customers will churn (switch to a competitor) in the next month. The data file has 41 unique data fields that help to understand each customer. These variables include both continuous data such as months they have had service, their age, and years at their current address, as well as categorical data such as marital status, level of education, and whether they rent equipment or not.

The data file has a single target variable called churn. This variable identifies whether or not the customer switched to a competitor in the past month (yes or no). When attempting to understand customer preferences and develop a customer service model, we usually begin by trying to identify which customers are most likely to churn.

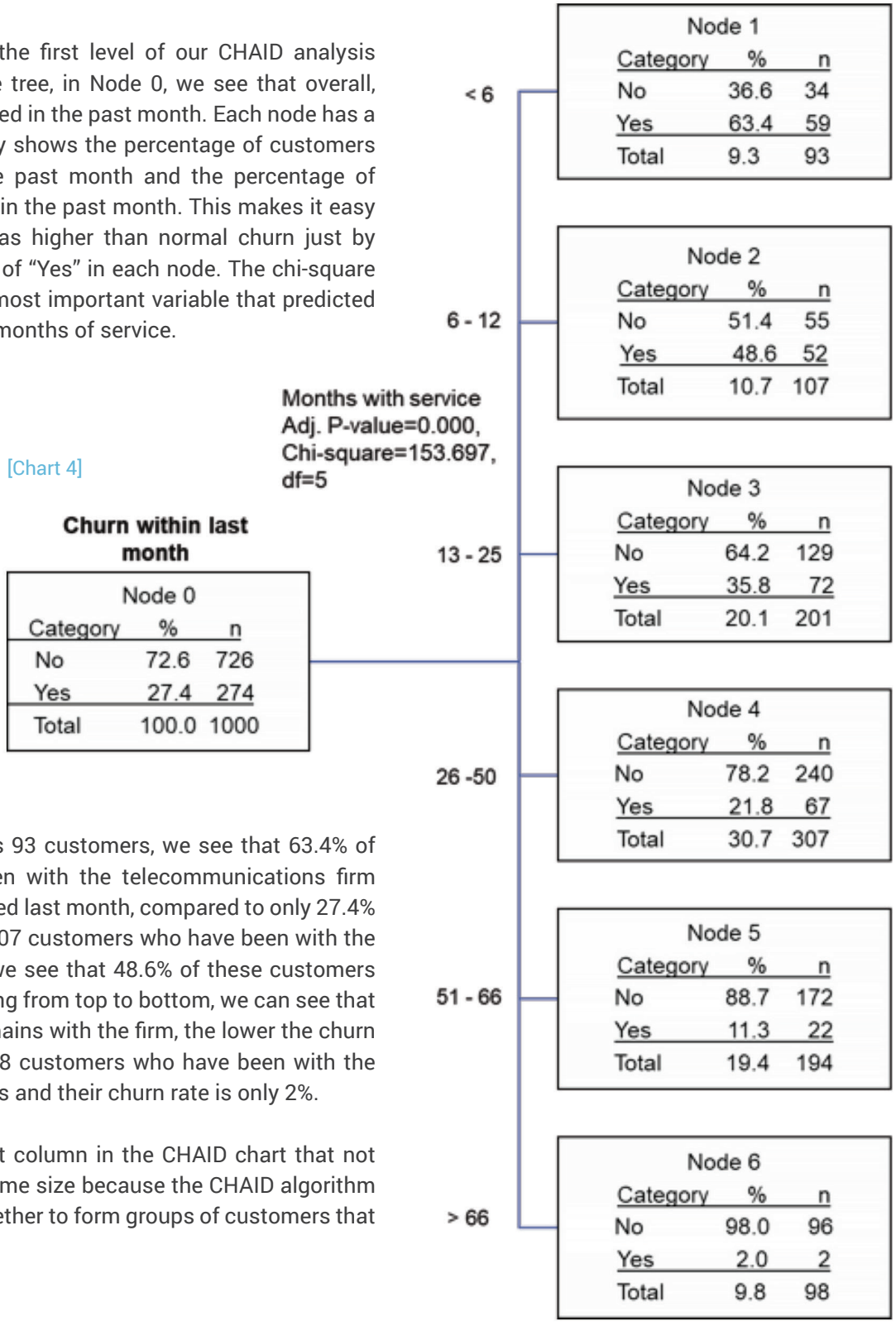
One of the easiest things about CHAID decision tree analysis is that to start the analysis we can simply select all 41 unique data fields as independent variables and specify the churn variable as the dependent variable.

Because chi-square can't be run on continuous variables, such as months of service or age, the CHAID algorithm starts the process by converting any continuous variables into deciles. CHAID will then evaluate the data and determine whether to combine some of the deciles together to better fit the data. This process is driven entirely by the data without analyst intervention.

For ease of visualizing the decision tree diagrams, we have oriented our decision tree in a left to right chart rather than top to bottom to better fit on the page.

In **Chart 4** here, we see the first level of our CHAID analysis graphic. At the left of the tree, in Node 0, we see that overall, 27.4% of customers churned in the past month. Each node has a simple table which quickly shows the percentage of customers who did not churn in the past month and the percentage of customers who did churn in the past month. This makes it easy to identify a node that has higher than normal churn just by looking at the percentage of "Yes" in each node. The chi-square algorithm found that the most important variable that predicted churn was the number of months of service.

[Chart 4]



In **Node 1**, which contains 93 customers, we see that 63.4% of customers who have been with the telecommunications firm less than 6 months, churned last month, compared to only 27.4% overall. Node 2 includes 107 customers who have been with the firm 6 – 12 months and we see that 48.6% of these customers churned last month. Moving from top to bottom, we can see that the longer a customer remains with the firm, the lower the churn rate. Node 6 represents 98 customers who have been with the firm longer than 66 months and their churn rate is only 2%.

We can see from this first column in the CHAID chart that not all the nodes are of the same size because the CHAID algorithm has combined deciles together to form groups of customers that behave similarly.

The CHAID algorithm will continue to go through all the independent variables to see if other variables help to explain fluctuations in the churn rate.

In **Chart 5** we see that the next variable that the CHAID decision tree algorithm found helpful was whether or not a customer rented equipment.

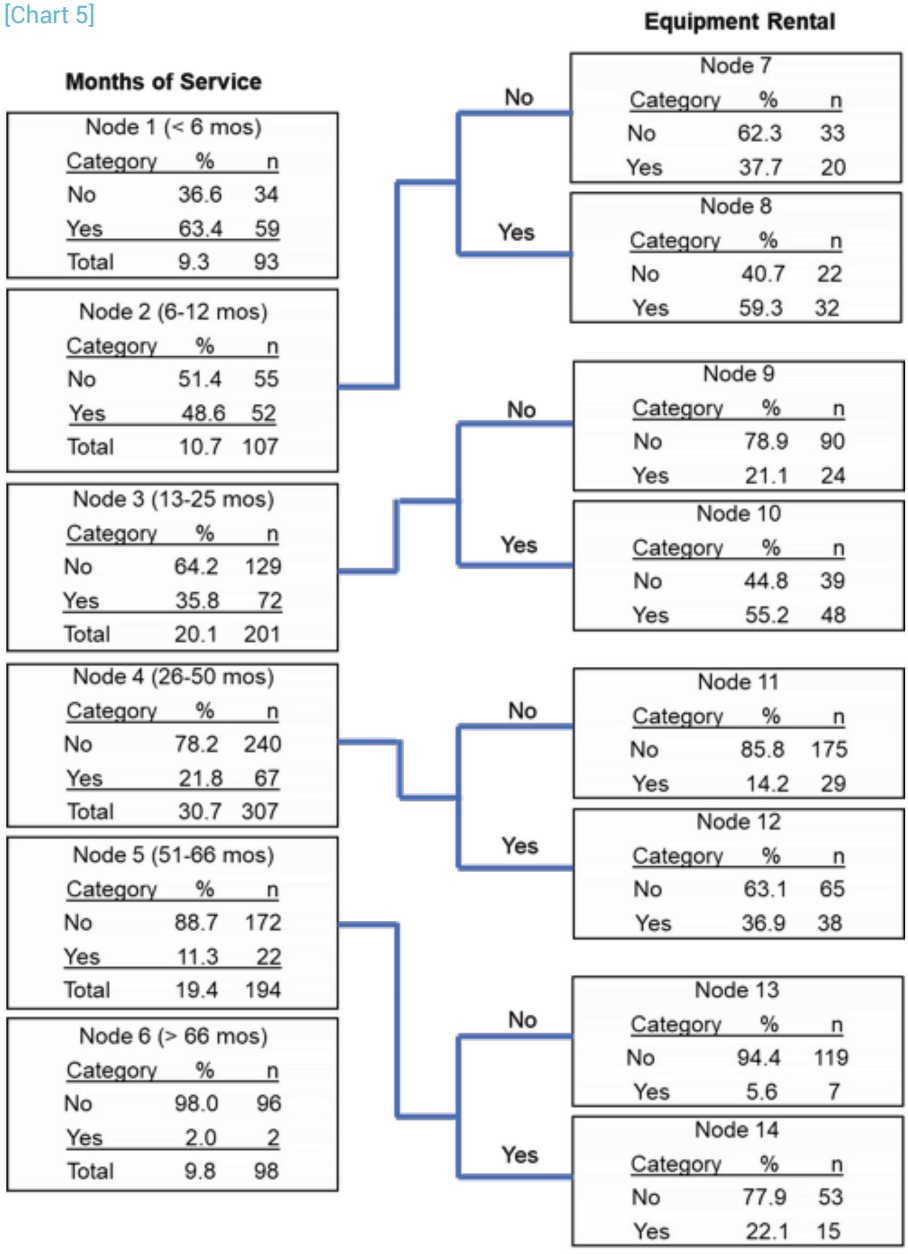
We see that in **Node 2**, the churn rate is 48.6%. But if the customers in this group rented equipment from the telecommunications firm, then the churn rate goes up to 59.3% (Node 8).

In **Node 3** the churn rate is 35.8%, but if customers in this group rented equipment, then the churn rate goes up to 55.2% (Node 10).

In **Node 4** the churn rate is 21.8%, but if customers in this group rented equipment, then the churn rate goes up to 36.9% (Node 12).

In **Node 5** the churn rate is only 11.3%, but if customers in this group rented equipment, then the churn rate goes up to 22.1% (Node 14).

[Chart 5]



Our decision tree chart quickly shows us that newer customers are the most likely to switch telecommunications providers. It also shows us that even customers that have been with the firm 6 months or longer are more likely to churn if they are renting equipment.

A telecommunications firm with this type of information can easily develop client retention programs targeting their newest customers as well as customers who rent their equipment.

An additional result from the CHAID analysis can be seen below in Table 3. This is the classification table which looks at the data model and evaluates the accuracy of the new model. In our scenario, we see that the results as displayed in Chart 5 will accurately predict which customers will churn 77% of the time.

CHAID has some simple stopping rules. The first stopping rule is that a tree branch will not split unless it has at least 100 records in it. As we can see in Chart 5 above, Nodes 1 and 6 were not split further because they had fewer than 100 records.

[Table 3 – Classification]

	Predicted		
Observed	No	Yes	Percent Correct
No	631	95	86.9%
Yes	135	139	50.7%
Overall Percentage	76.6%	23.4%	77.0%

The second stopping rule is tree depth. This is the number of levels the tree is allowed to continue until it simply halts at that maximum. The default in most software programs is five. CHAID typically wants to grow wide trees. But for most medical, financial, or marketing scenarios, going beyond five levels will rarely reveal additional meaningful insights.

The third stopping rule is confidence level, or alpha. This is the level of statistical significance that we desire for the purpose of our research. If we set the confidence level at .05, then our tree should grow larger. If we set the confidence level at .01, then the tree will grow less.

Conclusion

Machine learning consists of building models based on mathematical algorithms in order to better understand the data. One of the most important steps in understanding the data problem is to decide how the data needs to be analyzed in order to yield the desired results. In our grocery store example, we used cluster analysis to group the shoppers into clusters that would help us to understand the buying preferences of these shoppers. In our telecommunications company example, we used decision tree analysis to develop a prediction model to determine which customer characteristics will help us to predict the customers that are most likely to defect (or churn) and go to one of our competitors.

While today's machine learning algorithms make it easier to analyze large arrays of data without much programming from the analyst, it still requires that the analyst knows which type of algorithm to use for the problem and how to interpret the results.

Data Sources:

1. http://archive.ics.uci.edu/ml/datasets/Wholesale+customers
2. Telco.sav is an SPSS file that is supplied with the latest versions of the SPSS software.