DECISION TREE ANALYSIS
GEORGIA STATE UNIVERSITY

AnalyticsIQ, Inc.

Gregg Weldon

1. Explanation of CART Decision Trees
2. Decision Tree Uses and Limitations
3. Project Examples:
    a. Crosstabs
    b. Segmentation
    c. Continuous Dependent Variables
    d. Combination Variables
    e. Modeling

CART Decision Trees – Decision trees using CART (Classification and Regression Trees) create a maximum of 2 branches at a time, but allow the same variable to be used at multiple levels.

Tree splits are determined by "deviance", a mathematical equation that identifies the maximum difference in the Dependent variable of all potential branches. The "best" split (biggest Dependent variable difference) is chosen, and the resulting Nodes are then tested for the next split.

Branching will continue until A). there are no further branching possibilities, or B). there are no possible branches that meet the minimum observations-per-branch rule.

The user determines the minimum number of observations for each branch. This is usually a percentage of the total sample (5%-10% is the rule of thumb). If there are fewer observations than that, the tree will no longer branch, and the existing branch will serve as the "End Node".

There are a wide range of reasons for utilizing decision trees. These reasons include the following:

1. Crosstabs
2. Segmentation
3. Predicting a continuous dependent variable ($)
4. Combination variables
5. Modeling

EXAMPLE 1 – Crosstabs

This client was a sub-prime mortgage loan provider whose primary concern was finding new customers for mortgage re-financing.

HST39X is a variable that looks at the number of trades on someone's credit bureau report that have ever been 60 days past due or worse.

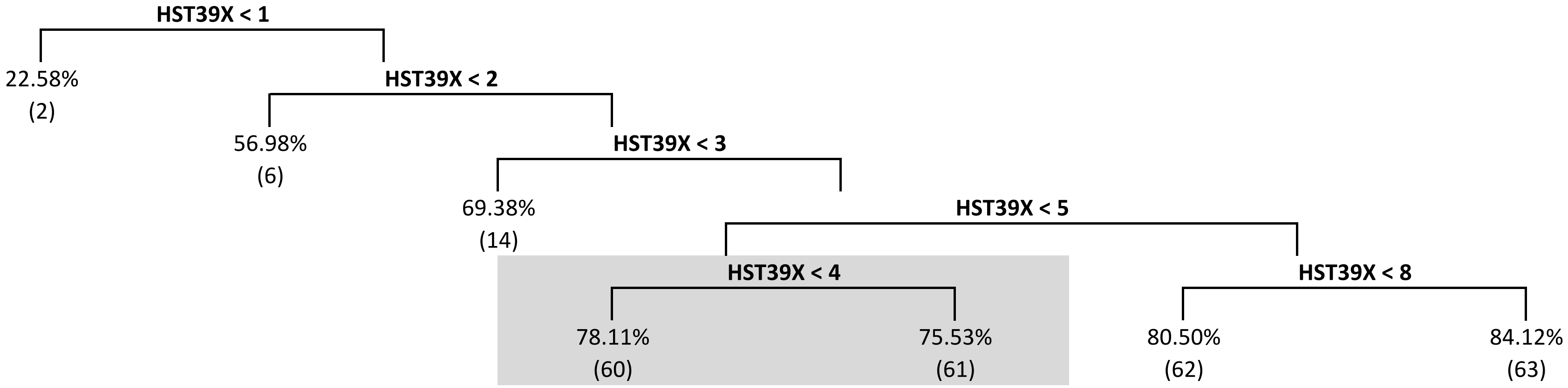
We had a modeling sample of 20,145 observations, with a 50/50 split of Goods (people who responded to an offer) and Bads (people who did not respond). The higher the ratio, the higher % of people responded to an offer.

By using a decision tree, we were able to see a very clear trend of higher delinquencies indicating a higher likelihood of response. For example, people with no history of delinquencies (HST39X < 0.5) only had a ratio of 22.58% (less than half of the sample's 50% average). People who had HST39X = 1 (HST39X < 1.5) had a ratio of 56.98%. Those people who had HST39X > 7.5 had a ratio of 84.12%.

Using CART can work for running crosstabs, but it's probably NOT the best (or easiest) way!

Example 1- Crosstabs

20,145 Observations in Sample
"Good" = a high % **Response** in the End Nodes
50/50 Sample (50% Goods, 50% Bads)



The shaded Nodes "don't work", in that the Good rate is higher at Node 60 than at Node 61 (backwards from the trend we see in the variable overall.
We would "trim" this Decision Tree limb and instead go up one level (HST39X < 5). That End Node is # 30, and the Good Rate there is 76.93%.

Final End Nodes for this variable are as follows:

Node 2	HST39X < 1 (i.e. HST39X = 0)	22.58% Good (response)Rate
Node 6	HST39X < 2 (i.e. HST39X = 1)	56.98% Good (response)Rate
Node 14	HST39X < 3 (i.e. HST39X = 2)	69.38% Good (response)Rate
Node 30	HST39X < 5 (i.e. HST39X = 3,4)	76.93% Good (response)Rate
Node 62	HST39X < 8 (i.e. HST39X = 5-7)	80.50% Good (response)Rate
Node 63	HST39X >= 8	84.12% Good (response)Rate

EXAMPLE 2 – Segmentation

Segmentation is how we use CART Decision Trees about 90% of the time. When a client needs help with modeling, it's not always clear if a single model with suffice or if multiple models are required. A small credit union may offer both consumer loans (cars, boats, vacations, etc.) and a VISA credit card. Would a single risk model work for both of these lines of business, or would separate models be better? Segmentation allows us to analyze a clients' data prior to the modeling process in order to determine this.

Decision trees can help if we need to determine the best split for potential models.

In this example, we are again focusing in on a mortgage loan provider. This client was convinced that segmenting sub-prime customers could yield better response results from their mailings. We looked at those existing loan customers whose FICO scores ranged from 470-539 (sub-prime, but not the absolute WORST). The client wanted to know if there were other aspects about these customers that would indicate response. We looked at a handful of mortgage-related variables for our answer.

MTHICR = high credit on mortgage

MTNEWT = age in months of newest mortgage trade

RESYRS = number of years at current residence

MTBALN = amount of mortgage balance

MTTRDS = number of mortgage trades (ever)

The tree on the next page shows that MTHICR was the best splitter, with those customers with high credit <= $37,825 having a much lower likelihood of responding than average (average = 80.24% here).

Among customers with "high" MTHICR, those with fairly recent MTNEWT (<= 54.5 months) are slightly more likely to respond than otherwise. RESYRS gave a split that was pretty insignificant.

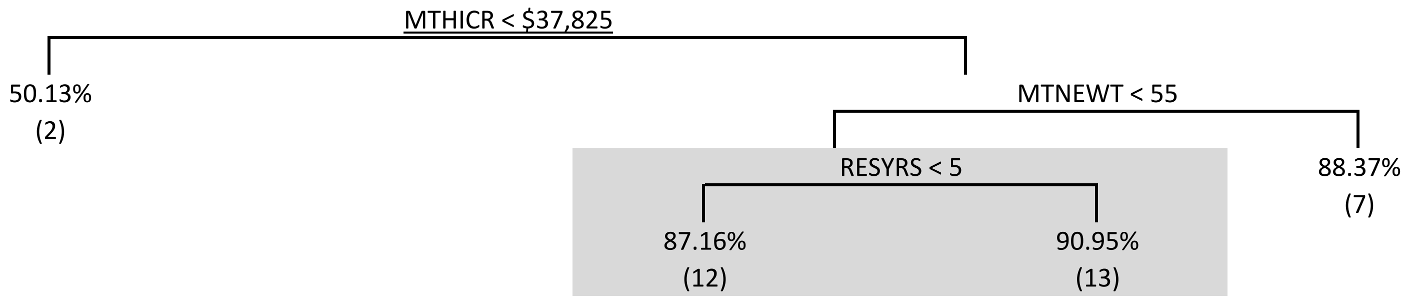
Mathematically, the "best" split would be at MTHICR <= $37,825 vs. MTHICR > $37,825. In the end, the client asked us to build models where MTHICR <= $75,000 vs. MTHICR > $75,000. Why the difference? $75,000 was a policy limit the company had traditionally used for what they considered "big" loans vs. "small" loans. So, for business reasons, we made this change.

Segmentation is a combination of art and science. CART will give us the science of finding the statistically best split, but the client's business needs (the art) will trump this every time. Incidentally, the split at $75,000 also made sense from a statistical standpoint. It just wasn't the optimal mathematical solution.

Example 2 - Segmentation

3790 Observations in Sample
"Good" = high % **Response** in End Nodes
Mean Good = 80.24%

FICO Scores from 470 to 539



Nodes 12 and 13 were dropped because their Good rates were so similar. Their "parent" Node - Node 6- was also too similar to Node 7, so these were dropped as well. This left Node 2 (50.13%) vs. Node 3 (87.73%)

EXAMPLE 3 – Predicting a Continuous Dependent Variable

Let’s assume that we have an insurance client who’s trying to predict the amount of claim dollars they’ll have to pay out over the next 12 months for various customers. People with a high likelihood of having claims would obviously pay higher premiums in order to make up for their risk. A model predicting Good (no claims) vs. Bad (claim) would work. However, the prediction of *the amount of the claim* can be equally important.

Since most insurance customers don’t have a claim within a given year, it’s very hard to build a model that can predict a claim amount. This is because 99% of the population has a claim of $0. The remaining 1% of people have claims ranging from $1000 to $100,000. Linear and logistic regression models will not work in these cases. Tobit modeling is designed for this kind of problem, but CART can work here as well.

A CART decision tree is able to segment among these insurance customers, with some End Nodes averaging $0 (or nearly $0) and other Nodes consisting of those customers who will have claims.

Please note, however, that much care must be taken with these kinds of trees, as the results need to be checked for logic and consistency. Generally, a higher minimum number of observations for branching is called for, as conservative analysis is needed to not “over-fit” the data. Any kind of analysis involving a high % of observations with one value and a handful of observations with a series of other values is prone to instability and error.

EXAMPLE 4 – Creating Combination Variables

Linear and logistic modeling use multivariate analysis to predict an outcome. A model, for example, may include variables such as Length of Time at Residence (LOR), Age, Income, Rent vs. Own, Dwelling Type, etc. However, in some cases, it may make sense to take several of these variables and attempt to create a “combination” variable- a variable that includes all of these factors in one.

Why would you do this, if modeling, by definition, creates a score that acts as a combination of all variables in the model?

We had a client in the fundraising industry (a major charity) that was trying to find “lapsed donors” (former contributors to the charity) who would be most likely to begin contributing again, if mailed. Our variables included all the demographic variables for these donors but, more importantly, all of their past history with the charity (# of past donations, amount of donations, ratio of donations to # or promotions, time since last donation, donations in 39 months/48 months/60 months/96 months, etc.). This past history information was incredibly important. In fact, it was so important that it drowned out the demographic variables that also told us part of the story as to who would begin contributing again. On their own, variables like Age, LOR, and Income weren’t significant enough to last in the model. However, a combination variable, tying the "best" of each of the variables together, was very important.

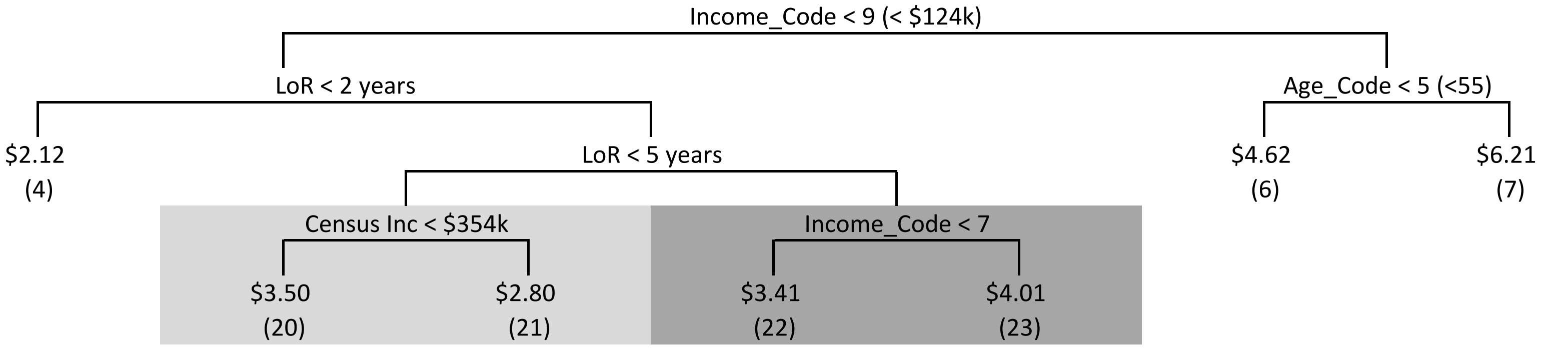
On the next page, you can see that we had a sample of 27,055 observations, and contributions averaged $3.62. 11.5% of this sample were Goods (began contributing again). The average of $3.62 is low because most people didn't contribute again, giving a value of $0.

The variables that came into this combination variable were Income, Age, and LOR. I ended up with 5 Nodes, ranging from $2.12 to $6.21 in terms of average contribution.

This new variable with 5 values (called TGroupA) was very significant in the final model. In fact, this one "combination" variable was one of the most significant variables in the entire model. The reason? We took the best information from several related, semi-significant variables and combined them into one *really* significant variable.

Example 4 - Combination Variables

27,055 Observations
$3.62 average amount of donation
11.50% "Good" rate (% of returned donors)



The first shaded box (Census Inc) was trimmed because the dollar amount returned for the 2 Nodes (20 and 21) are backwards from the overall tend. The Node above those (Node 10 - LoR < 5 years) had a dollar return of $3.04.

The second shaded box (Income_Code) works fine, but was dropped purely for ease of implementation. 5 End Nodes worked for this variable better than 6, so these nodes were trimmed up to Node 11 - LoR >= 5 years (dollar return of $3.65).

Final Results:

Node 4	$2.12	Good rate = 6.3%
Node 10	$3.04	Goods rate = 11.1%
Node 11	$3.65	Good rate = 14.3%
Node 6	$4.62	Good rate = 12.0%
Node 7	$6.21	Good rate = 13.0%

Note that the Good Rates for the final Nodes don't rank in the same order as the $ amount returned. This is because Nodes 6 and 7 had lower rates of returning donors, but they donated larger amounts of money. Our goal was to maximize dollars, not Good rates in this project.

<u>EXAMPLE 5</u> – Modeling

In the past, many people have tried to replace regression analysis (multivariate analysis) with decision trees (bivariate analysis). This doesn't work well, as trees leave "holes" that observations can slip through. Usually, trees need to be supplemented with a large number of "rules" or upfront criteria to try and catch these people before they slip through. It's not very efficient or effective.

There are occasions, however, where the use of a tree can actually work well.

Net Worth is the difference between a person's assets and liabilities. Predicting an individual's net worth is very important for marketers, as they can choose which consumers should be sent mailings for various products. The wealthy generally don't respond to sub-prime offers, and the poor don't respond (or don't qualify anyway) for luxury goods. Identifying net worth (or wealth) is a valuable ability. Net worth, for most people, is closely aligned with home value, mortgage amounts, income, spending habits, etc.

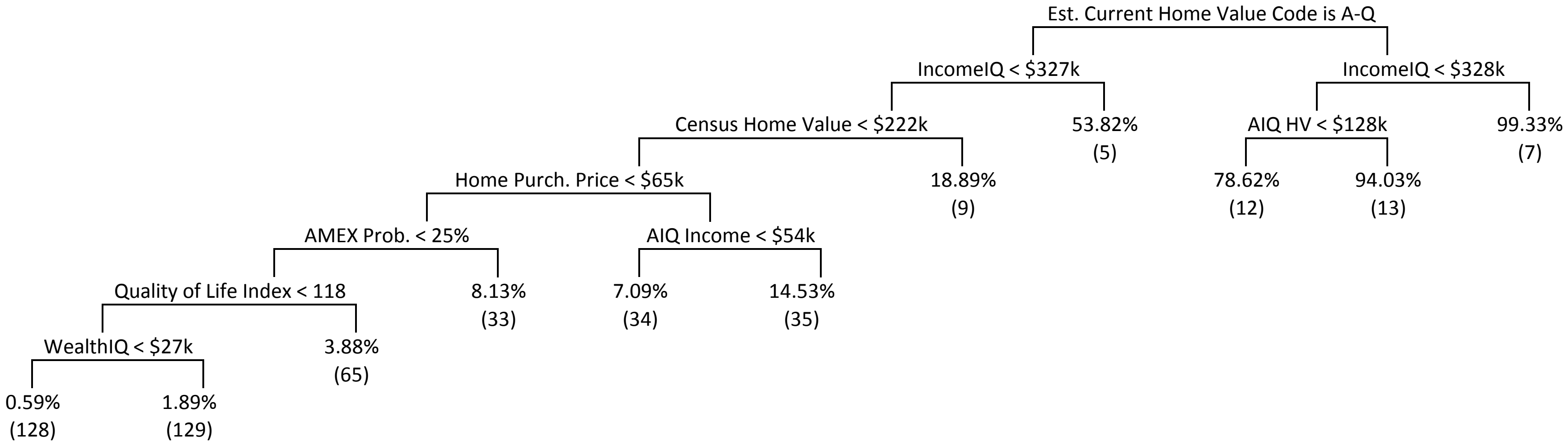
Our company has several modeled scores that attempt to predict a person's net worth. In addition, we have several variables from different data sources that look at home value, mortgage balance, original home purchase price, spending, credit card use, etc. How can we combine all of these variables together to get one really accurate variable that measures a person's wealth?

The problem is that all of these variables are closely related to each other. They're highly correlated because they're telling us the same story—who's rich, who's middle class, who's poor. Attempting to build a traditional model using these variables means having to throw most of them out due to multi-collinearity. What we may be left with is a model with 2-3 variables that all say basically the same thing.

A decision tree allowed us to look at all of these variables without having to worry about correlation. In the end, each consumer fell into 11 Nodes. We had 14,289 observations and the mean Good rate (Good = high wealth people) was 39.76%. Our Nodes ranged from 0.5% to 99.3% (a huge range). Because we used trees and bivariate analysis, the correlation between variables was not a factor. This tree was only the first phase of this particular project. After much more analysis and work on these 11 Nodes, we ended up with a product that predicts the actual net worth of each consumer in America. These values range from $0 to $20 million+.

Example 5 - Modeling

14,289 Observations
39.76% "Good" Rate


Est. Current Home Value Code is A-Q
IncomeIQ < $327k
IncomeIQ < $328k
Census Home Value < $222k
53.82%
(5)
AIQ HV < $128k
99.33%
(7)
Home Purch. Price < $65k
18.89%
(9)
78.62%
(12)
94.03%
(13)
AMEX Prob. < 25%
AIQ Income < $54k
Quality of Life Index < 118
8.13%
(33)
7.09%
(34)
14.53%
(35)
WealthIQ < $27k
3.88%
(65)
0.59%
(128)
1.89%
(129)