Paper 98-2019

Multi-Way (more than two) Splits In Decision Trees

Russ Lavery, YuTing Tian

ABSTRACT

Decision trees are no longer new tools of Data Scientists and are frequently used to split people into binary groups (two way splits). However SAS Enterprise Miner has the ability to create decision trees that split into more than two levels. This can be very useful if an analyst is trying to assign observations into more than two groups. This paper uses examples to explore this powerful feature (multi-way splits) of SAS Enterprise Miner to predict both n-way categorical variables and continuous variables

AN EXCITING 3-WAY SPLIT

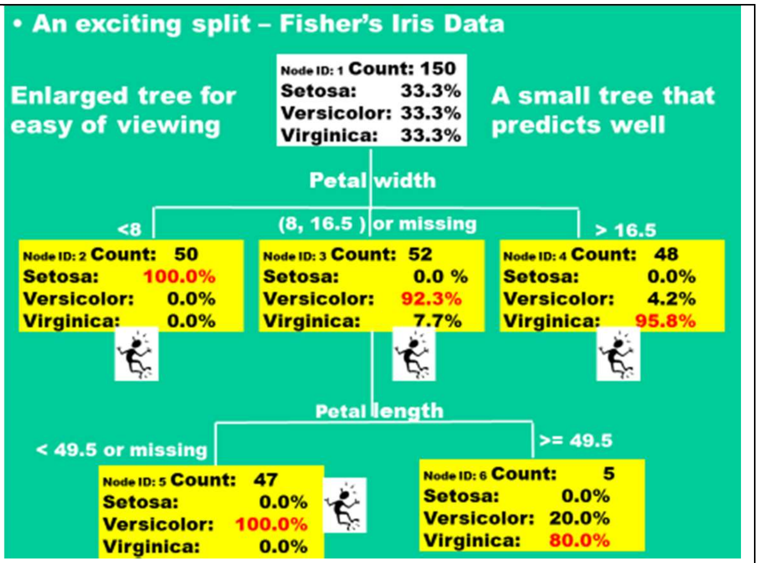
Many people are familiar with Fisher's Iris data. It gained a lot of attention in Fisher's 1936 paper on discriminant analysis and has been a favorite statistics and machine learning example. A colleague collected150 observations on three different species of Iris. Many machine learning techniques have been applied to this data in an effort to recover the three groups/species.

Figure 1 shows an exciting split. The Decision Tree tab in SAS Enterprise Miner was asked to split the data in three groups.

The first level of split does a very good job of making leaves of high level of purity.

The surprisingly accurate and parsimonious results coming from a single level of split motivated this paper.

This paper is an attempt to discover if the excellent recovery in Figure 1 is common and to discover any rules of thumb for applying N-way splits in decision trees.



- An exciting split – Fisher's Iris Data

Enlarged tree for easy of viewing

A small tree that predicts well

Node ID: 1 Count: 150
Setosa: 33.3%
Versicolor: 33.3%
Virginica: 33.3%

Petal width

<8

(8, 16.5) or missing

> 16.5

Node ID: 2 Count: 50
Setosa: 100.0%
Versicolor: 0.0%
Virginica: 0.0%

Node ID: 3 Count: 52
Setosa: 0.0 %
Versicolor: 92.3%
Virginica: 7.7%

Node ID: 4 Count: 48
Setosa: 0.0%
Versicolor: 4.2%
Virginica: 95.8%

Petal length

< 49.5 or missing

>= 49.5

Node ID: 5 Count: 47
Setosa: 0.0%
Versicolor: 100.0%
Virginica: 0.0%

Node ID: 6 Count: 5
Setosa: 0.0%
Versicolor: 20.0%
Virginica: 80.0%

Figure 1

The paper will have the following structure:

Some observations on the way SAS has implemented decision trees

Splitting three categories using a **three-way** split: Fisher's Iris data
Splitting three categories using a **two-way** split: Fisher's Iris data
Splitting three categories using a **four-way** split: Fisher's Iris data

Splitting six categories using a **six-way** split: SASHelp.cars
Splitting six categories using a **four-way** split: SASHelp.cars
Splitting six categories using a **two-way** split: SASHelp.cars

Splitting a continuous Y with multi-way splits

A bit of the geometry of multi-way splitting

SOME THOUGHTS ON THE WAY SAS HAS IMPLEMENTED DECISION TREES

The authors were astounded with the complexity of the way that SAS has implemented decision trees. SAS has implemented many adaptive algorithms to allow an analyst to quickly get a result, an accurate result, without having to handle many of the details of the process.

As an example of the helpful automation provided by SAS, if the number of levels of an X variable is small—ish (maybe on the order of 5,000 levels) SAS will examine every level as a splitting point. I am not certain of the number of splitting points handled by SAS, but JMP handles 5,000 levels. If the number of levels is greater than the threshold, SAS will automatically switch to sampling the levels of the X and use that sample to create splitting points

Another example is that, when appropriate, SAS will automatically apply a Bonferroni adjustment when comparing X variables with differing numbers of levels.

The level of automation provided by SAS is truly impressive and impossible to summarize in the limitations of a SUG paper.

SPLITTING A Y VARIABLE THAT HAS THREE LEVELS

EXAMPLE 1: APPLYING A THREE WAY SPLIT TO A Y VARIABLE WITH THREE LEVELS

The results of this type of action can be seen in Figure 1. It is applying a three way split to the Fisher's Iris data. This type of analysis is more complicated, and is expected to have a longer run time, than a decision tree using two way splits. However; this creates a tree that might be easier to explain to a subject matter expert than a tree using two-way splits.

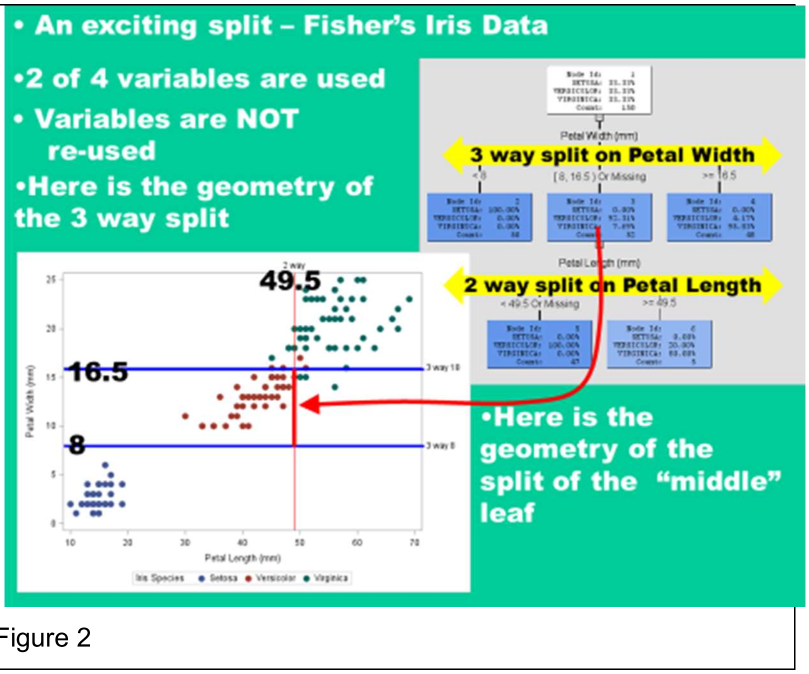
It might be that, for an analyst with access to enough computing power, it is worth running a multi-way split. It could be that ease of explanation to a client compensates for any additional run time.

Since this example only used two variables it is possible to show the geometry of the splits and this is done in Figure 2.

Notice that a decision tree uses lines that are perpendicular to a variable to create decision rules for classifying observations.

This means that data sets that can be separated by lines perpendicular to an axis are good subjects for decision trees and result in accurate and small trees.

If the line that separates two groups must be at an angle to an axis, then that angle is made by several splits making a step function and the tree would, likely, not be parsimonious.



- An exciting split – Fisher's Iris Data

•2 of 4 variables are used

- Variables are NOT re-used

•Here is the geometry of the 3 way split

•Here is the geometry of the split of the "middle" leaf

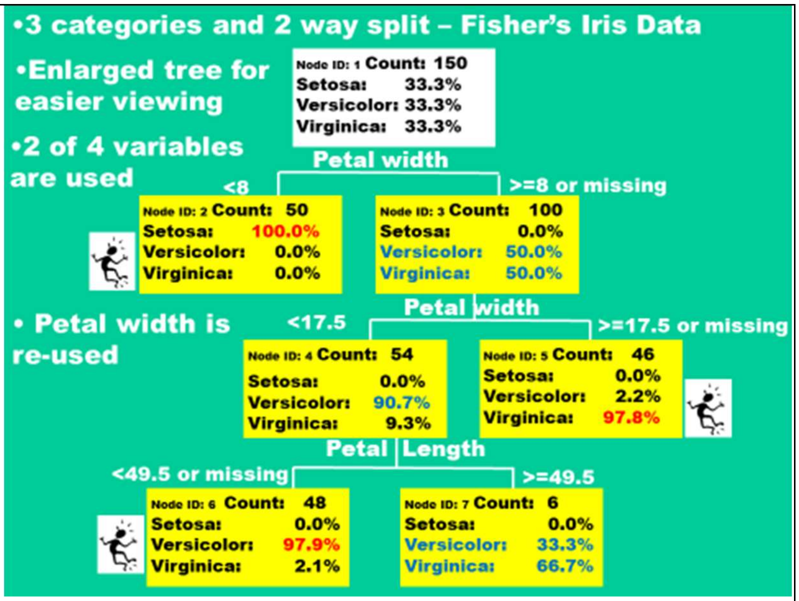
Figure 2

EXAMPLE 2: APPLYING A TWO-WAY SPLIT TO A Y VARIABLE WITH THREE LEVELS

Most books that discuss N-way splits do mention that N-way splits can be effectively mimicked by trees with two-way splits.

Figure 3 shows Fisher's Iris data set being analyzed by a series of two-way splits.

The accuracy and splitting points are very similar to a three-way split, but the pattern is a little less obvious and might be a little harder to explain.



•3 categories and 2 way split – Fisher's Iris Data

•Enlarged tree for easier viewing

•2 of 4 variables are used

• Petal width is re-used

Node ID: 1 Count: 150
Setosa: 33.3%
Versicolor: 33.3%
Virginica: 33.3%

Petal width

<8

>=8 or missing

Node ID: 2 Count: 50
Setosa: 100.0%
Versicolor: 0.0%
Virginica: 0.0%

Node ID: 3 Count: 100
Setosa: 0.0%
Versicolor: 50.0%
Virginica: 50.0%

Petal width

<17.5

>=17.5 or missing

Node ID: 4 Count: 54
Setosa: 0.0%
Versicolor: 90.7%
Virginica: 9.3%

Node ID: 5 Count: 46
Setosa: 0.0%
Versicolor: 2.2%
Virginica: 97.8%

Petal Length

<49.5 or missing

>=49.5

Node ID: 6 Count: 48
Setosa: 0.0%
Versicolor: 97.9%
Virginica: 2.1%

Node ID: 7 Count: 6
Setosa: 0.0%
Versicolor: 33.3%
Virginica: 66.7%

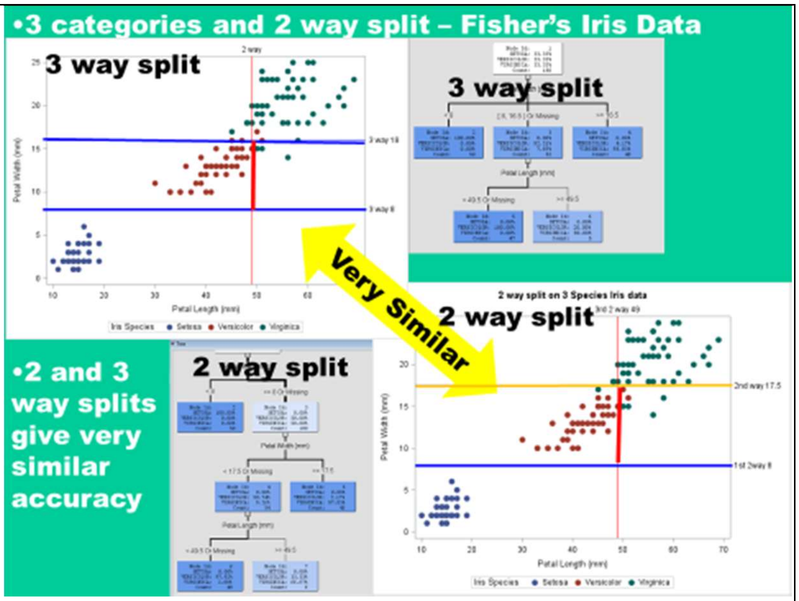
Figure 3

This paper does not judge the accuracy of decision trees using holdout samples, and the reason for that will be explained at the end, but we will try to give some understanding why accuracies might be similar.

Figure 4 shows the geometry of splits resulting from the three-way splits and two-way splits.

It is easy to see that the geometries created by the two different algorithms are very similar.

This is not offered as a mathematical proof but as an example to support comments commonly made in decision tree books.



•3 categories and 2 way split – Fisher's Iris Data

•2 and 3 way splits give very similar accuracy

Figure 4

Books often say that N-way splits can be approximated by several 2-way splits. We see this above. The issue to consider is if the tree showing a 3-way split is easier to explain to a client.

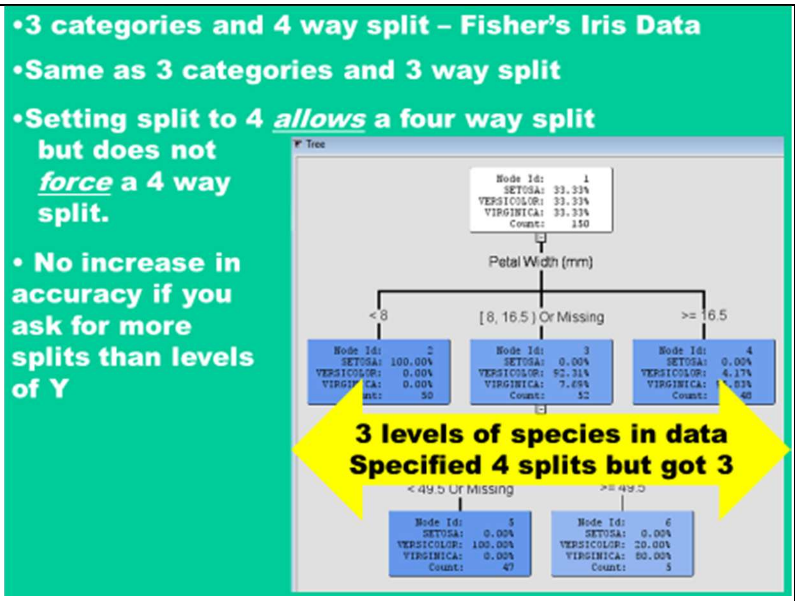
EXAMPLE 3: APPLYING A FOUR-WAY SPLIT TO A Y VARIABLE WITH THREE LEVELS

Figure 5 shows a tree that is the result of requesting four-way splits for a Y variable that only has three levels.

If the Y is categorical, and an analyst requests more splits than categories in the Y, the decision tree algorithm will not apply more splits than levels.

The authors suggest a rule-of-thumb. Setting splits to N ***allows*** an N-way split but cannot ***force*** an N-way split.

This suggests that, if this N-way technique is to be applied, then an analyst should previously run a PROC Freq to determine the number of levels of the Y variable.



- **3 categories and 4 way split – Fisher's Iris Data**
- **Same as 3 categories and 3 way split**
- **Setting split to 4 *allows* a four way split but does not *force* a 4 way split.**
- **No increase in accuracy if you ask for more splits than levels of Y**

Tree
Node Id: 1 SETOSA: 33.33% VERSICOLOR: 33.33% VIRGINICA: 33.33% Count: 150
Petal Width (mm)
< 8
[8, 16.5) Or Missing
>= 16.5
Node Id: 2 SETOSA: 100.00% VERSICOLOR: 0.00% VIRGINICA: 0.00% Count: 50
Node Id: 3 SETOSA: 0.00% VERSICOLOR: 92.31% VIRGINICA: 7.69% Count: 52
Node Id: 4 SETOSA: 0.00% VERSICOLOR: 4.17% VIRGINICA: Count: 48
3 levels of species in data
Specified 4 splits but got 3
< 49.5 Or Missing
>= 49.5
Node Id: 5 SETOSA: 0.00% VERSICOLOR: 100.00% VIRGINICA: 0.00% Count: 47
Node Id: 6 SETOSA: 0.00% VERSICOLOR: 20.00% VIRGINICA: 80.00% Count: 5

Figure 5

SPLITTING A Y VARIABLE THAT HAS SIX LEVELS

Admittedly, this paper is light on theory, and on math derivations, and simply attempts to use a few examples to create heuristics that an analyst might use in trying to decide if an N–way split is appropriate for their particular case.

With this in mind, the paper will use a convenient data set, SASHelp.cars, to see if patterns previously seen will recur.

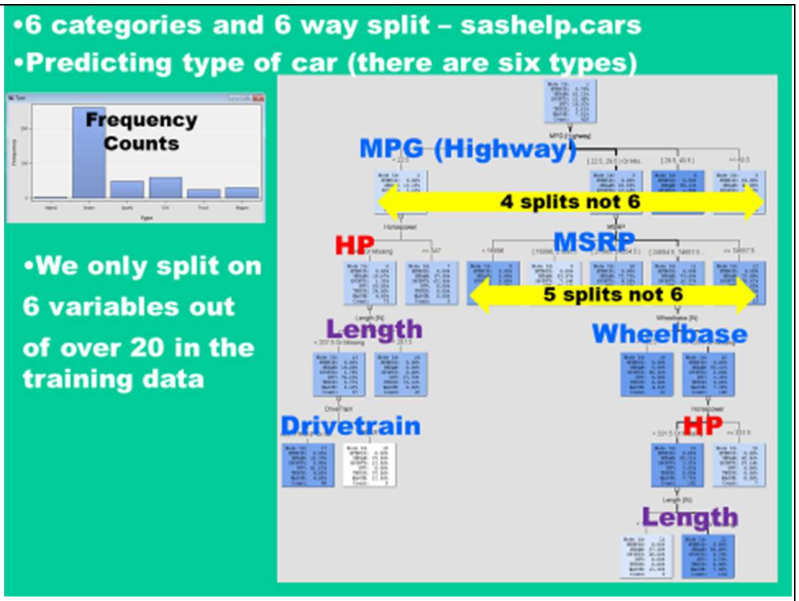
EXAMPLE 4: APPLYING A SIX-WAY SPLIT TO A Y VARIABLE WITH SIX LEVELS

Figure 6 shows the results of trying to predict the six types of car in the data set SASHELP.cars.

Importantly, we see no six way splits in this tree. The first level is a four-way split. The second level contains a five-way split.

This confirms the observation, in Figure 5, that a decision tree interprets the request for an N-way split as ***permission to split N-ways*** but ***NOT an instruction or requirement to split*** N-ways.

Note that the decision tree reuses variables and the hope that an N-way tree would reduce the reuse of variables might be incorrect.



- **6 categories and 6 way split – sashelp.cars**
- **Predicting type of car (there are six types)**
- **We only split on 6 variables out of over 20 in the training data**

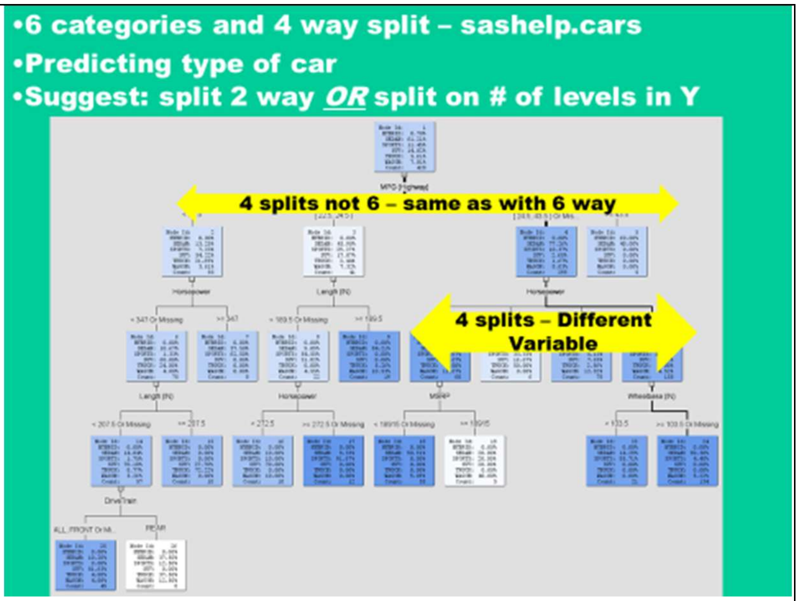
Figure 6

EXAMPLE 5: APPLYING A FOUR-WAY SPLIT TO A Y VARIABLE WITH SIX LEVELS

Figure 7 shows the results of applying a four-way split to a Y variable that has six levels.

As shown before, none of the levels contains a six way split. Requesting fewer splits than levels of the Y variable seems, on face value, to be a bad idea. It artificially limits the ability of a variable to, if the analyst is very lucky in the data geometry, find a split like the 3-way Iris split in Figure 1.

The authors suggest that any analysts, planning to use an N–way split, start the analysis with a PROC Freq to determine the number of levels of the Y variable so that the proper number of splits can be requested.



- 6 categories and 4 way split – sashelp.cars
- Predicting type of car
- Suggest: split 2 way ***OR*** split on # of levels in Y

4 splits not 6 – same as with 6 way

4 splits – Different Variable

Figure 7

It seems that requesting *more* splits than the number of levels in the Y variable is a waste of computing power. The decision tree will never create more splits than the number of levels in the Y variable.

It seems that requesting fewer splits than the number of levels in the Y variable is unlikely to create a parsimonious, and simple to explain, decision tree. However; most decision trees are created with two-way splits and the authors decided to compare trees from six-way splits and two-way splits where the Y variable contained six levels/groups.

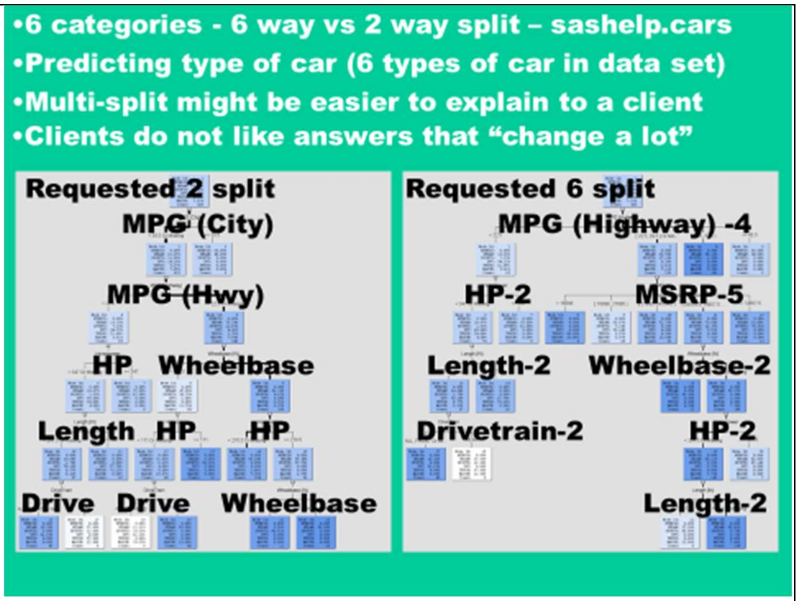
EXAMPLE 6: APPLYING A TWO-WAY SPLIT TO A Y VARIABLE WITH SIX LEVELS

Figure 8 shows the trees created by two-way splits versus six-way splits. The trees are quite different and this is to be expected – everyone knows decision trees are unstable.

The authors are tempted to prefer the six way split to the two–way split.

It would be very interesting to see if a subject matter expert could see types of cars “jumping out” of the decision tree with six splits.

It is suggested that having a tree that mimics, at least to some extent, the belief of subject matter experts make “selling” the tree to clients easier.



- 6 categories - 6 way vs 2 way split – sashelp.cars
- Predicting type of car (6 types of car in data set)
- Multi-split might be easier to explain to a client
- Clients do not like answers that “change a lot”

Figure 8

There is a bit of a conflict in the above idea. A decision tree that tells people things they already know is a tree that is: believable, easy to sell, likely true but not of much use. A decision tree that tells people true things they don't know is harder to sell but can be of great use. Possibly, an ideal tree would be one where part of the tree confirms what subject matter experts already believe and part tells them something new that can be applied.

SPLITTING A CONTINUOUS Y VARIABLE

Splitting a categorical variable had an easy "logical entry". It made sense to investigate splitting on the number of levels in the Y, and fewer levels than are in the Y and on more levels than are in the Y. Splitting on a continuous variable has less of an obvious line of attack.

EXAMPLE7: SPLITTING A CONTINUOUS Y VARIABLE

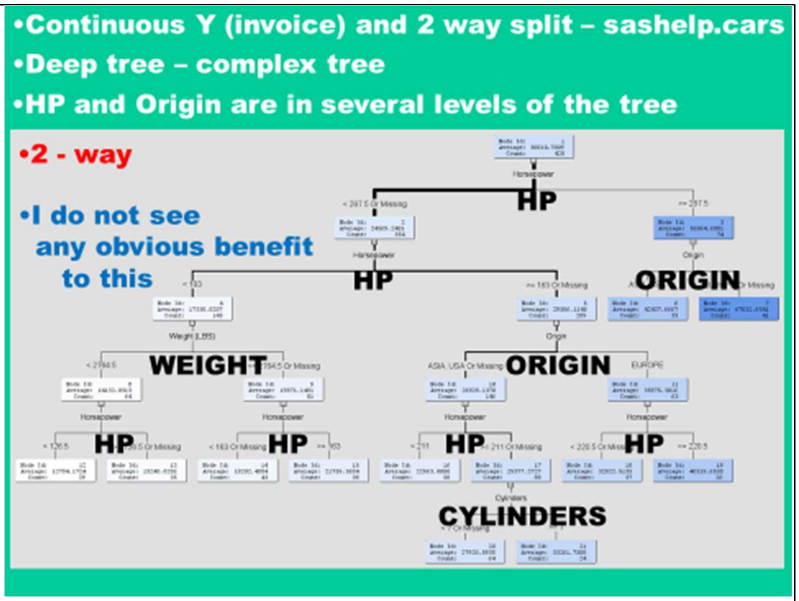
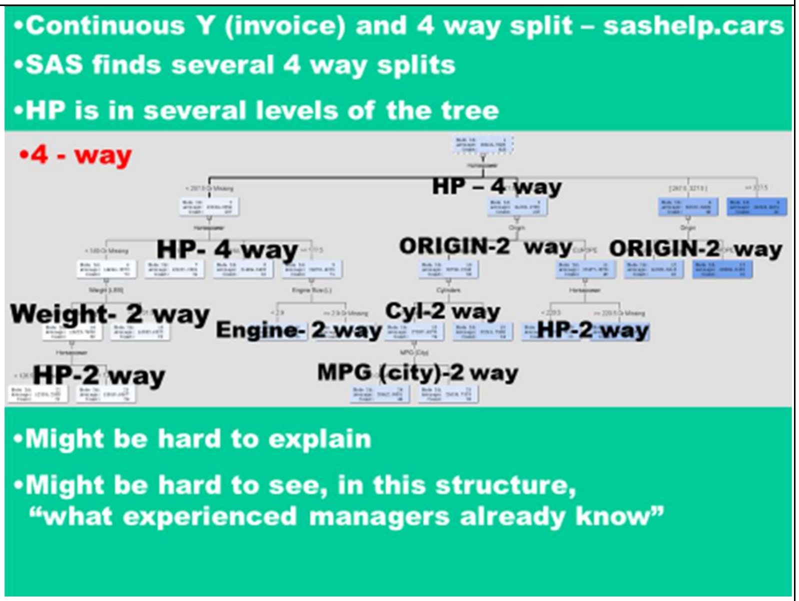
Figure 9 shows an example of how a decision tree might use two-way splits when a categorical X variable contributes to the values of Y. An attempt is made to predict invoice price in SASHelp.cars. Horsepower and origin end up in several levels of the tree. Origin and Cylinders are involved in two-way splits and those variables have more than two levels.	 •Continuous Y (invoice) and 2 way split – sashelp.cars •Deep tree – complex tree •HP and Origin are in several levels of the tree •2 - way •I do not see any obvious benefit to this HP HP ORIGIN WEIGHT ORIGIN HP HP HP HP CYLINDERS Figure 9
Figure 10 shows another example of how a decision tree might use N-way splits when a categorical X variable contributes to the values of Y. Increasing the number of splits changes the tree and allows horsepower to drive a four-way split at the top of the tree. The authors were tempted to take this tree to a subject matter expert and see if the decision tree showed the "truth of the market".	 •Continuous Y (invoice) and 4 way split – sashelp.cars •SAS finds several 4 way splits •HP is in several levels of the tree •4 - way HP – 4 way HP- 4 way ORIGIN-2 way ORIGIN-2 way Weight- 2 way Engine- 2 way Cyl-2 way HP-2 way HP-2 way MPG (city)-2 way •Might be hard to explain •Might be hard to see, in this structure, "what experienced managers already know" Figure 10

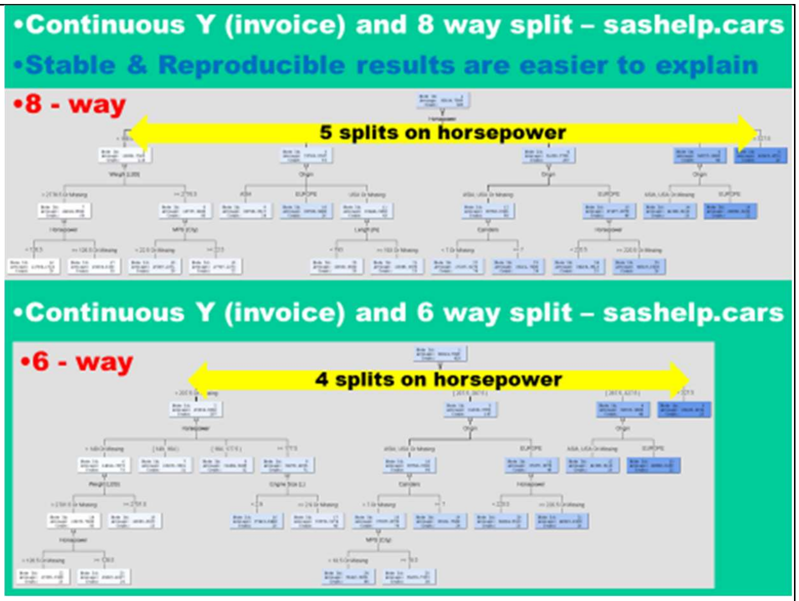
Figure 11 compares the decision trees for eight–way and six–way splits.

Both decision trees considered horsepower to be the most important predictor of invoice but differed on the number of splits.

The authors had expected that these two trees would find the same splitting points in the first level.

The disagreement between algorithms, shown in Figure 11, is difficult to explain to clients and makes them uncomfortable.

.


•Continuous Y (invoice) and 8 way split – sashelp.cars

•Stable & Reproducible results are easier to explain

•8 - way

•Continuous Y (invoice) and 6 way split – sashelp.cars

•6 - way

Figure 11

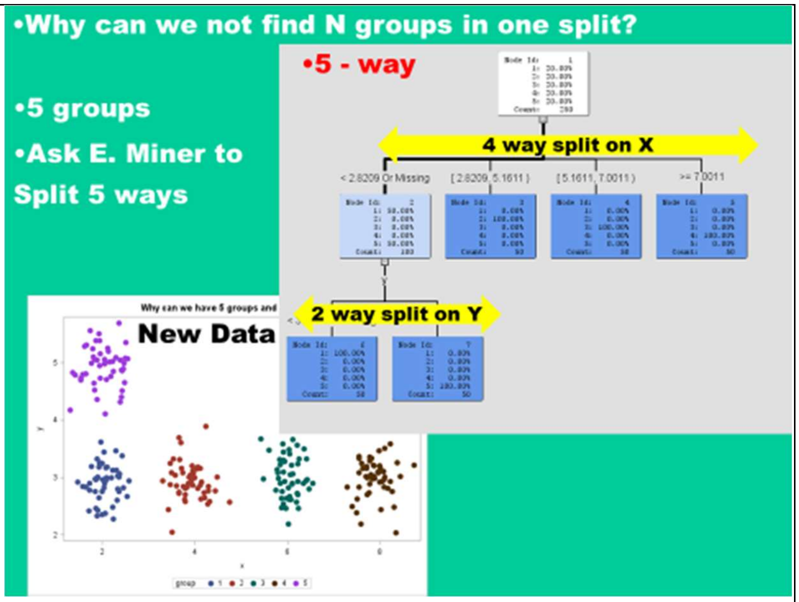
THE GEOMETRY OF MULTI-WAY SPLITS

Figure 12 introduces a new, and made up, dataset that allows us to explore the geometry of multi-way splits.

The new data set can be seen in the lower left-hand corner of this Figure. It is easy to see that there are five groups of data. Imagine that each of the colors has a name associated with it (possibly G1 to G5).

If an analyst were to run a PROC Freq on the group variable he would find five levels of group(G1 to G5).

This would lead him to request a five way split in the decision tree. The tree only finds four and the Figure 13 explores why that happens.


•Why can we not find N groups in one split?

•5 - way

•5 groups

•Ask E. Miner to Split 5 ways

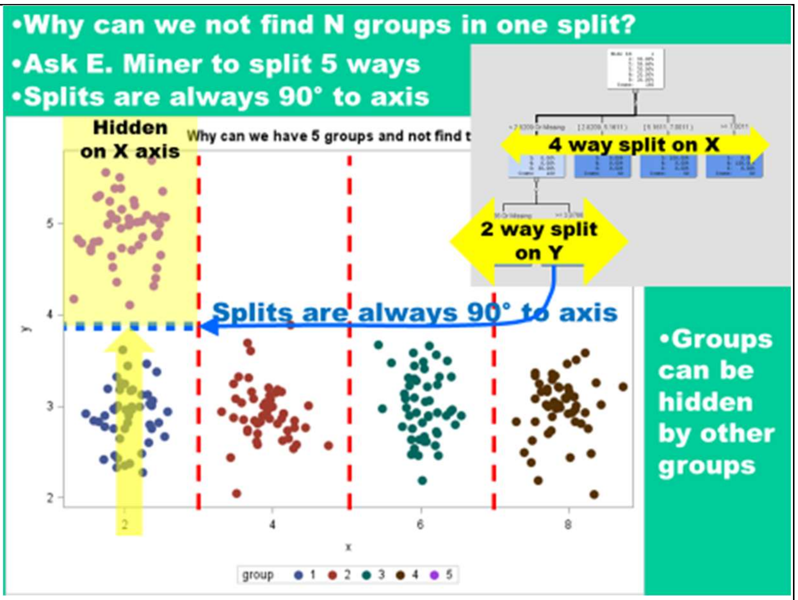
New Data

Figure 12

Figure 13 shows the geometry of the decision tree. The first level of the tree splits on levels of the X variable. The splits are shown by the dashed red lines perpendicular to the x-axis.

If you look above X equals 2, you can see that a group is hidden behind another group.

Splitting on X, with a split boundaries being perpendicular to an axis, cannot separate these two groups. The second level split is on the Y axis and is shown by the dashed blue line.



•Why can we not find N groups in one split?

•Ask E. Miner to split 5 ways

•Splits are always 90° to axis

•Groups can be hidden by other groups

Figure 13

It is thought that the phenomenon of groups hiding behind each other is the explanation why a request for an N-way split does not return N splits. In Figure 13 we requested a five way split and the algorithm gave us a four-way split.

The authors use this slide as an explanation for lack of mathematical rigor in this paper. The ability of a multi-way split to outperform a two-way split is determined by positions of the points in hyperspace. If the points line up “nicely”, as they do in Figure 13, a multi-way split can be very effective and easy to interpret. However; a proof of how frequently a multi-way split is likely to outperform a two-way split – in terms of accuracy – would be a proof of how groups of points are distributed in hyperspace. This is beyond our skills and not to our interest.

CONCLUSION

Multi-way splits can be very powerful especially if the analyst starts off the modeling process with a few PROC Freqs to find out how many levels there are in categorical variables (both Y and X).

Setting the number of splits to be equal to the number of levels in categorical variables, either X or Y, might result in trees that are easy to explain to subject matter experts.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Russ Lavery
Contractor
Russ.lavery@verizon.net

YuTing Tian
YT899963@WCUPA.edu